JITEL
Valencia, 2017

Libro de Actas

XIII Jornadas de Ingeniería Telemática (JITEL 2017)

27-29 de septiembre de 2017

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE UPV INGENIEROS DE TELECOMUNICACIÓN

IGIC
Instituto de investigación para la gestión integrada de zonas costeras

Instituto ITACA
Tecnologías de la Información y Comunicaciones

Colección Congresos UPV
Proceedings XIII Jornadas de Ingenieria Telematica - JITEL2017

Los contenidos de esta publicación han sido evaluados por el Comité Científico que en ella se relaciona y según el procedimiento que se recoge en http://ocs.editorial.upv.es/index.php/JITEL/JITEL2017

www.lalibreria.upv.es Ref.: 6380_01_01_01

ISBN: 978-84-9048-595-8

DOI: http://dx.doi.org/10.4995/JITEL2017.2017.7061

Mensaje de Bienvenida

Presentación JITEL 2017

En esta ocasión, la Universidad Politécnica de Valencia es la encargada de servir de anfitriona a las XIII Jornadas de Ingeniería Telemática (JITEL 2017), que se celebran del 27 al 29 de Septiembre de 2017. JITEL, organizadas por la Asociación de Telemática (ATEL), están constituyendo un foro consolidado de reunión, debate y divulgación para los grupos que imparten docencia e investigan en temas relacionados con las redes y los servicios telemáticos. Este evento fomenta, por un lado el intercambio de experiencias y resultados, además de la comunicación y cooperación entre los grupos de investigación que trabajan en temas relacionados con la Ingeniería Telemática.

JITEL 2017 se celebra conjuntamente con las V Jornadas de Innovación Educativa (V JIE 2017) y el III Workshop QoS y QoE en Comunicación Multimedia (III QQCM 2017). Además, por segunda edición consecutiva, se ha contado con un FastTrack para la presentación de trabajos de investigación de interés para el área de Ingeniería Telemática básicamente sustentado por los resultados relevantes en los dos últimos años.

En la presente edición se ha contado con los ponentes invitados Albert Banchs (Universidad Carlos III, Madrid), Joel J. P. C. Rodrigues (National Institute of Telecommunications (Inatel), Brazil e Instituto de Telecomunicações, Portugal) y Carlos Delgado Kloos (Universidad Carlos III, Madrid).

Estas actas contienen 56 contribuciones, organizadas en 12 sesiones e incluyen los siguientes ámbitos temáticos: calidad de servicio, calidad de experiencia, transmisión de vídeo, redes inalámbricas de sensores y de área local, seguridad, tráfico, movilidad, computación en la nube, 5G, virtualización, redes definidas por software, multicast, broadcast, e internet de las cosas.

Los presidentes del comité de programa queremos expresar nuestro agradecimiento a todos los patrocinadores y colaboradores del evento, Asociación de Telemática (ATEL), Universidad Politécnica de Valencia (UPV), Escuela Técnica Superior de Ingenieros de Telecomunicación (ETSIT-UPV), Instituto de Investigación para la Gestión Integrada de Zonas Costeras (IGIC-UPV) e Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA-UPV). También, nuestro agradecimiento a los ponentes invitados, a los autores de los artículos enviados y a los ponentes de los artículos presentados. Igualmente extender nuestro agradecimiento a los distintos comités; al comité de programa por la gestión de las revisiones, a los revisores anónimos, al comité editorial (Editorial UPV) por su soporte informático, al comité ejecutivo por su asesoramiento y al comité organización por su tiempo y apoyo para llevar adelante la conferencia. Finalmente agradecer a todos los asistentes su participación. Todos los colectivos citados han hecho posible el evento JITEL 2017. Muchísimas gracias.

Esperamos que el programa técnico de JITEL 2017 sea de vuestro agrado e interés y que podáis disfrutar de la ciudad de Valencia.

Jaime Lloret Mauri
Vicente Casares-Giner
Presidentes del Comité de Programa JITEL 2017

Presentación JIE 2017

Se cumple ya la quinta edición de las Jornadas de Innovación Educativa en Ingeniería Telemática (JIE V), celebradas en esta ocasión en Valencia, del 27 al 29 de septiembre de 2017. Al igual que en las ediciones anteriores, que tuvieron lugar en Valladolid, Santander, Granada y Palma, estas Jornadas se presentan como una oportunidad para crear un foro de encuentro e intercambio de experiencias de innovación docente entre profesores en el ámbito de la Ingeniería Telemática.

Las temáticas abarcadas se han estructurado en tres grandes ámbitos: metodologías docentes y organización, evaluación y seguimiento en la docencia y, por último, las tecnologías de la información y las comunicaciones aplicadas a la docencia. Se trata de un amplio abanico bajo el que tienen cabida todos aquellos trabajos y experiencias que se refieren a la innovación en la docencia de las disciplinas relacionadas con la telemàtica.

Los trabajos presentados para su posible aceptación en estas Jornadas se sometieron a un riguroso proceso de revisión por pares, en el que cada trabajo fue evaluado por tres revisores, garantizando así, bajo la supervisión del Comité de Programa, la originalidad y la calidad de su contenido. En estas actas se recogen las ocho contribuciones que fueron finalmente aceptadas para su presentación en dichas Jornadas. En ellas se refleja como la adaptación al Espacio Europeo de Educación Superior (EEES) y la implantación progresiva de los nuevos títulos de grado y posgrado en las universidades españolas, han propiciado la aplicación de nuevas tecnologías en el diseño y preparación de asignaturas, la planificación docente, la conducción de las clases y el seguimiento, tutorización y evaluación de los estudiantes. Estas presentaciones se han estructurado en 2 sesiones desarrolladas conjuntamente con las XIII Jornadas de Ingeniería Telemática (JITEL 2017).

Para concluir, quisiera agradecer al Comité de Programa su colaboración en todas las tareas organizativas de estas Jornadas, así como a los revisores su inestimable trabajo. Sin su participación no hubiera sido posible llevar a cabo estas Jornadas que, edición tras edición, se consolidan como un referente en la innovación educativa del área de Ingeniería Telemática.

Desde la Universitat Politècnica de València os damos la bienvenida y os animamos a disfrutar de estas Jornadas.

Jaume Ramis Bibiloni
Comité de Programa de JIE 2017

Presentación QQCM 2017

Internet se encuentra en una encrucijada, sus usuarios se han acostumbrado a la rápida implantación de acceso a contenidos multimedia, transferencia de información multimedia entre usuarios particulares, videoconferencias, juegos online... La ingente cantidad de información personalizada y los servicios disponibles crecen exponencialmente, pero los consumidores se ven atados a una red best effort que ofrece una respuesta muy inestable. Proporcionar nuevos servicios multimedia a través de Internet permite una gran flexibilidad, pero la calidad ofrecida puede ser bastante pobre. Por eso es necesario un análisis de la calidad ofrecida por la red, QoS (Quality of Service) y la calidad percibida por el usuario del servicio, QoE (Quality of Experience).

En esta línea, III Workshop QoS y QoE en Comunicación Multimedia (III QQCM 2017) abarca los siguientes tópicos de interés: repaso de las principales arquitecturas desarrolladas para asegurar la QoS en Internet y si están preparadas para soportar la implantación masiva de nuevos servicios multimedia. Propuesta de nuevos protocolos de comunicación para los servicios multimedia y evaluación de los mismos. Análisis de cómo la calidad percibida está en consonancia con los parámetros de QoS asegurados por las redes para servicios multimedia. Propuesta de nuevos parámetros adicionales (facilidad de manejo, receptividad, preferencias...) que introducen el factor humano y definen precisamente el consumo del servicio multimedia o la forma en que el usuario interactúa con el mismo.

Julián Fernández Navajas
Comité de Programa de QQCM 2017

Comité Ejecutivo

- Jaime Lloret Mauri, Universidad Politécnica de Valencia, España (JITEL 2017)
- Vicente Casares-Giner, Universidad Politécnica de Valencia, España (JITEL 2017)
- Magdalena Payeras Capellá, Universitat de les Illes Balears, España (JITEL 2015)
- Julián Fernández Navajas, Universidad de Zaragoza, España (JITEL 2019)
- Álvaro Suárez Sarmiento, Universidad de las Palmas de Gran Canaria, España (ATEL)

Comité de Programa

Presidentes del Comité de Programa:

- Jaime Lloret Mauri, Universidad Politécnica de Valencia, España
- Vicente Casares-Giner, Universidad Politécnica de Valencia, España

Miembros del Comité de Programa:

- Álvaro Suárez Sarmiento (Universidad de Las Palmas de Gran Canaria)
- Juanjo Unzilla Galán (Euskal Herriko Unibertsitatea)
- Enrique de la Hoz (Univesidad de Alcalá)
- Javier Aracil Rico (Universidad Autónoma de Madrid)
- Ramón Agüero (Universidad de Cantabria)
- Antonio de la Oliva Delgado (Universidad Carlos III de Madrid)
- Iria Manuela Estevez Ayres (Universidad Carlos III de Madrid)
- Victor Manuel Carneiro Díaz (Universidade da Coruña)
- Javier Vicente Sáez (Universidad de Deusto)
- José Luis González Sánchez (Universidad de Extremadura)
- Jorge Navarro Ortiz (Universidad de Granada)
- Sebastián García Galán (Universidad de Jaén)
- Pedro Merino Gómez (Universidad de Málaga)
- Javier Gozálvez Sempere (Universidad Miguel Hernández de Elche)
- Guiomar Corral Torruella (Universitat Ramon Llull)
- Pedro Miguel Ruiz (Universidad de Murcia)
- Roberto García Fernández (Universidad de Oviedo)
- Pilar Manzanares López (Universidad Politécnica de Cartagena)
- Xavier Hesselbach Serra (Universidad Politècnica de Catalunya)
- Gabriel Huecas Fernández-Toribio (Universidad Politécnica de Madrid)
- Jorge Martínez Bauset (Universitat Politècnica de Valencia)
- Daniel Morató Osés (Universidad Pública de Navarra)
- Santiago Felici Castell (Universitat de Valencia)
- Yannis Dimitriadis (Universidad de Valladolid)

- Raúl Fernando Rodríguez Rubio (Universidade de Vigo)
- Julián Fernández Navajas (Universidad de Zaragoza)

Comité Editorial

- Ramón Agüero (Universidad de Cantabria)
- Jaime Lloret Mauri (Universitat Politècnica de València)
- Vicente Casares-Giner (Universidad Politécnica de Valencia)
- Guillem Femenias Nadal (Universitat de les Illes Balears)

Comité Organización

Logistica Local:

- Oscar Romero Martínez, Universidad Politécnica de Valencia, España
- Jose Miguel Jiménez Herranz, Universidad Politécnica de Valencia, España
- Albert Rego, Universidad Politécnica de Valencia, España

Sitio web:

- Alejandro Canovas Solbes, Universidad Politécnica de Valencia, España

Difusión y publicidad de la llamada a ponencias:

- Sandra Sendra Compte, Universidad de Granada, España

Redes sociales:

- Miguel Garcia Pineda, Universidad de Valencia, España

Soporte envío de artículos:

- Lorena Parra Boronat, Universidad Politécnica de Valencia, España
- Laura Garcia Garcia, Universidad Politécnica de Valencia, España

Comité QQCM 2017

- Fidel Liberal Malaina
  Grupo de Investigación: Networking, Quality and Security (NQaS). Universidad del Pais Vasco (EHU)
- María Ángeles Vázquez Castro
  Grupo de Investigación: Wireless and Satellite Communications (WSC).
  Universidad autónoma de Barcelona (UAB)
- Juan José Ramos Muñoz
  Grupo de Investigación: Wireless and multimedia Networking Lab (WMNL).
  Universidad de Granada (UGR)
- Elsa María Macías López
  Grupo de Investigación: Grupo de Arquitectura y Concurrencia (GAC).
  Universidad de Las Palmas de Gran Canaria (ULPGC)
- Matías Toril Genovés
  Grupo de Investigación: Mobile Network Optimization (Mobilenet)
  Universidad de Málaga (UM)
- Mónica Aguilar Igartua
  Grupo de Investigación: Servicios Telemáticos (SerTel)
  Universidad Politécnica de Cataluña (UPC)
- María Dolores Cano Baños
  Grupo de Investigación: Ingeniería Telemática (IT)
  Universidad Politécnica de Cartagena (UPCT)
- Juan Carlos Guerri Cebollada
  Grupo de Investigación: Communications Multimedia (COMM)
  Universidad Politécnica de Valencia (UPV)
- Jaime Lloret Mauri
  Grupo de Investigación: Comunicaciones y teledetección. Instituto de investigación IGIC
  Universidad Politécnica de Valencia (UPV)
- Ramón Agüero Calvo
  Grupo de investigación: Grupo de ingeniería telemática.
  Universidad de Cantabria
- Fernando Boronat Seguí
  Grupo de investigación: Immersive Interactive Media (IIM) R&D Group
  Universitat Politècnica de València - Campus de Gandia
- Miguel García Pineda
  Grupo de Investigación: High Performance and Intelligent Systems (HiPIS)
  Universitat de València (UV)
- Julián Fernández Navajas
  Grupo de investigación: Communications Networks and Information Technologies for e-Health and Quality of Experience (CeNITEQ).
  Universidad de Zaragoza (UZ)

Comité JIE 2017

- Raquel M. Crespo García
  Universidad Carlos III de Madrid
- José Manuel Giménez Guzmán
  Universidad de Alcalá
- Isaías Martinez Yelmo
  Universidad de Alcalá
- José Ángel Irastorza Teja
  Universidad de Cantabria
- Alberto Eloy García Gutiérrez
  Universidad de Cantabria
- Ramón Agüero Calvo
  Universidad de Cantabria
- Francisco Barceló Arroyo
  Universitat Politècnica de Catalunya
- Elsa María Macías López
  Universidad de las Palmas de Gran Canaria
- Diego Fernández Iglesias
  Universidade da Coruña
- Fidel Cacheda Seijo
  Universidade da Coruña
- Jorge Navarro Ortiz
  Universidad de Granada
- Jesús Martínez Cruz
  Universidad de Málaga
- Jaume Ramis Bibiloni
  Universitat de les Illes Balears
- María Jesús Verdú Pérez
  Universidad de Valladolid
- Martín Llamas Nistal
  Universidad de Vigo
- Guillermo Azuara Guillén
  Universidad de Zaragoza

Lista de Revisores

- Ramón Agüero Calvo
- Mónica Aguilar Igartua
- Javier Aracil Rico
- Jasone Astorga Burgo
- Guillermo Azuara Guillén
- Francisco Barceló Arroyo
- Fernando Boronat-Segui
- María Victoria Bueno Delgado
- Fidel Cacheda Seijo
- Maria-Dolores Cano
- Javier Carmona Murillo
- Victor Manuel Carneiro Díaz
- Vicente Casares-Giner
- Cristina Cervelló Pastor
- Guiomar Corral Torruella
- David Cortés Polo
- Antonio de la Oliva Delgado
- Yannis Dimitriadis
- Antonio José Estepa Alonso
- Iria Manuela Estevez Ayres
- Santiago Felici Castell
- Diego Fernández Iglesias
- Julián Fernández Navajas
- Gabriel Huecas Fernández-Toribio
- José Ramón Gállego-Martínez
- Laura Garcia
- Marta García
- Roberto García Fernández
- Sebastián García Galán
- Alberto Eloy García Gutiérrez
- Miguel García-Pineda
- José Manuel Giménez Guzmán
- Juan Carlos Guerri-Cebollada
- Ángela Hernández-Solana
- Xavier Hesselbach Serra
- José Ángel Irastorza Teja
- Eduardo Jacob Taquet
- Jose Miguel Jimenez
- Jorge Lanza
- Fidel Liberal-Malaina
- Jaime Lloret
- Javier López
- Elsa María Macías López
- Pilar Manzanares López
- Domingo Marrero Marrero
- Iván Marsá Maestre

- Jorge Martínez Bauset
- Pedro Merino Gómez
- Pedro Miguel Ruiz
- Daniel Morató Osés
- Juan Pedro Muñoz Gea
- Jorge Navarro Ortiz
- Lorena Parra
- Jaume Ramis Bibiloni
- Juanjo Ramos
- Albert Rego
- Javier Rodríguez Pérez
- Raúl Fernando Rodríguez Rubio
- Jose Oscar Romero
- Luis Sánchez
- José Antonio Sánchez Sánchez
- Miquel Soriano
- Ignacio Soto
- Álvaro Suárez Sarmiento
- Miran Taha
- Matías Toril-Genovés
- Juanjo Unzilla Galán
- Francisco Valera
- Mª Ángeles Vazquez-Castro
- María Jesús Verdú Pérez
- Javier Vicente Sáez

KEYNOTE SPEAKERS

Keynote Speaker 1

Network slicing: Enabling Customization in 5G Mobile Networks

Albert Banchs, Univeristy Carlos III, Madrid, Spain

Abstract: There is consensus among the relevant industry and standardization communities that a key element in 5G mobile networks will be network slicing. The idea is to allow the mobile infrastructure to be "sliced" into logical networks, where each slice is a collection of resources and functions that includes software modules running at different locations as well as the nodes' computational and communication resources. The intention is to tailor each slice to support a specific service, providing only what is necessary for the service while avoiding unnecessary overheads and complexity. This provides a basis for efficient infrastructure sharing among diverse entities, ranging from classical or virtual mobile network operators to new players that simply view connectivity as a service, where each of these entities may be running one or more slices. This talk will focus on the key enablers for network slicing and the research challenges involved in realizing this technology. Current standardization activities will be reviewed along with the contributions of major research projects, such as the H2020 5G-NORMA and 5G-MoNArch projects. A key problem underlying network slicing is enabling efficient sharing of mobile network resources. Various approaches considered in 3GPP will be analysed, ranging from per-reservation based schemes (where network slices reserve the required resources in advance) to others based on network shares (where resources are allocated based on pre-determined shares). The performance and behavior of the various approaches will be studied based on analytical tools including optimization, game theory and machine learning. Buildig on these analyses, we will provide some insights on the stability, peformance, optimality and level of customization enabled by the various approaches.

Short Bio: Dr. Albert Banchs received his Telecommunications Engineering degree from UPC in 1997, and the PhD degree from the same university in 2002. He visited the ICSI, Berkeley, in 1997, worked for Telefonica I+D in 1998, and for the Network Laboratories of NEC Europe Ltd., Germany, from 1998 to 2003. Since 2003, he is with Univeristy Carlos III of Madrid, where he is currently a Full Professor, and since 2009 he has a double affiliation as Deputy director of the IMDEA Networks research institute. He was Academic Guest at ETHZ in 2012, and Visiting Professor at EPFL in 2013 and 2015. Prof. Banchs is currently editor for IEEE/ACM Transactions

on Networking and IEEE Transactions on Wireless Communications, and regularly serves in the Technical Programme Committee of many conferences in the area, including IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, IEEE WoWMoM and IEEE WCNC, among others. Prof. Banchs has been a key contributor to a number of EU projects, and has been the coordinator of two of them. He has also run several industry contracts and is the inventor of 5 granted patents. Currently, his main effort is the technical management of the EU H2020 5G-MoNArch project, which is one of the flagship 5GPPP phase 2 projects focusing on network architecture. Prof. Banchs has received a number of awards, including the national prize to the best PhD thesis on broadband networks and the runner-up award to the best collaborative project in the region of Madrid, in addition to several paper awards at conferences as well as outstanding awards for supervised PhD theses. He a Senior Member of IEEE and an IEEE Distinguished Lecturer.

Keynote Speaker 2

Challenges and Perspectives Towards IoT Deployment

Dr. Joel J. P. C. Rodrigues, National Institute of Telecommunications (Inatel), Brazil // Instituto de Telecomunicações, Portugal.

Abstract: This keynote addresses a hot and updated topic focusing on Internet of Things (IoT), considering their most relevant challenges and opportunities. It starts with an introduction to IoT and its typical application scenarios considering different verticals. After, an initiative to prepare ICT professionals for new challenges regarding this new generation technologies for IoT will be presented. A special attention will be given to the Inatel Smart Campus, an open Campus for research on IoT, experiments, and concepts and technology validation. Inatel has sponsored this project, started in August 2016, open for companies' participation and promoting the academy-enterprise interaction. It is a true living lab for several IoT verticals, including smart cities and smart homes. New challenges and opportunities on IoT are discussed. The communication ends with new trends and issues on Internet of Things, suggesting further research topics.

Short Bio: Joel J. P. C. Rodrigues (joeljr@ieee.org) [S'01, M'06, SM'06] is a professor and senior researcher at the National Institute of Telecommunications (Inatel), Brazil and senior researcher at the Instituto de Telecomunicações, Portugal. He has been professor at the University of Beira Interior (UBI), Portugal and visiting professor at the University of Fortaleza (UNIFOR), Brazil. He received the Academic Title of Aggregated Professor in informatics engineering from UBI, the Habilitation in computer science and engineering from the University of Haute Alsace, France, a PhD degree in informatics engineering and an MSc degree from the UBI, and a five-year BSc degree (licentiate) in informatics engineering from the University of Coimbra, Portugal. His main research interests include e-health, sensor networks and IoT, vehicular communications, and mobile and ubiquitous computing. Prof. Joel is the leader of NetGNA Research Group (http://netgna.it.ubi.pt), the President of the scientific council at ParkUrbis – Covilhã Science and Technology Park, the Past-Chair of the IEEE ComSoc Technical Committee on eHealth, the Past-chair of the IEEE ComSoc Technical Committee on Communications Software, Steering Committee member of the IEEE Life Sciences Technical Community and Publications co-Chair,

and Member Representative of the IEEE Communications Society on the IEEE Biometrics Council. He is the editor-in-chief of the International Journal on E-Health and Medical Communications, the editor-in-chief of the Recent Advances on Communications and Networking Technology, the editor-in-chief of the Journal of Multimedia Information Systems, and editorial board member of several high-reputed journals. He has been general chair and TPC Chair of many international conferences, including IEEE ICC, GLOBECOM, and HEALTHCOM. He is a member of many international TPCs and participated in several international conferences organization. He has authored or coauthored over 500 papers in refereed international journals and conferences, 3 books, and 2 patents. He had been awarded several Outstanding Leadership and Outstanding Service Awards by IEEE Communications Society and several best papers awards. Prof. Rodrigues is a licensed professional engineer (as senior member), member of the Internet Society, an IARIA fellow, and a senior member ACM and IEEE.

Keynote Speaker 3

¿Le están moviendo el queso a la Universidad?

Prof. Dr. Carlos Delgado Kloos, Universidad Carlos III de Madrid, Spain (UC3M)

Abstract: Estamos hartos de escuchar cómo las TIC están transformando todos los sectores de la sociedad. Sin embargo, a la universidad española parece que no le afecta o le afecta poco. Las estructuras existentes, la normativa y la inercia no son propicias para la adaptación a un nuevo contexto, si lo hubiere.

En la presentación analizaremos si efectivamente nos encontramos en un nuevo contexto para la educación superior, presentaremos casos de éxito y reflexionaremos sobre posibles formas para pasar de la educación de la era industrial a la de la sociedad de la información.

Short Bio: Carlos Delgado Kloos es Ingeniero de Telecomunicación por la Universidad Politécnica de Madrid y Doctor en Informática por la Universidad Técnica de Múnich. Es Catedrático de Ingeniería Telemática en la Universidad Carlos III de Madrid, donde es también Director del Grupo de Investigación GAST y Director de la Cátedra UNESCO sobre "Educación Digital Escalable para Todos". Además, es Vicerrector de Estrategia y Educación Digital en su Universidad y es el representante español en el comité TC3 sobre Educación de IFIP. Su área principal de investigación es la tecnología educativa. Ha liderado una multitud de proyectos de investigación tanto a nivel europeo, como nacional y bilateral y coordina la red eMadrid sobre Tecnología Educativa en la Comunidad de Madrid. El número de contribuciones científicas en congresos o revistas nacionales e internacionales supera las 400. Además, ha escrito un libro y co-editado más de una docena. También ha coordinado e impartido varios MOOCs, entre los que cabe destacar uno sobre programación con Java en la plataforma edX (www.edx.org/professional-certificate/uc3mx-introduction-java-programming).

TABLA DE CONTENIDOS

Miércoles, 27 de septiembre 2017

Sesión 1: QoS & QoE & Video

Session Chair: Antonio Estepa

Sesión 2: WSN & WLAN

Session Chair: Alvaro Suarez

Sesión 3: Seguridad

Session Chair: Mª Dolores Cano

Sesión 4: Fast Track

Session Chair: Vicente Casares

Jueves, 28 de septiembre 2017

Sesión 5: Tráfico & Movilidad

Session Chair: Jorge E. López de Vergara

Sesión 6: QQCM-I

Session Chair: Juan C. Guerri

Sesión 7: Cloud Computing & 5G

Session Chair: Ramón Agüero

Sesión 8: QQCM-II

Session Chair: Julián Fernández Navajas

Sesión 9: Virtualizacion & SDN

Session Chair: Xavier Hesselbach

Viernes, 29 de septiembre 2017

Sesión 11: JIE-I

Session Chair: Maria Magdalena Payeras Capellà

Sesión 12: JIE-II

Session Chair: Santiago Felici

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6497

FlexiTop: sistema escalable y flexible de medidas de calidad para servicios *Over-The-Top*

Daniel Perdices*†, Jorge E. López de Vergara*†, Paula Roquero*, Carlos Vega*†, Javier Aracil*†
*Departamento de Tecnología Electrónica y de las Comunicaciones, Universidad Autónoma de Madrid
Escuela Politécnica Superior, Calle Francisco Tomás y Valiente, 11, 28049 Madrid
†Naudit High Performance Computing and Networking, S.L.
Calle Faraday, 7, 28049 Madrid
daniel.perdices@naudit.es, {jorge.lopez_vergara, paula.roquero, javier.aracil}@uam.es, carlos.vega@naudit.es

***Resumen*—Hoy en día, el uso de los servicios *Over-The-Top* tales como *streaming* de vídeo, servicios web y redes sociales es dominante en el tráfico de Internet. En consecuencia, existe un amplio interés por parte de los proveedores de servicios de Internet en conocer la calidad de este tipo de comunicaciones para poder proporcionar el mejor servicio y por tanto la mejor experiencia a los usuarios. Para este propósito, se propone FlexiTop, un sistema flexible y escalable de medidas activas de calidad para este tipo de servicios, que permite obtener métricas con un consumo de recursos contenido. El diseño propuesto ha sido implementado y validado mediante pruebas. Así mismo se ha realizado una campaña de medidas durante varios meses en diferentes equipos, pudiéndose concluir que el sistema cumple con las expectativas de los actores involucrados.**

***Palabras Clave*—QoS, medidas activas, servicios OTT, streaming, servicios web.**

I. INTRODUCCIÓN

Un servicio *Over-The-Top* (OTT) se define como aquel que se provee a través de la red sin que ningún operador de telecomunicaciones esté involucrado en su envío, desarrollo o planificación [1], es decir, un servicio que utiliza la red como mero canal de transporte sin ningún otro tipo de consideración.

Los servicios OTT tienen necesidades y naturalezas muy heterogéneas. Mientras que, por un lado, los servicios de *streaming* multimedia requieren un ancho de banda dado y poca variabilidad del mismo, los servicios web tienen como prioridad el tiempo de respuesta. Además, desde el punto de vista de Calidad de la Experiencia (*Quality of Experience*, QoE) [2], existen otras métricas más relevantes que las tradicionalmente utilizadas en el análisis de tráfico de red. Por ejemplo, en estos casos es de interés no únicamente el tiempo de respuesta de las peticiones, sino también el tiempo total para finalizar una operación de alto nivel o una transacción completa. Estas diferencias llevan a la necesidad de analizar parámetros distintos según cada servicio.

Más allá de los parámetros que se desean analizar, muchos servicios actualmente funcionan bajo HTTPS, lo que dificulta el análisis pasivo de los mismos. Por ejemplo, es necesario poder ver el contenido de las peticiones HTTP para obtener métricas que son de gran interés para los demandantes de este sistema, tales como el porcentaje de fallos HTTP o los códigos de respuesta de las peticiones. Por ello, las medidas activas —en las cuales se puede monitorizar completamente el tráfico que envía y recibe el agente de medición— son prácticamente la única forma de identificar y analizar el tráfico que generan estos servicios sin recurrir a mecanismos para descifrar la conexión, solución que está limitada a escenarios muy concretos.

Además, el cambio constante de estos servicios, no solo en términos de prestaciones sino también en cuanto a funcionalidad y mecanismos de comunicación, hace necesario que los sistemas de medición evolucionen al mismo ritmo que lo hacen dichos servicios. Actualmente, centrarse en herramientas que midan un solo protocolo, o que se basen en una arquitectura diseñada exclusivamente sobre el protocolo que se quiere analizar, tendrán un periodo de vida corto.

Adicionalmente, el interés por estas medidas no proviene únicamente de los Proveedores de Servicio de Internet (*Internet Service Providers*, ISP), sino también de los usuarios. Como respuesta a esta necesidad, en este artículo se plantea FlexiTop, un sistema comercializado por naudit que se puede ejecutar en equipos de muy bajo coste y que permite un despliegue masivo de manera que se puedan registrar, agregar y analizar las medidas tomadas en cada uno de los dispositivos de una red de miles de puntos de medida.

Este análisis de servicios como YouTube [3] u otros servicios web [4] se ha realizado previamente de manera bastante exhaustiva en entornos de pruebas controlados. Sin embargo, la propuesta de este artículo es un sistema que permita tomar medidas de Indicadores Clave de

Prestaciones (*Key Performance Indicator*, KPI) de la aplicación y medidas a nivel de red, centrándose en aquellas métricas que son de interés para los ISP y los usuarios. Además, esta propuesta también aborda aspectos como la escalabilidad o el consumo de recursos, que en otros trabajos no han sido objeto de estudio.

La propuesta ha sido probada en diferentes tipos de dispositivos, abarcando PCs de sobremesa, entornos virtualizados y sistemas empotrados de bajo coste. Los resultados han sido satisfactorios en todos los casos, las métricas se han validado con distintas alternativas y se ha logrado elaborar comparativas y análisis de estos servicios que permiten identificar los problemas que pueden causar una deficiente calidad de experiencia.

Para explicar el funcionamiento de FlexiTop, el artículo se ha estructurado de la siguiente manera: La sección II realiza un estudio de las tecnologías más usadas en la actualidad en los servicios OTT, explicándose a continuación en la sección III las diferentes alternativas a esta propuesta. Posteriormente, la sección IV enumera las métricas que son de interés tanto para los usuarios como para los ISP. Tras ello, en la sección V se describe la arquitectura del agente de medición. En relación con las métricas, en la sección VI se analizan la configuración de pruebas y los diferentes *tests* desarrollados, para así exponer los diferentes métodos y aproximaciones necesarias para medir un servicio a la vez que se muestran y comentan algunos de los resultados obtenidos en la sección VII. Finalmente se presentan las conclusiones y líneas de trabajo futuro.

II. PRELIMINARES

Para poder analizar los servicios, es necesario conocer las tecnologías y estándares que permiten que estos se presten. En el caso de los servicios de *streaming* multimedia, actualmente se utilizan ampliamente distintas tecnologías con bastantes similitudes. Los ejemplos más usados de estas son:

1) *Dynamic Adaptive Streaming over HTTP* (DASH, *Streaming* Adaptativo Dinámico sobre HTTP): técnica de *streaming* con tasa binaria adaptativa desde servidores HTTP convencionales [5].
2) *HTTP Live Streaming* (HLS, *Streaming* en Vivo con HTTP): técnica similar a la anterior con servidores convencionales que almacenan fragmentos con distintos tipos de codificación [6].

La mayor parte de las tecnologías anteriores comparten la característica de utilizar HTTP(S) como canal de transporte. Esta naturaleza permite que medir la calidad de estos servicios se base, en parte, en medir la calidad de estas conexiones HTTP. Sin embargo, se encuentran dificultades cuando se resuelven las URLs de los recursos solicitados ya que en la mayoría de casos estos fragmentos de vídeo son servidos por una Red de Distribución de Contenidos (*Content Delivery Network*, CDN), luego su resolución depende de la geolocalización del cliente y del servicio de nombres (*Domain Name System*, DNS).

Un factor importante aparte de la tecnología que provee el servicio es la configuración del reproductor o de la aplicación cliente. En este caso, existen parámetros como el tamaño de los *buffers*, los cambios programados de una calidad de vídeo a otra o el códec que se utiliza que tienen una gran importancia en el correcto funcionamiento de los servicios. Sin embargo, estos no son aspectos que estén relacionados con el funcionamiento de la red y, por tanto, no son de aplicación en este artículo.

Por otro lado, existen diferentes métodos de comunicación para servicios web. Los más utilizados son:

1) *Representational State Transfer* (REST, Transferencia de Estados Figurativos): arquitectura de comunicación basada en recursos HTTP.
2) *Simple Object Access Protocol* (SOAP, Protocolo Sencillo de Acceso a Objetos): protocolo estándar de intercambio de datos y llamadas a procedimientos sobre HTTP.

Actualmente, tanto aplicaciones en páginas web como aplicaciones para dispositivos móviles utilizan protocolos basados en REST para la comunicación entre cliente y servidor, por lo que existe un claro interés en el análisis de los tiempos de respuesta de estas *RESTful APIs*.

III. ESTADO DEL ARTE

En la actualidad, existe una gran variedad de métricas estandarizadas a nivel de red y de transporte. Estas métricas proporcionan información relevante acerca de los flujos originados por un servicio y pueden proporcionar información de interés para analizar las causas de las métricas observadas a nivel de aplicación y de la QoE. Muchas de estas métricas han sido estandarizadas por el grupo de trabajo IPPM (*Internet Protocol Performance Metrics*) del IETF (*Internet Engineering Task Force*). En concreto, son de relevancia para este trabajo la medición de latencia en un sentido [7] y en ambos [8], que estandarizan las métricas de latencia a nivel de red y que por tanto muestran metodologías de medición que son de interés.

Otros trabajos [9] siguen una aproximación radicalmente distinta a las medidas activas para resolver el problema del tráfico cifrado, basada en caracterizar el tráfico de estos servicios a base de encontrar patrones a nivel de red y de flujo [10]. Esta aproximación, también plausible, requiere sin embargo un mayor cálculo y un desarrollo más extenso de modelos matemáticos que se adapten a estos patrones deseados. Se pueden encontrar caracterizaciones ya realizadas para Facebook [4] y YouTube [11].

Para los servicios OTT, se ha observado ampliamente que existe una alta correlación entre los parámetros de Calidad de Servicio (*Quality of Service*, QoS) y de QoE [12]. Por tanto, está probado el interés en el análisis de la calidad de servicio de las conexiones que utilizan estos servicios. También existen diferentes trabajos que analizan la calidad de experiencia y de servicio de YouTube [3] o de otros servicios de vídeo o televisión bajo demanda [13]. Sin embargo, ninguno de los casos

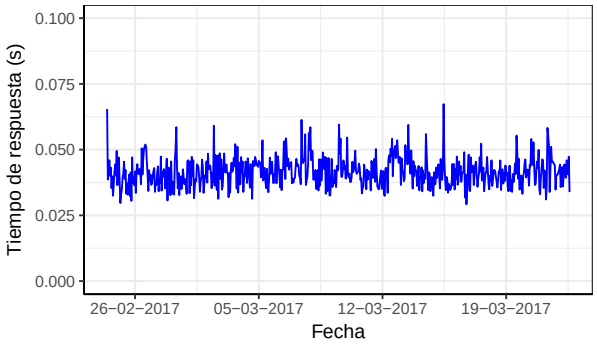


Fig. 1. Series de los tiempos de respuesta de DNS

adopta un enfoque más amplio que se centre en múltiples servicios.

Más allá de estas métricas clásicas y con el fin de poder obtener resultados más correlados con la QoE y de relevancia para los usuarios, se analizan KPIs del nivel de aplicación. Existen análisis previos de algunos de estos, especialmente en el impacto de aspectos de configuración de la aplicación como el tamaño del *buffer* en estas métricas y KPIs percibidos a nivel de aplicación. Queda fuera del ámbito de este trabajo analizar estos parámetros. En la medida de lo posible se proveerá una aproximación agnóstica a estos aspectos o en caso contrario, se proporcionará una configuración modificable para poder ver el impacto de estos parámetros.

IV. MÉTRICAS DE INTERÉS

Como se ha motivado en la introducción, las métricas han de diferir según el tipo de servicio que se desea medir. La caracterización de estas métricas y garantizar su uniformidad (estandarización) es por tanto necesario.

A. DNS

En primer lugar, es prioritario medir el rendimiento de las peticiones DNS por el interés que tiene conocer cómo se resuelven los nombres de dominio al acceder a un servicio. Con este propósito, se toma el tiempo de respuesta como principal métrica y además se almacenan las direcciones IP resultantes de la respuesta como apoyo para un posible análisis posterior.

En la Fig. 1 se muestra un ejemplo de serie temporal de la mediana de los tiempos de respuesta del servicio DNS para los 20 dominios más visitados en España. Para esta tarea de identificar estos dominios más visitados se puede utilizar un listado público. En este caso se ha utilizado Alexa[1]. Estas métricas se pueden comprobar mediante la herramienta `dig`[2].

B. HTTP

Ya que la mayoría de servicios funcionan sobre HTTP(S), se requiere de un sistema de medición específico para este protocolo.

La aproximación seguida se basa en medir el tiempo de envío de una petición (*Request spread*) y posteriormente medir el tiempo de recepción diferenciando dos marcas temporales: la llegada del primer paquete y la llegada del último. Estas dos marcas temporales van a servir para estimar lo que sería el tiempo de servicio y el tiempo de transmisión de la respuesta (*Response Spread*). Además, también se pueden obtener otras métricas del nivel de red (p.e., fragmentación) y de transporte (p.e., retransmisiones o ventanas 0) al analizar las conexiones asociadas a las peticiones HTTP.

Estas métricas se extienden al principal uso del protocolo analizado: la navegación web. Para este protocolo se pretenden dar estas métricas separadas según sea el documento base o las dependencias del mismo. Las dependencias se descargan concurrentemente. Se ha seguido una aproximación similar a la propuesta en TRANSUM [14], un disector para Wireshark que realiza un análisis de tiempos de respuesta de HTTP y otros protocolos como FTP o SMB2.

C. Streaming multimedia

Para servicios de *streaming* de contenido multimedia, un parámetro de probada relevancia es el *throughput*. No solo es un factor limitante para la calidad del contenido, sino que la variabilidad del mismo puede generar parones, cambios de calidad y otros problemas asociados al proceso de *buffering* que causan una mala calidad de experiencia.

Por tanto, para este tipo de servicios se realizan mediciones del ancho de banda consumido, del tiempo de descarga y el tiempo total de procesado (análisis de los manifiestos, obtención de las URLs, etc.) que se combinan con las medidas anteriormente mencionadas para HTTP.

D. Servicios web

Hay una variedad de servicios que funcionan sobre HTTP. Estos servicios abarcan desde redes sociales, servicios de mensajería hasta servicios de almacenamiento en la nube. Debido a su heterogeneidad, estos servicios van a tener unas necesidades variadas.

En general, los parámetros de interés en este caso van a ser el tiempo de acceso, tiempo de completado de las transacciones y velocidad de descarga. La naturaleza de las transacciones es un factor dependiente del servicio, por lo que en ciertos casos realizar estas transacciones puede ser excesivamente complejo. A diferencia de los servicios anteriores, el tiempo de procesado de la petición en el servidor y en el cliente es también un factor relevante.

E. Otros servicios

Más allá de los servicios que funcionan sobre HTTP, existen otros servicios que funcionan directamente sobre TCP/IP. Para estos servicios, se podrían dar, cómo se ha mencionado antes, métricas del nivel de red y transporte.

Sin embargo, el análisis del rendimiento de la aplicación resulta de mayor interés de cara a medidas más correladas con la QoE.

Uno de los principales servicios con estas características ha sido el correo electrónico a través de los protocolos más comunes de acceso: SMTP, IMAP y POP3.

[1] http://www.alexa.com/topsites/countries/ES
[2] https://ftp.isc.org/isc/bind9/cur/9.11/doc/arm/man.dig.html

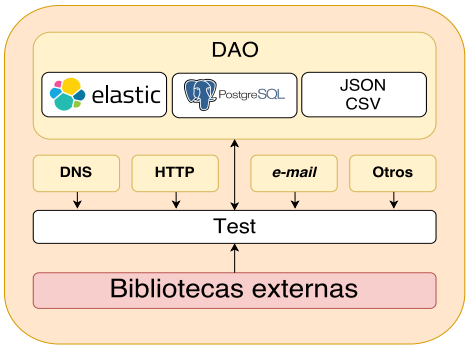


Fig. 2. Arquitectura del agente de medición.

V. ARQUITECTURA

En el diseño del sistema, se han tenido en cuenta los requisitos del sistema para dar la solución más adecuada. Los más relevantes para esta tarea han sido:

- **Flexibilidad**: como se ha motivado previamente, la flexibilidad del sistema tiene que ser acorde con el nivel de cambio de los servicios.
- **Escalabilidad**: de igual manera, el sistema ha de poder desplegarse en diferentes localizaciones sin requerir grandes procesos de configuración, de tal manera que los resultados estén disponibles de manera centralizada y de manera local.
- **Bajo consumo de recursos**: con el fin de ejecutar el sistema con el menor impacto posible y además en equipos de bajo coste, se pretende limitar el consumo del sistema así como proveer soluciones eficientes.

Estas prioridades resultan en una arquitectura del agente que realiza las medidas según se muestra en la Fig. 2:

1) **Módulo de *DNS***: este módulo es utilizado por todos los *tests* para medir el tiempo de resolución de las peticiones al servicio de nombres.
2) **Módulo de *HTTP***: este módulo permite realizar pruebas de manera concurrente o con un único hilo para estimar la máxima velocidad de descarga que se puede alcanzar. Soporta múltiples URLs, configuración de número máximo de hilos y distintas opciones para los *buffers* y repeticiones.
3) ***Database Access Object* (DAO, Objeto de Acceso a Bases de Datos)**: módulo que se encarga de abstraer la recolección (y el envío) de resultados y las herramientas de registro de errores.
4) **Otras bibliotecas y utilidades**: además se proveen bibliotecas propias (p.e. e-mail) y se pueden utilizar otras externas para construir las peticiones de las diferentes APIs del servicio o para el procesado del resultado de la petición.

Esta arquitectura permite no solo el desarrollo adecuado del proyecto, sino que garantiza que se puedan medir nuevos servicios en el futuro, así como modificar los ya implementados. Cada test se implementa según el siguiente esquema genérico: se obtienen las URLs del servicio, se resuelven mediante el módulo de DNS, se utiliza el modulo de HTTP u otro para calcular las diferentes métricas y se utiliza el DAO para guardar los resultados. En principio, los *tests* son independientes, pero pueden llamarse unos a otros o depender unos de otros para casos de prueba complejos.

Aparte de los agentes de medición, se tiene un servidor que puede realizar tareas tales como actualización de los *tests*, recepción de resultados, asignación de identificadores, activación o desactivación de *tests* o análisis estadístico. Esto permite que el agente no sea solo interesante para el cliente que instale el sistema, sino también para el proveedor que recibirá estadísticas y datos de interés sobre el funcionamiento de los servicios en la red que sean de ayuda para proporcionar un nivel superior de calidad a los usuarios.

VI. ENTORNO EXPERIMENTAL

El software del agente de medición ha sido desarrollado en Python, por la portabilidad y versatilidad que ofrece. Éste se ha probado en diferentes dispositivos:

1) Orange Pi Plus: H3 Quad-core Cortex-A7, 1GB DDR3, Gigabit Ethernet.
2) Equipo de pruebas (*host*): Intel Core i7 860, 8GB DDR3, Gigabit Ethernet.
3) Máquina virtual (*guest*): dos núcleos, 1GB DDR3, 2GB de disco duro, Gigabit Ethernet.

La propuesta completa ha estado ejecutándose de manera continuada en los diferentes equipos. En el primero de los mencionados, el sistema se dedica exclusivamente al sistema de medición, aunque nunca se utilizan todos los recursos disponibles. En el segundo caso, se ha ejecutado el sistema sobre el propio *host* y en diferentes entornos virtualizados: mediante una máquina virtual y mediante un contenedor Docker. Ambos equipos se encuentran conectados a la red de la Universidad Autónoma de Madrid.

El entorno virtualizado es interesante, porque permite no solo registrar los recursos consumidos sino también limitar y controlar el consumo de los mismos y el uso del *hardware* del *host*, p.e., el número de núcleos, la memoria RAM disponible o el espacio en disco.

El sistema servidor se encuentra instalado en el equipo de pruebas. Este sistema incluye visualización en tiempo real y análisis de datos.

VII. PRUEBAS Y RESULTADOS

Una vez se ha implementado la arquitectura propuesta, se implementan *tests* de diferentes servicios para validar el diseño realizado. Estos servicios son:

- **Servicios de *streaming* multimedia**: Netflix, YouTube, Yomvi, Spotify.
- **Servicios web**: Twitter, Facebook, Dropbox, Speedtest, Google (buscador), Navegación web.
- **Otros**: Correo electrónico.

En la implementación de cada *test* se han utilizado técnicas variadas para simular la utilización de cada uno de los servicios. A continuación se explican las particularidades de cada servicio, así como los principales procesos usados dependiendo del servicio, y se muestran algunos resultados obtenidos.

A. *Servicios de streaming multimedia*

El procedimiento para tomar medidas está orientado a obtener el ancho de banda utilizado. Para ello, se obtienen varias URLs de fragmentos de contenido multimedia y se descargan de manera simultánea utilizando varios hilos y conexiones.

La manera de obtener las URLs difiere según el servicio. Para YouTube, existen tanto bibliotecas como formas de descargar el vídeo analizando el manifiesto DASH. De manera muy similar, Spotify provee una API REST[3] que permite obtener URLs de muestras de cada canción. Estos servicios exponen alguna forma de acceso a las URLs, por lo que no es necesario utilizar técnicas más avanzadas.

No obstante, para otros servicios implementados, se ha requerido de la técnica de interposición (*Man in the Middle*) para analizar las conexiones HTTPS que se utilizan en el servicio. Esto permite hallar las URLs y analizar la lógica del servicio. Con el mismo propósito, se ha estudiado también el código de reproductores multimedia *online* para determinar con mayor precisión el funcionamiento del servicio y validar la información obtenida. Un ejemplo de este procedimiento ha sido el caso de Yomvi, en el que el análisis del código y de las conexiones HTTPS ha permitido realizar pruebas de rendimiento del servicio con un consumo de recursos reducido y sin la necesidad de ejecutar un navegador web.

Sin embargo, para otros casos este proceso puede ser tedioso e incluso resultar intensivo para el agente de medición. En estos casos no queda otra alternativa más que ejecutar un navegador durante un corto periodo de tiempo para probar el servicio. Aunque esta aproximación es viable y se han desarrollado herramientas que se integran con el navegador para capturar estas peticiones, no es la situación deseable para mantener un bajo consumo de recursos. Este es el caso de Netflix, en el cual se envía el manifiesto DASH cifrado y la única manera de obtener URLs sería realizar el proceso descrito. No obstante, tras la publicación de la herramienta fast.com[4], proporcionada por Netflix para probar la velocidad de descarga de su servicio de vídeo, se ha realizado ingeniería inversa y así se ha construido un método similar para obtener estas URLs sin necesidad de ejecutar el reproductor de Netflix.

En la Fig. 3 se muestran los resultados de *throughput* en Mbps obtenidos durante varias semanas. En esta figura se puede ver la variabilidad de estos servicios así como las notables diferencias de rendimiento en media. Se puede observar una mejora clara en la velocidad de descarga de YouTube en torno al 4 de febrero. Este tipo de análisis permite averiguar cambios en la configuración del servicio, como es el caso anterior, momentos de mayor congestión a lo largo de la semana o del día, etc.

En los casos en los que ha sido posible, se han validado los resultados de estas medidas con ayuda de los reproductores o de otras herramientas.

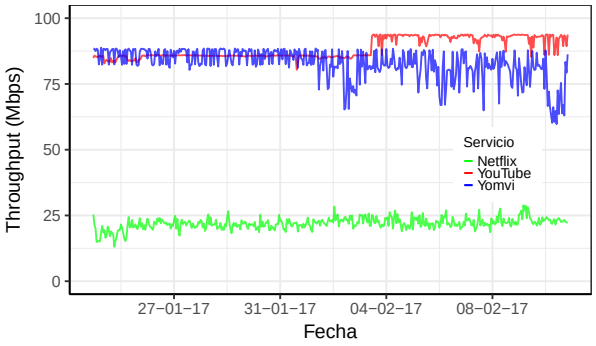


Fig. 3. Series temporales del ancho de banda (mbps) para Netflix, YouTube y Yomvi.

B. *Servicios web*

Para los servicios web, se realiza un proceso análogo al anterior. Para los servicios que proveen un API pública basada en REST, se construye un sistema de medición en torno al tiempo de respuesta de las peticiones. Esto sucede por ejemplo con Twitter[5] o Dropbox[6].

Por otro lado, para el resto de servicios se ha realizado un análisis de las peticiones utilizando de nuevo la técnica de interposición para analizar la lógica del protocolo. Para ciertos servicios como WhatsApp esta aproximación es la única posible. Se encuentran ya diferentes bibliotecas que permiten el acceso al API de WhatsApp. Sin embargo, en ocasiones se impone en los términos y condiciones del servicio que el acceso al mismo debe ser únicamente realizado a través de su cliente oficial. Esto impide, al menos desde un punto de vista legal, la medición de estos servicios salvo a través de la ejecución del cliente oficial y por tanto forzando la integración del sistema de medición en la aplicación cliente, creando en general un sobrecoste de consumo de recursos y de tiempo de desarrollo a tener en cuenta.

Para el caso de la navegación web, se ha analizado el tiempo de descarga de cualquier página. Para esta tarea, se descarga el documento principal HTML, se obtienen las dependencias de ese documento y se procede a su descarga en paralelo. Durante la descarga se contabiliza el tiempo de envío de todas la peticiones a la vez, así como el tiempo de recepción (desde que se recibe el primer fragmento de la primera dependencia hasta que se ha completado la descarga de todas las dependencias). Esto permite simular el mecanismo de carga de una página web.

Por ahora, los documentos *JavaScript* se descargan, pero no se interpretan evitando así fallos de seguridad y manteniendo el consumo de recursos al mínimo. No obstante, esto provoca que las medidas sean poco fieles para páginas con un contenido mayoritariamente dinámico. En el resto de webs, los tiempos han sido validados satisfactoriamente con varios navegadores webs. Para futuras versiones, se plantea la posibilidad de ejecutar código *JavaScript* con ayuda de tecnologías que simulan un navegador sin interfaz gráfica[7]. Otra opción es hacer

[3]https://developer.spotify.com/web-api/
[4]https://fast.com

[5]https://dev.twitter.com/rest/public
[6]https://www.dropbox.com/developers/documentation/http/overview
[7]http://phantomjs.org/

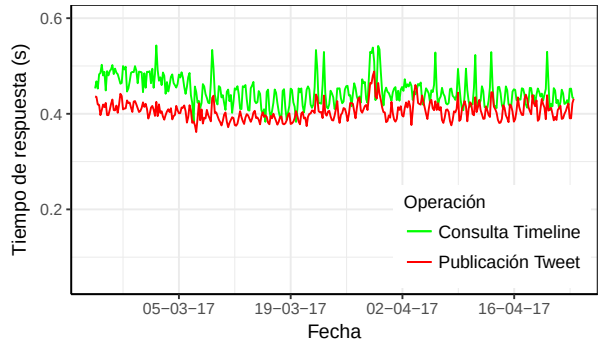


Fig. 4. Serie temporal del tiempo de publicación de un *tweet* y del tiempo de recuperación del *timeline*

que el agente ejecute un navegador completo donde se tomen estas medidas con ayuda de un complemento.

En la Fig. 4 se representan las medianas de los tiempos de respuesta del servicio Twitter. Sin entrar en aspectos de funcionamiento interno de las aplicaciones, esta métrica debería estar mucho más cercana a la calidad de experiencia percibida por el usuario que el análisis del tiempo de establecimiento de la conexión.

C. Otros servicios

Para este caso concreto se realizan medidas del correo electrónico. Igual que en los casos anteriores, se proveen algunos KPIs y métricas. Algunos ejemplos son: tiempo que se tarda en establecer una sesión por IMAP o SMTP, tiempo de envío o recepción y descarga de un correo de un tamaño dado y velocidad de descarga de un correo. Esto muestra simplemente cómo gran parte de las herramientas desarrolladas también son de utilidad en protocolos que no utilizan HTTP como medio de transporte.

De igual manera, el análisis pasivo sería una alternativa. Sin embargo, pese a poder identificar los flujos al utilizar puertos conocidos, es imposible analizar los comandos del protocolo si el tráfico va cifrado. Por tanto, es inviable cuantificar el tiempo de envío de un correo o el tiempo de consulta de la bandeja de entrada con un análisis pasivo. Por ello, las medidas activas son una opción más adecuada. En particular, este sistema proporciona la capacidad de análisis de los KPIs, así como la identificación de los flujos para su posterior análisis.

VIII. CONCLUSIONES

Las medidas activas han demostrado ser una alternativa al puro análisis pasivo del tráfico. No solo se evita el problema del análisis del tráfico cifrado y el de la caracterización y detección de los flujos generados por los servicios, sino que además se tiene control y conocimiento de ciertos parámetros del nivel de aplicación como el número de conexiones utilizadas o el tamaño del *buffer* utilizado.

Como prueba de concepto, se ha realizado una implementación de esta arquitectura así como de varios *tests* relevantes para los usuarios. Los resultados obtenidos son satisfactorios y han sido validados con diferentes herramientas específicas de cada servicio. Según se ha mostrado en la sección anterior, el sistema implementado es completamente funcional y las medidas obtenidas son relevantes para el rendimiento de los servicios.

Además, la utilidad de estas medidas va más allá del análisis del servicio en sí. Con la caracterización de las métricas y KPIs de interés y mediante el tratamiento de los datos obtenidos, se podría utilizar el sistema para buscar también la causa de los cambios de rendimiento, llegando a poder caracterizar mejoras en la infraestructura, detectando horas puntas de uso o posibles anomalías a nivel de red.

La arquitectura propuesta permite el cumplimiento de los requisitos y, además, realizar implementaciones de servicios muy heterogéneos sin necesidad de que esto implique un coste de desarrollo desmedido. También se ha realizado un análisis de los servicios que son de interés para los usuarios. A partir de ello, se han propuesto procedimientos de elaboración de futuros *tests* que permitirán actualizar el sistema conforme surjan nuevas aplicaciones y servicios.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional a través del proyecto TRÁFICA (MINECO/FEDER TEC2015-69417-C2-1-R).

REFERENCIAS

[1] W. Green, B. Lancaster, and J. Sladek, "Over-the-Top Services," in *Pipeline*, vol. 4, no. 7, 2006.

[2] K. Brunnström, S. A. Beker, K. De Moor, and A. Dooms et al, "Qualinet White Paper on Definitions of Quality of Experience," Mar. 2013, qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, March 12, 2013.

[3] G. Dimopoulos, "YouTube Traffic Monitoring and Analysis," Master's thesis, Technical University of Catalonia, 2012.

[4] M. Kihl, R. Larsson, N. Unnervik, J. Haberkamm, A. Arvidsson, and A. Aurelius, "Analysis of facebook content demand patterns," in *2014 International Conference on Smart Communications in Network Technologies (SaCoNeT)*, June 2014, pp. 1–6.

[5] I. Sodagar, "MPEG-DASH: The Standard for Multimedia Streaming Over Internet," Tech. Rep., 04 2012.

[6] R. Pantos and W. May, "HTTP Live Streaming," Internet Engineering Task Force, Internet-Draft draft-pantos-http-live-streaming-21, Mar. 2017, work in Progress.

[7] S. Kalidindi, M. J. Zekauskas, and D. G. T. Almes, "A One-way Delay Metric for IPPM," RFC 2679, Sep. 1999.

[8] G. Almes, S. Kalidindi, and M. J. Zekauskas, "A Round-trip Delay Metric for IPPM," RFC 2681, Sep. 1999.

[9] Ixia, "IxChariot," https://www.ixiacom.com/products/ixchariot.

[10] A. Cuadra-Sanchez and J. Aracil, "A novel blind traffic analysis technique for detection of WhatsApp VoIP calls," *International Journal of Network Management*, vol. 27, no. 2, 2017.

[11] J. P. Laulajainen, A. Arvidsson, T. Ojala, J. Seppanen, and M. Du, "Study of youtube demand patterns in mixed public and campus wifi network," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug 2014, pp. 635–641.

[12] D. Rivera, N. Kushik, C. Fuenzalida, A. Cavalli, and N. Yevtushenko, "QoE Evaluation Based on QoS and QoBiz Parameters Applied to an OTT Service," in *2015 IEEE International Conference on Web Services*, June 2015, pp. 607–614.

[13] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of HTTP video streaming," in *Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management, IM 2011, Dublin, Ireland, 23-27 May 2011*, 2011, pp. 485–492.

[14] "TRANSUM Wireshark Plugin: Analyzing a Website Problem," https://community.tribelab.com/mod/page/view.php?id=492.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6575

MediaDASH Tool: Plataforma Web para la Codificación, Difusión y Recepción de Videos DASH

Miguel García-Pineda, Daniel García-Costa, Jonatan Hannecke-Esteve, Santiago Felici-Castell, Jaume Segura-Garcia.
Departament d'Informàtica
Universitat de València
Av. de la Universitat, s/n. 46100 – Burjassot. Valencia. Spain.
migarpi@uv.es, daniel.garcia@uv.es, jonidhe@gmail.com, felici@uv.es, jsegura@uv.es

***Resumen*- Hoy en día el tráfico de video en Internet significa alrededor del 65-70% del tráfico total. Se prevee que en 2019 aumente hasta el 80%. Debido a esto se hace necesario optimizar y mejorar las tecnologías relacionadas con la transmisión de video para garantizar una calidad de experiencia adecuada. Es por ello que se están desarrollando nuevos protocolos de transmisión de video adaptativos basados en DASH. Debido a que su uso está aumentando en los últimos años para realizar streaming sobre HTTP. En este artículo presentamos la herramienta MediaDASH Tool, la cual es una aplicación web que permite de manera muy intuitiva poder comprimir y preparar videos para su difusión con DASH. Así como su posterior visualización a través de la misma plataforma. Esta herramienta puede ser de gran utilidad para usuarios finales que quieran disponer de su sistema DASH como también para investigadores que quieran testear sus contenidos multimedia haciendo uso de estas técnicas de streaming.**

***Palabras Clave*- DASH, MPEG-DASH, WebM-DASH, web, herramienta, codificación, streaming**

I. INTRODUCCIÓN

Según datos del último estudio [1] realizado por Cisco Systems, en el año 2020 necesitaremos más de 5 millones de años para poder visualizar el tráfico de video que circulará en un mes por las redes IP. En ese mismo año, más del 82% de todo el tráfico consumido en Internet será tráfico de video sobre IP. Además, los usuarios finales cada vez demandan mayor calidad, tanto a nivel de mayores resoluciones y se estima que para el 2020 más del 20% de video bajo demanda sea UHD (Ultra High Definition), como en el proceso de difusión del video a través de las redes IP.

Para poder satisfacer estos requerimientos de los usuarios finales en el campo de la difusión de contenido multimedia sobre redes IP, se tienen que llevar a cabo diversas tareas en diveros campos [2]. El primero de ellos es la resolución de las imágenes. Cada vez más se disponen de dispositivos con pantallas con mayor resolución y por tanto el usuario quiere disponer del contenido adecuado a esa misma resolución. Actualmente ya existen dispositivos con resoluciones de 8K UHDTV (Ultra High Definition TeleVision, 4320p), lo cual conlleva a crear contenido con estas resoluciones.

Ligado a esta característica, están los codecs de video y sus contenedores. Los codecs de video más utilizados para resoluciones inferiores a HD (High Definition, 720p) o FHD (Full High Definition, 1080p) son VP8 y H264, en cambio cuando las resoluciones aumentan los codecs que mejor se comportan son VP9 y H265 [3]. El último códec que pretende ser un éxito debido a su eficiencia y a su característica de software libre será AV1, ya que es la evolución de VP9 y existen grandes empresas, como son Cisco, Google, Intel, Microsoft, Netflix, que están apoyando el proyecto [4]. Unidos a los codecs están los contenedores y para la familia MPEG (H264 y H265) se utiliza el contenedor mp4, mientras que para la familia WebM (VP8 y VP9) se utiliza el contenedor webm.

Una vez preparado el video lo que se realiza es seleccionar los protocolos de transporte para la difusión del contenido multimedia. A día de hoy existen diversos protocolos para la difusión de contenido multimedia, pero en este articulo vamos a focalizarnos en los protocolos sobre HTTP, como son: HTTP Live Streaming (HLS), Adobe HTTP Dynamic Streaming (HDS), Microsoft Smooth Streaming (MSS) y Dynamic Adaptative Streaming over HTTP (DASH) [5]. Concretamente seleccionaremos DASH ya que es el protocolo estandarizado que mayor compatibilidad posee y además se trata del protocolo utilizado por plataformas como YouTube, Netflix, Amazon, Hulu, etc.

Dado que el uso del protocolo DASH esta aumentando en los últimos años y es uno de los protocolos más utilizados para realizar streaming sobre HTTP. En este articulo presentamos la herramienta MediaDASH Tool, la cual es una aplicación web que permite de manera muy intuitiva poder comprimir y preparar videos para su difusión con DASH, así como su posterior visualización a través de la misma plataforma. Para ello abordaremos los siguientes objetivos:

- Desarrollar una aplicación web que permita comprimir videos especificando diferentes características finales que el usuario desee que tenga el video, como resolución, tamaño, tasa binaria, etc.
- Preparar y reproducir videos en WebM, MP4, MPEG-DASH [6] y WebM-DASH [7], a través de la aplicación web.
- Llevar a cabo una evaluación de la herramienta para analizar el comportamiento de la misma haciento uso de los diversos métodos de video streaming.

Este artículo esta organizado de la siguiente manera. En la sección II trataremos el estado del arte. En primer lugar, hablaremos de los protocolos de streaming adaptativo sobre http más utilizados a día de hoy y posteriormente mostraremos plataformas para generar y difundir este tipo de contenido multimedia. En la sección III mostraremos el diseño de la aplicación MediaDASH Tool, sus diagramas de sequencia y la implementación. La implementación de la herramienta presentada en este artículo se mostrará en la sección IV y por último en la sección V se presentarán las conclusiones y los trabajos futuros que están naciendo a partir de esta aplicación.

II. Estado del Arte

En este punto vamos a analizar dos aspectos claves para este articulo. En primer lugar, analizaremos las técnicas más utilizadas para la difusión adaptativa de video sobre HTTP y en segundo lugar comentaremos diversos trabajos que existen donde se presentan otras herramientas para la difusión de este tipo de contenidos, analizando cuales los son los puntos fuertes y débiles respecto a nuestra propuesta.

A. *Protocolos de streaming adaptativo sobre HTTP*

Actualmente, existen diversos protocolos para el streaming adaptativo que ofrecen diversas mejoras sobre los protocolos de transmisión tradicionales, como RTP o RTMP, entre las que se cuentan:

- Reducción de los costes de infraestructura.
- Posibilidad de almacenamiento en memoria caché en redes CDN y otras infraestructuras de caché HTTP.
- Reducción de las amenazas procedentes de restricciones de proxy y cortafuegos.
- Optimización en tiempo real mediante heurística en el equipo cliente (tasa de bits adaptativa).
- Redundancia integrada.
- Implementación sencilla de reproductores HTML5.

HLS [8] utiliza video H.264 MPEG-2 TS segmentado y archivos descriptores M3U8 para difundir el video en directo y a la carta, con tasas de bits adaptativas. Un archivo M3U8 es un índice que permite al cliente saber qué secuencias y segmentos están disponibles en un momento dado. El dispositivo selecciona automáticamente la secuencia más adecuada desde el archivo de manifiesto primario, teniendo en cuenta las limitaciones de ancho de banda y de CPU. A continuación, descarga el segmento y lo añade al búfer de reproducción. Otra alternativa similar propuesta por Adobe es HDS [9]. Este es el protocolo para la difusión de video adaptativo compatible con Flash. Este método de streaming permite envío de video adaptativo bajo demanda y en directo, con la principal diferencia con respecto a HLS que utiliza el contenedor MP4. Por otro lado, esta MSS [10] que es un método de entrega de medios híbridos desarrollado por Microsoft. Actúa como streaming, pero se basa en la descarga progresiva en HTTP. Las descargas HTTP se realizan en una serie de pequeños fragmentos, permitiendo que los medios se almacenen en caché de forma fácil y barata a lo largo de los puntos finales de la red, más cerca de los clientes. Proporcionar varias velocidades de bits codificadas de la misma fuente de medios, también permite a los clientes cambiar de forma continua y dinámica entre las velocidades de bits dependiendo de las condiciones de la red y de la potencia de la CPU.

Finalmente, DASH es un estándar ISO [6] lanzado con la intención de unificar la metodología de transmisión adaptativa. Hasta entonces cada plataforma, Micorsoft, Apple y Adobe, generaba cada uno los segmentos y los archivos de descripción con diferentes formatos, con lo que los dispositivos y programas que quisieran reproducirlos debían implementarlos todos. Para DASH, los archivos de audio y video se llaman MP (Media Presentation) y los archivos de descripción se llaman MPD (Media Presentation Description) y están codificados en XML. Al igual que las anteriores tecnologías de transmisión adaptativa mencionadas, estas consisten en codificar y trocear el contenido en diferentes bitrate para que el cliente solicite el más conveniente en cada momento. El archivo de descripción indexa este contenido. Su funcionamiento es el siguiente. Primero el cliente DASH solicita el archivo MPD, y con él obtiene toda la información necesaria de los videos. A continuación, a través de peticiones HTTP el cliente captura los segmentos de los videos y a la vez monitoriza la red obteniendo así el stream más adecuado al estado de la red.

B. *Plataformas de video streaming adaptativo*

En el año 2015 YouTube decidió dejar de lado el reproductor web basado en Flash y pasar a un reproductor basado en HTML5. Desde ese momento también hubo un cambio en la forma de transmitir el video a través de HTTP, y se seleccionó DASH para realizar el streaming adaptativo del contenido que ofrecía YouTube

[11]. YouTube es una plataforma web cuyo objetivo es realizar la codificación, compresión y recepción de los videos que incluyen los usuarios. Estos aspectos son muy similares a la finalidad de la aplicación propuesta, pero la principal diferencia con el sistema propuesto es que el usuario no puedo modificar la codificación de los streams. En cambio, en la herramienta MediaDASH Tool desarrollada, el usuario tiene un control total sobre su contenido. Otra plataforma de video bajo demanda que está utilizando DASH es Netflix [12]. En este caso la plataforma web de Netflix sólo permite reproducir sus videos y los usuarios finales no pueden incluir su propio contenido. La última plataforma en utilizar DASH ha sido Hulu [13]. Esta plataforma era un poco reticente a cambiar su forma de realizar streaming basado en FLV sobre RTMP o HLS para dispositivos Apple, pero finalmente ha decidio cambiar a DASH para así tener una mayor compatibilidad. En esta plataforma ocurre lo mismo que con Netflix, donde el usuario es un simple consumidor de contenido multimedia.

Finalmente, hemos encontrado un trabajo [14] donde los autores presentan una plataforma para realizar streaming DASH. Esta aplicación permite la configuración de los pasos necesarios a lo largo de la comunicación multimedia extremo a extremo, desde la codificación, segmentación y almacenamiento del contenido multimedia en el lado del servidor, hasta la entrega y consumo adaptativo del streaming multimedia en el lado del cliente. Esta aplicación esta desarrollada mediante el framework Gstreamer y su programación ha sido en C. Las principales diferencias entre esta aplicación y la presentada en este articulo son: MediaDASH Tool es una herramienta web (uso masivo independiente del dispositivo y el sistema operativo) y las funcionalidades de servidor/cliente está en la misma aplicación, lo que aporta un uso más fácil e intuitivo de la aplicación por usuarios sin conocimientos sobre streaming. Nuestra herramienta puedo trabajar con codecs VP8, VP9, H64 y H265, en cambio el trabajo [14] solo utiliza H264. Por otra parte, el trabajo [14] puede emular condiciones de red para observar cambios en el streaming, aspecto que en nuestra aplicación no se ha incluido en la propia aplicación.

III. MEDIADASH TOOL

MediaDASH Tool es una aplicación web responsiva a la que se puede acceder desde cualquier PC o dispositivo móvil (Smartphone o Tablet) conectado a la red a través del protocolo de aplicación HTTP. Al acceder a la aplicación, primero se accederá a una página inicial donde el usuario se podrá registrar, iniciar sesión o ver los videos públicos disponibles. Cuando un usuario registrado inicia la sesión se pasa a la página inicial de usuario registrado donde se puede acceder a las distintas acciones que se pueden realizar, ya sea codificar o administrar videos o archivos. Las acciones que se pueden ejecutar son las siguientes:

- Cargar un video o archivo desde el dispositivo que accede a la aplicación.
- Ver la lista de archivos o videos de los que se es propietario y también de los que son públicos.
- Para la codificación se accede a un pequeño formulario donde se pueden especificar algunas de las características que se desea que tenga el video final. Se puede especificar el bitrate, la resolución, la calidad en la compresión, el tamaño de GOP, el estándar de codificación a utilizar y el formato de video de salida.
- Para la codificación en DASH también se puede especificar cuantos streams se quieren crear, y la resolución cada uno de ellos. Este proceso además de crear dichos videos creará automáticamente el archivo MPD necesario para acceder a los videos DASH.
- Se pueden reproducir archivos codificados en H.264, WebM, MPEG-DASH y WebM-DASH.

Existen una serie de restricciones para las acciones mencionadas anterioremente:

- La resolución de salida no podrá ser superior a la del archivo original.
- Se mantendrá la relación de aspecto (DAR) si la original se encuentra entre las disponibles, y en el caso que no exista en nuestra base de datos se utilizará la de 16:9 por defecto.
- Las características que se seleccionen se combinarán siempre que se pueda, para que así la codificación pueda llevarse a cabo de forma correcta. Si en una codificación, alguna de las características no es compatible, no se realizará y se le notificará al usuario.
- Los estándares de codificación, la resolución y el formato de video de salida se escogerán de las opciones que se indroducen en el formulario, pudiendo elegir una opción no especificada lo que provocará que algunas de las opciones se elijan automáticamente.

La parte del backend de la aplicación web está programada en PHP y JavaScript. Para las pruebas hemos utilizado un servidor Apache/MySQL. Para la codificación se necesitan instalar en el sistema operativo (SO) del servidor las herramientas FFmpeg [15] y MP4Box [16], las cuales se lanzarán desde la web por línea de comandos de forma transparente para el usuario, el cual especificará las características de la codificación a través de un formulario web.

A. Casos de uso

A raíz de la especificación del sistema definidos anteriormente, podemos obtener los siguientes perfiles que usarán el sistema y los casos de uso que habrá. los usuarios del sistema se han separado en tres perfiles:

- Administrador: es un usuario que podrá acceder a la totalidad de los videos y a la información de los usuarios y podrá borrar videos o usuarios si lo cree conveniente.

- Usuario no registrado: es el tipo de usuario con el que se accede a la página de inicio. Éste usuario podrá registrarse o hacer login. Es un usuario de visita de la página, que podrá ver y reproducir los videos que sean públicos.
- Usuario registrado: es el usuario que se registra para hacer uso de las funcionalidades de codificación de la página.

Debemos de diferenciar entre videos y archivos, siendo los videos archivos multimedia reproducibles, y los archivos corresponden a archivos (valga la redundancia) con extensión y4m, que son un conjunto de frames en crudo con formato video YCbCr. Los archivos una vez codificados pasan a crear un nuevo video manteniendo el archivo y4m original. En la Fig. 1 se observa la relación entre los perfiles de los usuarios y la relación de los casos de usos definidos en la aplicación.

B. Diseño de la estructura web

Para poder implementar la aplicación web vamos a presentar el diseño de la estructura de la web y el acceso a las diferentes funcionalidades basado en niveles. Como se puede observar en la Fig. 2, la estructura esta dividida en 3 niveles. En el Nivel 0, la página inicial que se verá cuando accedes a la web será la de inicio. Desde esa página, mostraremos los videos públicos y podremos acceder a la página de registro y de login. Estas dos páginas también pertenecerán al nivel 0, dado que son páginas que utilizaremos de paso al siguiente nivel.

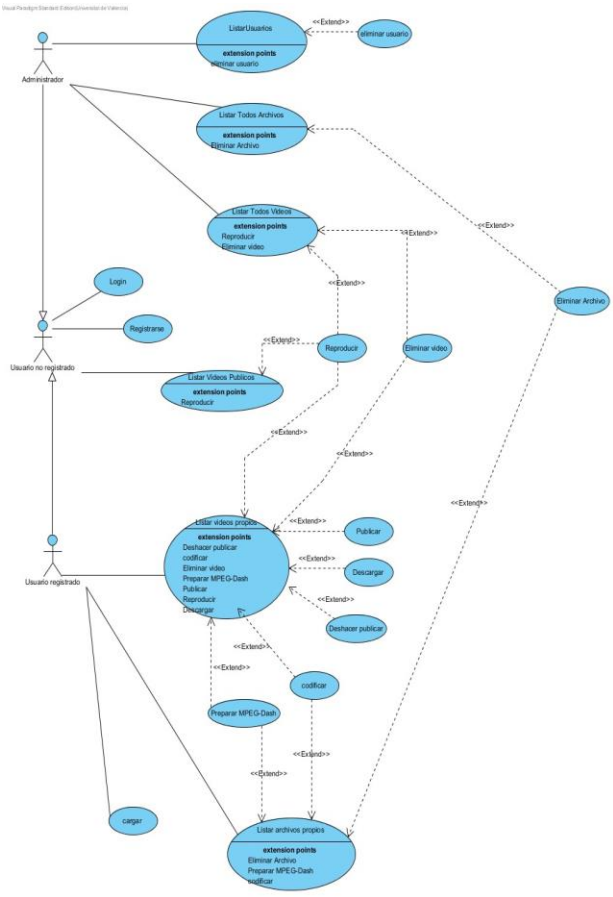


Fig. 1. Esquema de casos de uso de la aplicación MediaDASH Tool.

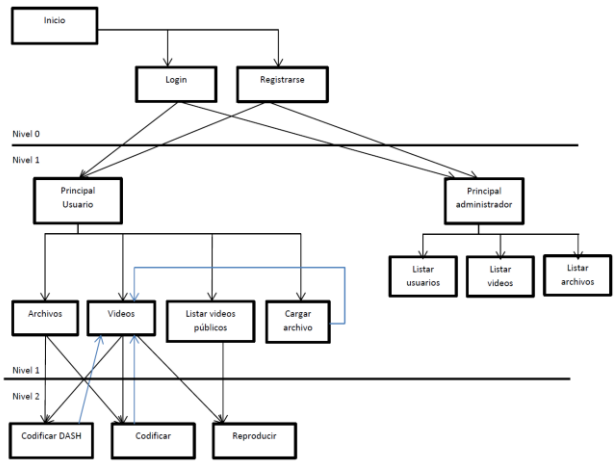


Fig. 2. Estructura de la web.

Desde el nivel 0 podremos acceder al nivel 1. Este nivel se separa en 2 partes, según el tipo de usuario con el que se haya realizado el login, administrador o usuario. En ambas partes se accederá a una página principal de presentación y de menú, desde la cual únicamente se podrá acceder a las páginas de este mismo nivel o cerrar la sesión. Si se trata de un usuario registrado, podrás acceder a las páginas de listar videos, listar archivos, cargar archivo y listar videos públicos. Por otro lado, si eres administrador podrás acceder a las páginas de listar videos, listar archivos y listar usuarios.

Por último, estarán las páginas del nivel 2. A estas páginas se podrá acceder sólo desde el nivel 1. Serán las páginas desde las cuales se lanzarán las codificaciones o las reproducciones.

C. Diseño de la estructura de datos

La estructura de los datos consta de 7 tablas implementadas en MySQL con la siguiente información. La tabla *usuarios* almacena la información de cada uno de los usuarios, como su usuario, que es con él que se realiza el login y una contraseña. También se guarda un correo electrónico que se utilizará como método de contacto en el caso que el administrador borre su cuenta o alguno de sus archivos. Para diferenciar si el usuario es administrador o no, habrá un último campo que será el de rol. Las funciones que se pueden realizar sobre esta tabla son añadir usuario, eliminar usuario, y consultar datos para hacer el login.

En la tabla *archivos subidos* se almacena la información necesaria para el manejo de los archivos y4m. En ella, guardaremos el nombre, el usuario al que pertenece, la URL relativa que hace referencia a la ubicación del archivo para poder trabajar con él y por último el formato. En ellas, se puede añadir, borrar, codificar archivos, descargar y ver la información. La codificación, la descarga y la consulta de la información no es sobre los datos, sino sobre el archivo al que hacen referencia.

En la tabla *videos* se almacena lo mismo que para archivos subidos, es decir, nombre, usuario, formato y URL, añadiéndole la característica de si es público o no. Esta última característica no se encuentra en la tabla

archivos porque estos archivos sólo pueden ser usados por sus propietarios. También se guarda el nombre de una imagen, que será el de una captura que realicemos sobre el video para poder realizar una previsualización del mismo. Sobre los videos podremos añadir y eliminar, cambiar el estado del atributo "público", reproducir, descargar, codificar y consultar información. Al igual que en *archivos*, la reproducción, la descarga, la codificación y la consulta de información se realizan sobre el archivo de video al que hacen referencia.

La tabla *MPD* guarda el archivo mdp que contiene la misma información que la tabla *videos*. Esto es porque el fichero mpd se utiliza para reproducir videos DASH, y es el archivo que se carga cuando elegimos la opción de reproducir, así que, aunque no sean archivos de video, los tratamos como tal. Por lo que guardamos el nombre, el usuario, si es público, una captura, la URL y el formato, si es WebM o MP4. Sobre la tabla *MPD* se podrá añadir, eliminar, reproducir, publicar y ver la información.

La tabla *Videos Dash* guarda información sobre los videos que se han codificado para transmitir en DASH. Se guardará el nombre, el usuario, el formato, la URL, si es público, el MPD que le hace referencia. A estos videos el usuario no podrá acceder directamente, sino solo a través de su MPD. Se puede añadir, borrar, reproducir y ver información.

La relación entre estas tablas la podemos ver en la Fig. 3. La estructura de datos utilizada es una estructura simple. La relación que hay entre ellas es que cada video, archivo, MPD y video Dash pertenece a un usuario. En la implementación de la base de datos, el atributo usuario que tiene cada uno de estas tablas será la clave ajena a la tabla *usuarios*, aplicando así la restricción de que cada archivo pertenece a un usuario obligatoriamente. También existe la relación entre la tabla *MPD* y la tabla *Videos DASH*. Esta relación es para que cada video DASH sea utilizado solamente por un mpd. Cada mpd puede referenciar a uno o más videos DASH.

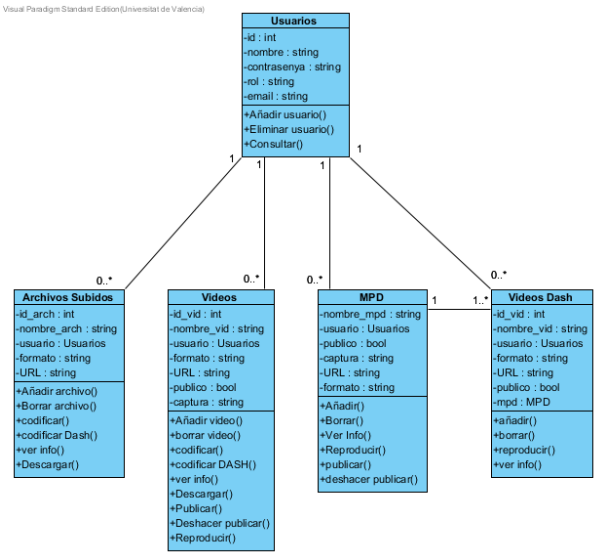


Fig. 3. Base de datos relacional utilizada en MediaDASH Tool.

A parte de estas tablas, hay dos tablas que no estarán relacionadas con ninguna. Son la tabla de *resoluciones* y de *nombres reservados*. La tabla *resoluciones* guarda una lista no modificable de resoluciones disponibles y sus respectivos DAR. La utilizaremos para proveer de una lista de resoluciones entre las que se podrá elegir para llevar a cabo una codificación. Cuando queramos hacer una codificación, buscaremos el DAR del archivo origen y buscaremos en la tabla qué resoluciones hay disponibles con ese DAR. Así podremos cambiar la resolución manteniendo el DAR original. Sobre esta tabla, sólo se podrán realizar consultas. La tabla *nombres reservados*, como su nombre indica, es una tabla donde se guardará temporalmente los nombres de archivos mientras se realiza una carga o codificación, para evitar posteriormente conflictos entre archivos con el mismo nombre.

D. Diagramas de Sequencia

Para entender cómo se implementan las codificaciones, vamos a mostrar los pasos que se siguen para llevarlas a cabo, y vamos a plasmarlos en unos diagramas de secuencia (ver Fig. 4 y Fig. 5).

En la Fig. 4 vemos una codificación de un video que no sea DASH. Para empezar el usuario, desde la lista de videos, selecciona la opción de codificar (suponemos un archivo o video ya subido al servidor). Al solicitar codificar, la aplicación envía al usuario registrado al formulario correspondiente para que rellene los campos con las características que desea que tenga el video. Cuando el usuario rellena los campos y lanza la codificación primero, se comprueba que el nombre que le quiere poner al nuevo video no existe. Si existe se le solicita que introduzca uno nuevo nombre. Si no existe, entonces la aplicación lanza FFmpeg pasándole los datos del formulario y el nombre del video de origen y éste empieza la codificación. Cuando acaba la codificación FFmpeg habrá creado un nuevo video. Este se añade a la lista de videos. Cuando se confirma la codificación, la aplicación lista los videos al usuario, incluyendo el nuevo video en la lista

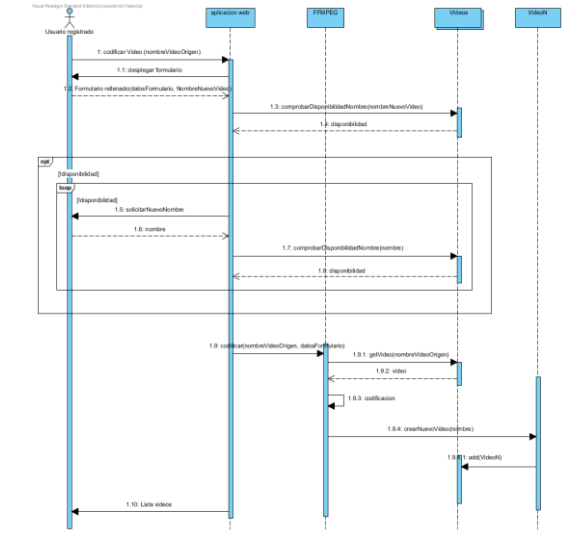


Fig. 4. Diagrama de sequencia de la codificación de un video.

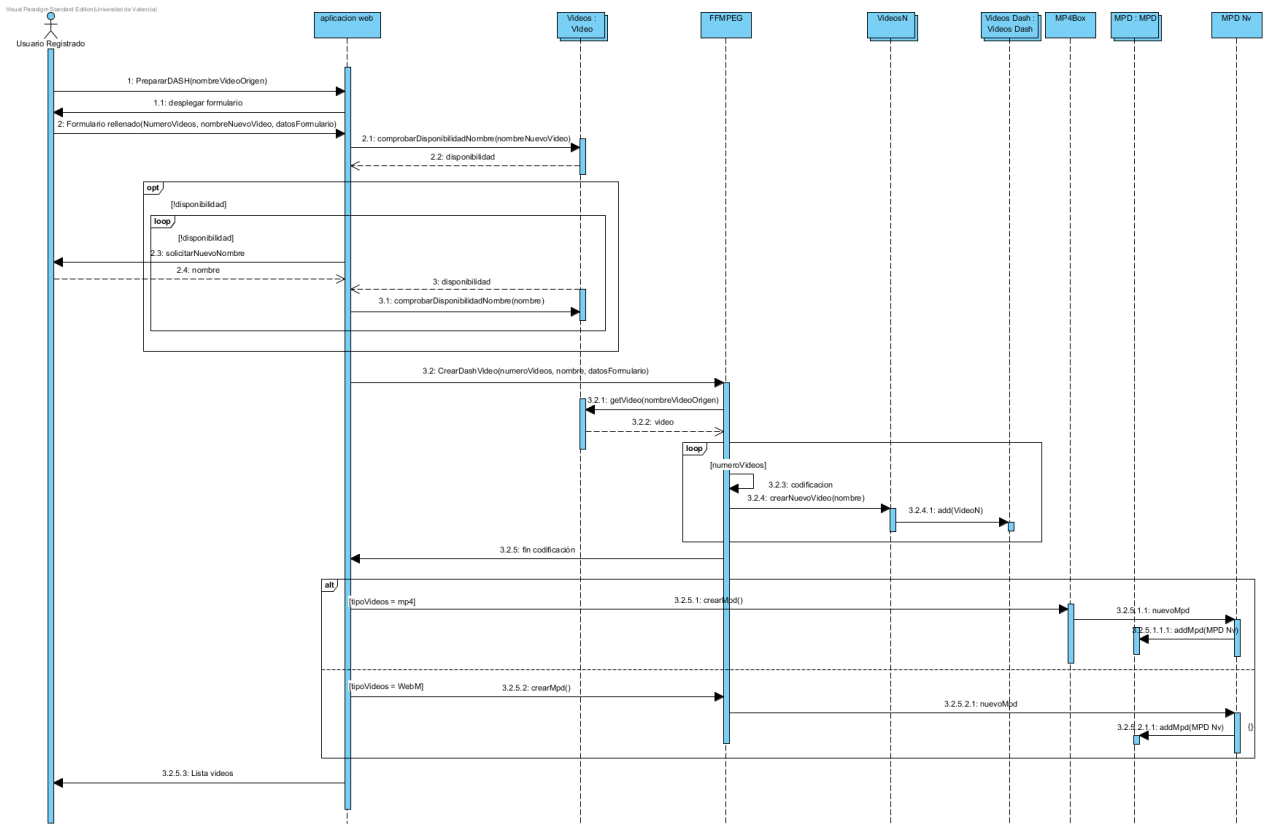


Fig. 5. Diagrama de sequencia de la codificación de un video DASH.

Cuando codificamos en DASH, se codifican varios videos y se generan los archivos MPD (ver Fig. 5). Después de que el usuario introduce los datos en el formulario y de comprobar que el nombre que desea ponerle está disponible, empieza la codificación y FFmpeg codifica varios videos. Después cuando FFmpeg confirma que ha terminado la codificació de todos los videos, entonces la aplicación lanza MP4Box para generar el archivo MPD si los videos son en formato mp4 o lanza otra vez FFmpeg si los videos son en formato WebM, para que genere el MPD correspondiente al DASH generado. Después los videos nuevos creados se añaden a la lista de videos Dash, y el MPD a la lista de MPDs.

IV. Implementación de MediaDASH Tool

La implementación web de la herramienta MediaDASH Tool se ha realizado utilizando HTML5, JavaScript, PHP y MySQL, mientras que la parte correspondiente a la codificación se ha utilizado el framework FFmpeg y MP4Box. Parte del fronted de la aplicación web MediaDASH Tool puede observarse en las Fig. 6 - 15. En la Fig. 6 puede verse la vista de la aplicación de un usuario no registrado, que podrá seleccionar el video público a reproducir y reproducirlo (veasé Fig. 7).

En la Fig. 8 vemos la aplicación tal y como la manejaría un usuario registrado. En ella se puede ver un menú a la izquierda donde se puede seleccionar los videos públicos, los videos precodificados, los videos en crudo (raw) y la acción de subir un video. Todos los videos disponen de una imagen previa para tener una visión previa del video excepto los videos o archivos que estén en fase de codificación, donde en lugar de la imagen del video aparecerá una imagen indicando que se están codificando o lo que no poseen imagen previa que se indicará que la imagen no está disponible. Además en la Fig. 9 se observan las acciones que se pueden realizar sobre un archivo (video raw) o un video, que son la codificación normal, codificación DASH, despublicar o publicar, descargar y obtener información del archivo.

En la Fig. 10 observamos el formulario para la codificación DASH de un video. En este formulario se puede seleccionar el códec, los frames por segundo, el tamaño de GOP, y para cada stream la resolución y el bitrate, además de incluir un nombre para el archivo MPD que será el que permite el streming del video DASH. En la Fig. 11 se puede observar el reproductor que dispone el usuario registrado.

En la Fig. 12 vemos el frontend del usuario administrador, en el aparecen 3 nuevos menús que son usuarios, todos los videos y todos los archivos. En el menú usuarios (ver Fig. 13) aparece la gestión de usuarios registrados en la herramienta. En el resto de menus nuevos se visualizan todos los videos o archivos y se pueden eliminar o mostrar la información (ver Fig. 14). Por último, en la Fig. 15 se observa el proceso de subido de un fichero, el cual puede realizar el usuario regisrado y el administrador.

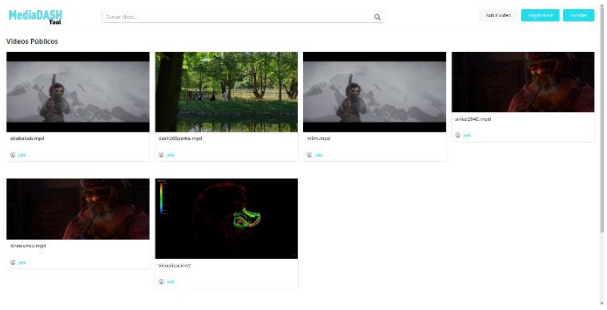


Fig. 6. Frontend con usuario público.

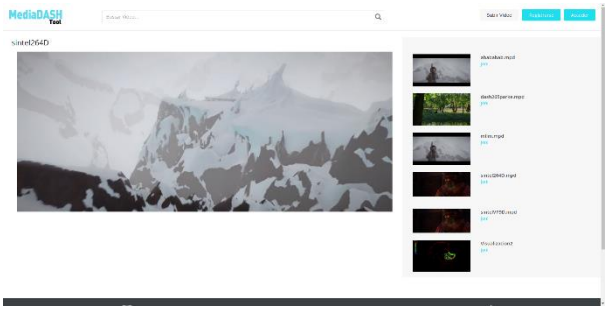


Fig. 7. Visualización de video DASH con usuario público.

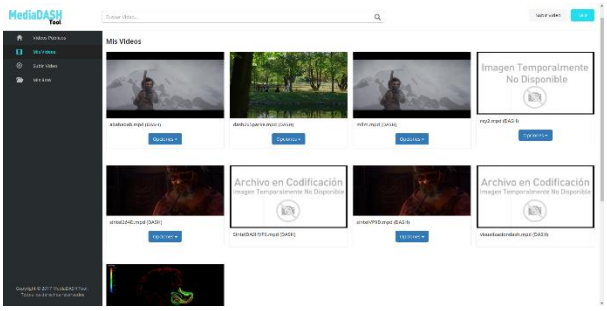


Fig. 8. Videos propiedad de un usuario registrado.

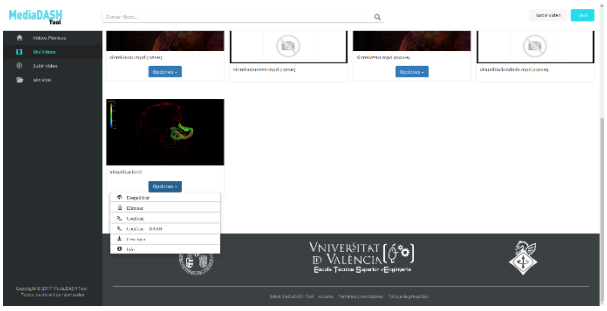


Fig. 9. Acciones aplicables a los videos de un usuario registrado.

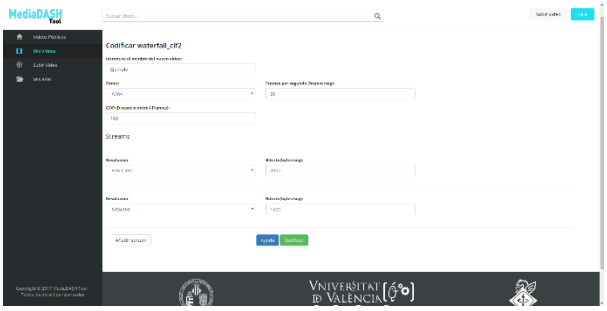


Fig. 10. Formulario para la codificación DASH.

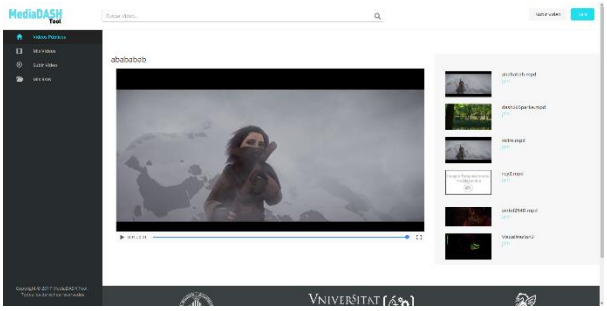


Fig. 11. Visualización de video DASH de un usuario registrado.

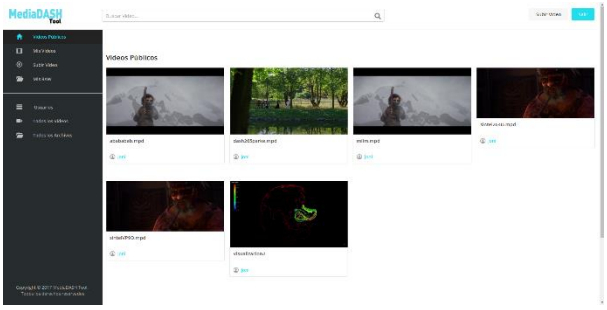


Fig. 12. Frontend para el usuario administrador.

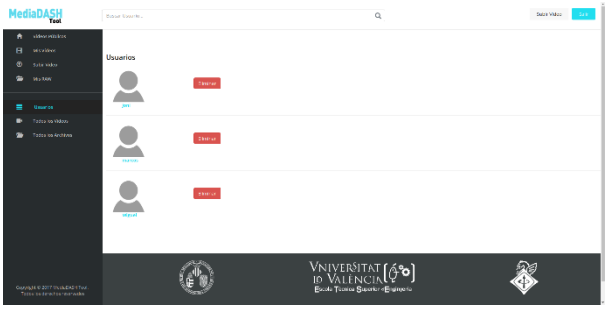


Fig. 13. Gestión de usuarios para el usuario administrador.

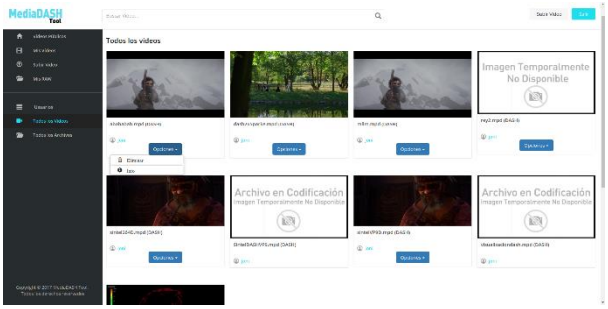


Fig. 14. Acciones que pueden realizar el usuario administrador.

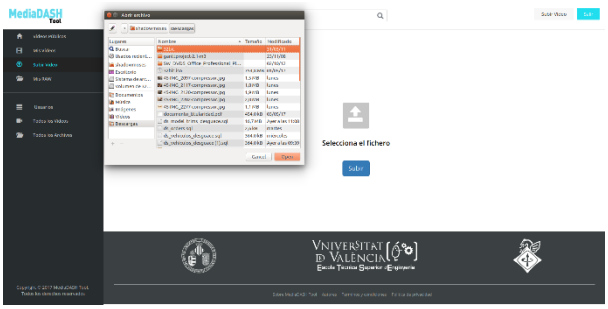


Fig. 15. Proceso de subida de archivos o videos al servidor.

V. Conclusiones y Trabajo Futuro

En este artículo hemos presentada la herramienta MediaDASH Tool, la cual es una aplicación web que permite de manera muy intuitiva poder comprimir y preparar videos para su difusión con DASH. Así como su posterior visualización a través de la misma plataforma. Esta herramienta puede ser de gran utilidad para usuarios finales que quieran disponer de su sistema DASH así como para investigadores que quieran testear sus contenidos multimedia haciendo uso de estas técnicas y servicios.

MediaDASH Tool es una herramienta web responsiva (uso masivo independiente del dispositivo y el sistema operativo) y las funcionalidades de servidor/cliente esta en la misma aplicación lo que aporta un uso más fácil e intuitivo de la aplicación por usuarios sin conocimientos sobre streaming. Nuestra herramienta puedo trabajar con codecs como VP8, VP9 encapsulados en WebM y H264/AVC, H265/HEVC encapsulados en MP4.

Como trabajo futuro estamos mejorando la aplicación para que disponga de más características como puede ser la monitorización del video recibido por el usuario final y el almacenamiento de esta información para su posible estudio posterior.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por la Universitat de València a través de los proyectos UV-INV-PRECOMP14-207134, UV-INVAE15-339582, por la Generalitat Valenciana a través del proyecto GV-2016-002 y por el Ministerio de Economía a través del proyecto BIA2016-76957-C3-1-R.

Referencias

[1] Cisco Systems, «Cisco Visual Networking Index: Forecast and Methodology, 2016–2021,» 7 Junio 2017. [En línea]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html. [Último acceso: Mayo 2017].

[2] B. Bing, Next-generation video coding and streaming, New Jersey: John Wiley & Sons, 2015.

[3] J. Bienik, M. Uhrina, M. Kuba y M. Vaculik, «Performance of H. 264, H. 265, VP8 and VP9 Compression Standards for High Resolutions,» de *19th International Conference on Network-Based Information Systems (NBiS)*, Ostrava, Czech Republic, 2016.

[4] Alliance for Open Media, «Alliance for Open Media,» 2015. [En línea]. Available: http://aomedia.org. [Último acceso: Mayo 2017].

[5] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld y P. Tran-Gia, «A survey on quality of experience of HTTP adaptive streaming,» *IEEE Communications Surveys & Tutorials*, vol. 17, nº 1, pp. 469-492, 2015.

[6] ISO, «ISO/IEC 23009-1:2014. Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats,» 2014.

[7] WebM Project, «WebM Dash Specification,» [En línea]. Available: http://wiki.webmproject.org/adaptive-streaming/webm-dash-specification. [Último acceso: 25 Julio 2017].

[8] Apple Inc., «HTTP Live Streaming,» 2016. [En línea]. Available: https://developer.apple.com/library/content/documentation/NetworkingInternet/Conceptual/StreamingMediaGuide/Introduction/Introduction.html. [Último acceso: 10 5 2017].

[9] Adobe, «HTTP Dynamic Streaming Specification,» 2013. [En línea]. Available: http://wwwimages.adobe.com/content/dam/Adobe/en/devnet/hds/pdfs/adobe-hds-specification.pdf. [Último acceso: 10 Mayo 2017].

[10] Microsoft, «Smooth streaming,» 2008. [En línea]. Available: https://www.iis.net/learn/media/on-demand-smooth-streaming/smooth-streaming-technical-overview . [Último acceso: 10 Mayo 2017].

[11] D. K. Krishnappa, D. Bhat y M. Zink, «DASHing YouTube: An analysis of using DASH in YouTube video service,» de *In IEEE 38th Conference on Local Computer Networks (LCN)*, Sydney, Australia, 2013.

[12] J. Martin, Y. Fu, N. Wourms y T. Shaw, «Characterizing Netflix bandwidth consumption,» de *n 2013 IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, USA, 2013.

[13] N. Weil , «Hulu's Move to DASH,» 2015. [En línea]. Available: http://www.streamingmediaglobal.com/Articles/ReadArticle.aspx?ArticleID=105110&PageNum=1. [Último acceso: 10 MAyo 2017].

[14] D. Gómez, F. Boronat, M. Montagud y C. Luzón, «End-to-end DASH platform including a network-based and client-based adaptive quality switching module,» de *In Proceedings of the 7th International Conference on Multimedia Systems*, Klagenfurt am Wörthersee, Austria, 2016.

[15] FFmpeg, «FFmpeg multimedia framework,» [En línea]. Available: https://ffmpeg.org. [Último acceso: 10 Mayo 2017].

[16] GPAC, «MP4Box multimedia packager,» [En línea]. Available: https://gpac.wp.imt.fr/mp4box/. [Último acceso: 10 Mayo 2017].

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6591

Mitigando Efectos Adversos de Interrupciones del Servicio de Video-vigilancia del Hogar en Clientes WiFi inalámbricos

Tatiana Gualotuña[1], Elsa Macías[2], Alvaro Suárez[2], Efraín R. Fonseca C.[1], Andrés Rivadeneira[1]

[1]Departamento de Ciencias de la Computación,
[2]Departamento de Ingeniería Telemática (DIT),
[2]Instituto de Ciencias y Tecnologías Cibernéticas (IUCTC)

[1]Universidad de las Fuerzas Armadas (ESPE)

[2]Universidad de Las Palmas de Gran Canaria (ULPGC)

[1]Av. General Rumiñahui S/N y Paseo Escénico Santa Clara. Sangolquí - Ecuador.

[2]Edificio de Electrónica y Telecomunicación – Campus universitario de Tafira – 35017 Las Palmas de G.C.

tmgualotunia@espe.edu.ec, erfonseca@espe.edu.ec, alvaro.suarez@ulpgc.es, elsa.macias@ulpgc.es, avrivadeneira@espe.edu.ec

***Resumen*- Actualmente los sistemas de videovigilancia en el Hogar se pueden combinar con sensores para formar un sistema de muy bajo coste y fácil manejo por parte del usuario final. Un componente importante en este sistema es el servidor de video streaming en tiempo real a clientes Web que usan *Wireless Fidelity*. El servidor se puede instalar en *open hardware* como el *Raspberry Pi* y se puede ayudar de sensores *arduino* para detectar alarmas de intrusión o condiciones domóticas adversas en el Hogar. Sin embargo, el *Wireless Fidelity* tiene conocidos problemas que provoca interrupciones esporádicas e impredecibles del servicio de video y de poca duración que disparan la pérdida de *frames* de video clave para observar el estado del Hogar en un momento dado, mientras el cliente se mueve (después de recibir una alarma). Mitigar los efectos adversos de estas interrupciones es una tarea complicada que hemos trabajado durante años. La novedad es que ahora construimos un sistema de *open hardware embebida* y de bajo coste (con utilidad práctica a los ciudadanos), y software libre íntegramente basado en servicios Web que es interoperable y basado en ontologías (para incluir decisiones sofisticadas, *smart*, de reconectar el servicio interrumpido). El elevado valor de parámetros de calidad de experiencia de usuario avalan los buenos resultados alcanzados.**

***Palabras Clave*- Video-vigilancia, WiFi, Streaming, patrón de diseño software, RapBerry Pi, arduino.**

I. Introducción

La cantidad de robos en domicilios es uno de los parámetros que se usan para observar la Calidad de Vida de los ciudadanos de un país. En España [1], en 2016, el número de robos con fuerza está en los 4.800 y con violencia en 1.800; cifras parecidas a las del año 2015. Así mismo en Ecuador al 12% de la población le preocupa la delincuencia e inseguridad sobre otros problemas [2]. Estos temas han hecho que los servicios de telecomunicación basados en videovigilancia en el Hogar hayan prosperado enormemente en los últimos años al amparo de la jurisprudencia de protección de datos personales [3]. Esto es, la videovigilancia es una Industria madura que ha producido tecnologías y equipamiento propietario de muy diverso tipo. Los operadores de Telecomunicación han aprovechado este auge para proponer servicios al Hogar sobre las antiguas redes de acceso a Internet de cobre (*Digital Subscriber Line, DSL*) o bien las nuevas redes de alta velocidad basadas en Fibra óptica (*Fiber to the Home*). De esta forma, cualquier usuario puede recibir una alerta en el caso que se detecte una entrada inesperada a su Hogar e incluso podría inmediatamente recibir video de lo que está pasando dentro de su Hogar en tiempo real.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

La popularidad de los sistemas de videovigilancia en el Hogar ha provocado la proliferación de sistemas de bajo coste de hardware y software libre que cualquier usuario puede desplegar (a base de tutoriales de la Web) en su propio Hogar y manejarlos a través de su conexión a Internet. El despliegue vertiginoso que ha tenido la *Internet of Things* (*IoT*) [4] ha convergido con los sistemas de videovigilancia propiciando nuevos servicios en los que además de la detección de intrusos inesperados se puede tener un control más completo (*domótico*) del Hogar observando como varían sus niveles de temperatura, humedad… El abaratamiento de los sistemas de computación y comunicación, en los últimos años ha sido espectacular. Computadores embebidos en una placa, como el *Raspberry Pi* [5] y plataformas de hardware abierto para diseño de redes de sensores como arduino [6] permiten un diseño muy barato de un sistema de videovigilancia, apoyado en sensores, de forma rápida y con un manejo muy sencillo por parte del usuario final.

Un componente importante de estos sistemas de videovigilancia es el servidor de video streaming en tiempo real. Estos servidores pueden ser desde *youtube* hasta servidores embebidos en las cámaras inalámbricas *Internet Protocol* (*IP*), o bien en los Raspberry Pi [7]. En todos estos casos el cliente puede recibir video directamente en un navegador Web.

Los sistemas de videovigilancia modernos son muy sofisticados y proponen un tratamiento del video enfocado, entre otros, a: a) el procesamiento de imágenes para obtener posiciones relativas de objetos soportando adversidades medioambientales y combinándolo con IoT para mejorar la adquisición de la información [8]. Nosotros no estamos interesados en la segmentación de objetos ni la afectación de condiciones medioambientales, sino en una recepción en tiempo real de imágenes que faciliten una primera visión por parte del usuario final que observa la intrusión a su Hogar. b) La implantación de sofisticados sistemas de compresión y codificación de objetos en un sistema de múltiples cámaras de videovigilancia para enviar únicamente información relevante al usuario final [9]. Nosotros tenemos un reducido número de cámaras en el Hogar con lo cual no tenemos necesidad de implantar un sistema de este tipo. c) Uso de computadores embebidos para alojar el sistema de video streaming conectado a una cámara de video a través de *Universal Serial Bus* (*USB*) y enviarlo a través de redes móviles. En [10] se utiliza un sistema empotrado basado en el procesador *ARM9* (de libre distribución), una tarjeta de red móvil 3G, una cámara USB que captura video utilizando el estándar H.264 y lo envía a un servidor de video. El usuario accede a él mediante un teléfono móvil Android. La construcción de un sistema empotrado como servidor y procesador de video es una opción importante debido a que se puede encontrar hardware y software de libre distribución manteniendo un esquema de comunicación de bajo costo que podría ser la solución para el Hogar. Sin embargo, en [10] no se enfocan, como nosotros en el video streaming en tiempo real. Nosotros usamos un computador embebido de propósito general (Raspberry Pi) por los buenos resultados de rendimiento que ofrece y su bajo coste.

El escenario que hemos contemplado es aquel en el que existe un Hogar controlado por sensores y una o varias cámaras de videovigilancia. Cuando los sensores detectan condiciones físicas fuera de un intervalo de normalidad entonces disparan la cámara de video que empieza a retransmitir en tiempo real a usuarios que disponen de teléfonos móviles *Wireless Fidelity* (*WiFi*). Estos clientes pueden estar en movimiento puesto que deben dirigirse hacia el Hogar. Durante ese movimiento se pueden dar situaciones en las que se interrumpa la recepción del video debido a múltiples factores revisados en [11]: comportamiento caótico de los canales radio, problemas de acceso al canal en WiFi, problemas de control incorrecto de congestión y flujo en protocolos de nivel de red y transporte en Internet para redes inalámbricas y mala gestión de la conexión de los servidores de video streaming que no detectan cuando un cliente ha perdido la conexión enviando frames de video que nunca serían recibidos por su destinatario. En este caso, un modelo simple de rendimiento demuestra que la *Quality of Service* (*QoS*) de la red se degrada enormemente, básicamente, porque por cada vez que se interrumpe el servicio, durante cierto tiempo, es necesario volver a iniciar la sesión de video streaming; pero además degrada enormemente la *Quality of Experience* (*QoE*) [12]. Nosotros en este artículo presentamos una solución novedosa a este problema que mitiga los efectos adversos de estas interrupciones de servicio en la QoE, a la vez que permite un mayor control de la seguridad en el Hogar.

Las novedades importantes de este trabajo son:

- Formulamos una solución interoperable, multiplataforma y basada íntegramente en la Web y agentes *Java Agent Development Framework* (*JADE*) [13] mediante el *addon* de JADE *Web Services Integration Gateway* (*WSIG*) [14] para transformar los comportamientos de los agentes JADE en servicios Web.
- Diseñamos ontologías para implantar comportamientos inteligentes de los agentes que controlan los sensores y las interrumpciones de servicio.
- Integramos el servidor de video en un Raspeberry Pi B+ e iniciamos la emisión de video si se producen alarmas en varios sensores que se han implantado en una plataforma arduino.
- Las medidas de la QoE demuestran el buen comportamiento y satisfacción de los usuarios entrevistados.

La estructura del artículo es la siguiente. En la Sección II revisamos brevemente un modelo sencillo de interrupciones de servicio y su efecto adverso en la QoE. En la Sección III se analiza el diseño del sistema de videovigilancia con control de las interrupciones de video streaming. En la Sección IV se presentan los resultados experimentales y los valores obtenidos para la métrica de *Mean Opinion Score* (*MOS*) para medir la QoE y finalmente presentamos algunas conclusiones y trabajo futuro.

II. LAS INTERRUPCIONES DEL SERVICIO DE VIDEO STREAMING

En la Fig. 1 se muestra un esquema de los componentes (modelo de cajas negras) principales de nuestro sistema de videovigilancia: los sensores, procesador de alarma de los sensores, el servidor de videostreaming y el cliente. El objetivo es introducir un modelo matemático sencillo que de idea de la problemática de las interrupciones del servicio y los efectos adversos que tiene en la QoE.

A grandes rasgos, el sistema de videovigilancia tiene las siguientes características:

- Los sensores están continuamente enviando datos sensados (con un periodo de muestreo determinado) al procesador de alarmas.
- La cámara de video se pone en funcionamiento una vez el procesador fusione los datos de los sensores y determine que se ha producido una alarma. En ese momento comienza la sesión de videostreaming con el cliente. A la vez, se para el procesamiento de nuevos datos sensados, hasta que se resuelva la alarma actual mediante una acción concreta del usuario (cliente).
- El servidor de video streaming recibe frames de la cámara de video y los almacena en un archivo (buffer) para posterior análisis de las grabaciones.
- Al mismo tiempo que se van almacenando los frames de video en el archivo, envía una notificación al cliente indicando la existencia de una intrusión u otra condición domótica excepcional en el Hogar.
- En ese momento el cliente inicia una sesión de videostreaming con el servidor, recibiendo, en tiempo real, los frames de video que va produciendo la cámara.
- En un momento determinado el cliente para la emisión de video streaming de la cámara a través de su aplicación móvil. En este momento el procesador de alarmas volverá a fusionar datos sensados en busca de nuevas alarmas.

Centrando ahora la atención en el envío de video streaming podemos tener dos escenarios distintos: a) desde que inició la sesión de video streaming hasta que la cancela el usuario, no se produce ninguna interrupción del servicio. b) Se producen una o varias interrupciones de video de una duración suficiente cómo para que el servidor cierre la sesión de video streaming, debiéndose iniciar de nuevo la sesión (esto implica volver a iniciar una nueva sesión y comenzar a enviar el video almacenado desde el principio). Cada vez que se produzca el cierre una sesión se debe volver a empezar la retransmisión desde el principio.

Analicemos el impacto en la QoS y la QoE con un modelo sencillo abstrayendo detalles de la comunicación y observando únicamente la cantidad de tiempo y energía necesaria para enviar todo el video almacenado y de manera informal cuál sería la satisfacción del usuario. En la Fig. 2 se muestra un esquema que facilita la comprensión del modelo. En el caso ideal (que no haya ninguna interrupción), el cliente (*C*) tarda *Ts* unidades de tiempo (*u.t.*) en iniciar la sesión de videostreaming e indicar el comienzo de la recepción de paquetes de video (*SETUP*, *OK* y *PLAY*). Después de t_p u.t. recibe el primer paquete de video. A partir de ahí recibe un nuevo paquete después de cada d_i u.t. (retraso de llegada del paquete *i* respecto al paquete *i-1*). Por tanto, el tiempo óptimo (T_{opt}) que tarda *C* en recibir todo el video es el que se muestra en la Ec. 1:

$$T_{opt} = T_s + t_p + \sum_{i=2}^{n} d_i \qquad (1)$$

Donde *n* es el número total de paquetes que se envían, T_s es el tiempo de inicialización de la sesión y t_p es el tiempo de llegada de un paquete.

En el caso que se produzcan *m* interrupciones del video entonces el tiempo (T_{int}) que tarda en recibirse el vídeo es considerablemente mayor y está en función del número de interrupciones que se producen y su duración (Ti_j), según indica la Ec. 2:

$$T_{\text{int}} = (T_s + t_p)m + \sum_{j=1}^{m} Ti_j + \sum_{i=2}^{K_j} d_i + T_{opt} \qquad (2)$$

Donde el factor K_i es la cantidad de paquetes que se recibe antes de producirse la interrupción *i* y suponemos que al menos se recibe uno antes de cualquier interrupción.

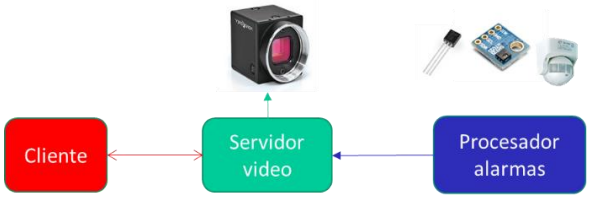


Fig. 1. Esquema del sistema de videovigilancia.

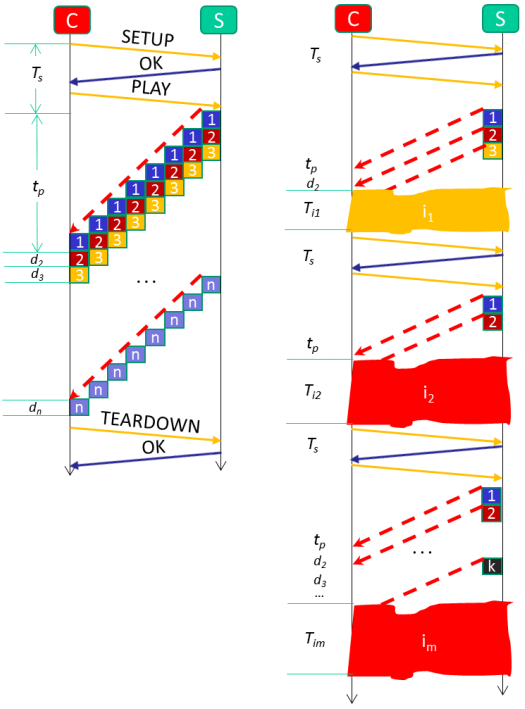


Fig. 2. Tiempos de comunicación sin y con interrupciones de servicio de video streaming.

Un efecto adverso importante es el número de veces que se retransmite un paquete de video que viene dado por la Ec. 3.

$$\sum_{i=2}^{K_i} d_i \tag{3}$$

En cada una de estas retransmisiones el paquete está ocupando el canal WiFi evitando que se pueda comunicar otros paquetes (aumentando por tanto el tiempo de contención). Además, si suponemos que el teléfono móvil del *C* tiene un gasto de *p* unidades de consumo de energía (u.e.), entonces el teléfono móvil estará haciendo un gasto (consumo extra ineficaz de batería) indicado en la Ec. 4.

$$\sum_{i=2}^{K_i} p \tag{4}$$

El efecto adverso más importante es la degradación de la QoE [15]. Es posible que el usuario acepte que se le retransmita una vez los mismos paquetes; pero no estaría dispuesto a que se lo retransmitieran más de tres veces. Y lo más probable es que decidiera no seguir reiniciando las sesiones de video streaming. Esto es importante porque en un sistema de videovigilancia como el que consideramos, precisamente un elemento distintivo es que el usuario pueda hacer una visualización del video lo más pronto posible mientras, por ejemplo, está camino a su Hogar.

Este no es un problema particular de nuestro sistema, sino que se produce en otros escenarios. Por ejemplo, en [16] se demuestra matemáticamente, que a nivel de enlace, las interrupciones cuya duración varía entre muchos segundos o pocos minutos producen *buffer underrun* y se propone mitigarlas mediante: control de *buffering*, monitorización de la red y la regulación de la inyección de paquetes de vídeo desde el Servidor. A nivel de enlace se puede tener un control de este tipo, pero cuando las interrupciones tienen cierta duración, el problema es que los servidores de videostreaming abortan la sesión de video y se debe recomenzar de nuevo como hemos indicado anterioremente. En [17] se analiza la optimización del uso de *caches* de video para lograr minimizar las interrupciones de video y más concretamente minimizar el T_s. En [18] se presenta un estudio de implantación de un sistema de videovigilancia sobre *Long Term Evolution* en el que se reconoce como una de las enseñanzas aprendidas que se debe proporcionar un sistema de *buffering* adecuado para poder soportar el videostreaming en tiempo real sin interrupciones de servicio, además de modificar el sistema de buffering de *Android*. Nosotros en [19] [20] analizamos el uso de varios buffers de video streaming y el empleo de proxies del servidor de video streaming que permitan monitotizar parámetros de la comunicación entre el cliente y el servidor para minimizar el T_s y las retransmisiones de paquetes, además de automatizar la reconexión de la sesión del videostreaming para evitar que el usuario lo deba hacer manualmente, reduciendo con ello el efecto adverso de las interrupciones de video streaming en la QoE. Nuestros esquemas pueden ser adaptados para construir la solución presentada en [18] teniendo en cuenta los mecanismos específicos de QoS.

III. Sistema de videovigilancia que mitiga las interrupciones de servicio de videostreaming

Dado que el sistema a implantar tiene una complejidad considerable, para implantar el software de la mitigación de los efectos adversos de las interrupciones de video streaming, en conjunción con el manejo de las alarmas de los sensores, se procedió a una especificación de los componentes del software basados en patrones de diseño de software [21]. En [19] se puede encontrar un análisis exhaustivo de la derivación formal de los mejores patrones para implantar cada componente. Hasta donde alcanza nuestro conocimiento esto representa una novedad importante en el diseño e implantación de sistemas de videovigilancia encaminados en la línea de sistematizar y estandarizar la parte de diseño software de servicios telemáticos, y va en la línea recomendada en [22] [23]. Como ejemplo, en la Fig. 3 se muestra los patrones utilizados para el diseño de los componentes de la Fig. 1. El patrón *Modelo Vista Controlador* (*MVC*) se ha usado para la interacción entre el cliente y el servidor

de video streaming al que se ha añadido un patrón *proxy* que interactúa con el servidor y el cliente. Para el manejo de los sensores se ha implantado el patrón *observador*. El *Modelo* maneja el buffer de video (almacenamiento temporal) que no puede ser transmitida en tiempo real porque se ha producido una interrupción. La *Vista* utiliza el patrón *Composite* que maneja la interfaz de usuario (*C*). Recibe un mensaje de alarma negocia la sesión de video y visualiza el video. Si se produce una interrupción, despliega un mensaje *reconectando* hasta que se reconecta automáticamente de nuevo que es cuando la visualización del video continúa. El *Controlador* utiliza los patrones *Proxy* y *Strategy*. El *proxy* solicita al *Observer* del servidor de procesado de alarmas (sistema externo) el estado de la información de los sensores procesada por él. Con ella genera alarmas y solicita que la cámara se ponga en funcionamiento. Cuando ocurre una interrupción de video invoca al patrón *Strategy* para que ejecute el algoritmo de tratamiento de la interrupción. Cuando se reestablece la sesión de video le solicita la ejecución del procedimiento de restablecimiento de la sesión. El *Observador* del sistema externo utiliza el patrón *Adapter* para transformar en servicios Web (JADE-WSIG) las acciones de los agentes inteligentes JADE. Estos agentes manejan ontologías para tratar de forma estándar los datos provenientes de los sensores. Estas ontologías facilitan el diseño e implantación de la inteligencia de los agentes para disparar alarmas. Aunque no se muestra en la Fig. 3, el video streaming *Proxy* también maneja ontologías de batería, posición y nivel de cobertura de los teléfonos móviles.

En [19] se presentan todos los diagramas de secuencia entre los objetos que implantan los componentes del sistema atendiendo a los patrones de diseño utilizados.

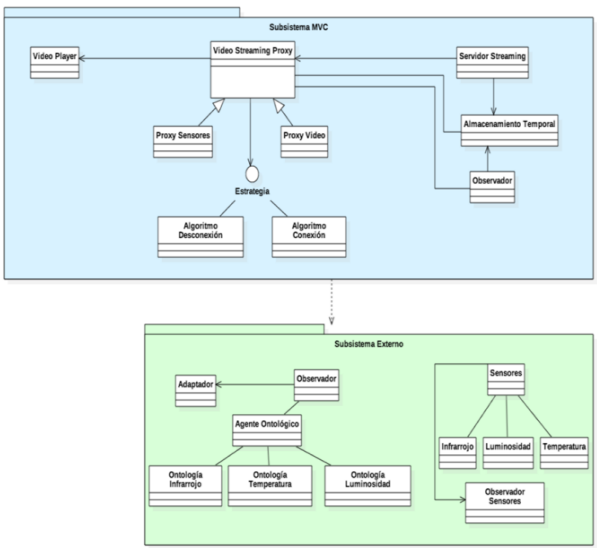


Fig. 3. Ejemplo de especificación de los patrones de diseño para el servidor de video, el cliente y el servidor de alarmas de los sensores.

A. Modelado de la solución a las interrupciones de video streaming

Con los mecanismos provistos en el sistema de videovigilancia, se mitigan los efectos adversos de las interrupciones de video streaming. En la Fig. 4 se muestra un esquema del tiempo que se invierte en el sistema de video streaming con el nuevo mecanismo. Se puede observar que ahora todos los paquetes de video se almacenan en el buffer la primera vez que se envían. Cuando se produce la i_1, los paquetes se siguen almacenando hasta completar todos los paquetes de video. Para evitar que el *C* envíe una orden de parar el video streaming se opta por almacenar el video durante una cantidad de tiempo prefijada pero configurable (entre 4 y 5 minutos, porque esta es la duración estimada que duran los atracos en el Hogar). En ese tiempo se almacenarían los *n* paquetes del video recogido por la cámara una vez el servidor de alarmas produce una nueva alarma.

Con el nuevo sistema el tiempo que se tarda en enviar el video completo está dado por la Ec. 5.

$$T_{buf} = T_s + t_p m + \sum_{j=1}^{m} Ti_j + \sum_{i=2}^{n-m} d_i \qquad (5)$$

Con lo cual, a partir de las Ec. 1, 2 y 5 se puede obsevar la relación de la Ec. 6.

$$T_{opt} < T_{buf} << T_{\text{int}} \qquad (6)$$

Lo que indica que el usuario ahorraría una cantidad de tiempo considerable en ver el vídeo en presencia de interrupciones en relación a si no se aplica nuestro método. Pero además, se le proporciona mayor comodidad al hacer que la reconexión se pueda manejar de forma automática. Esto indudablemente mejora la QoE.

Otro aspecto importante es el consumo de batería. Con este mecanismo se consigue minimizar el consumo extra de energía en retransmisiones. De forma indirecta esto también mejora la QoE debido a que el usuario dispone de mayor tiempo para estar conectado al sistema de videovigilancia. Por otro lado, es importante maximizar el tiempo de vida de la batería para asegurar una mayor tranquilidad del usuario de tal manera que no perciba que se puede desconectar del sistema y no saber que está pasado en en su Hogar.

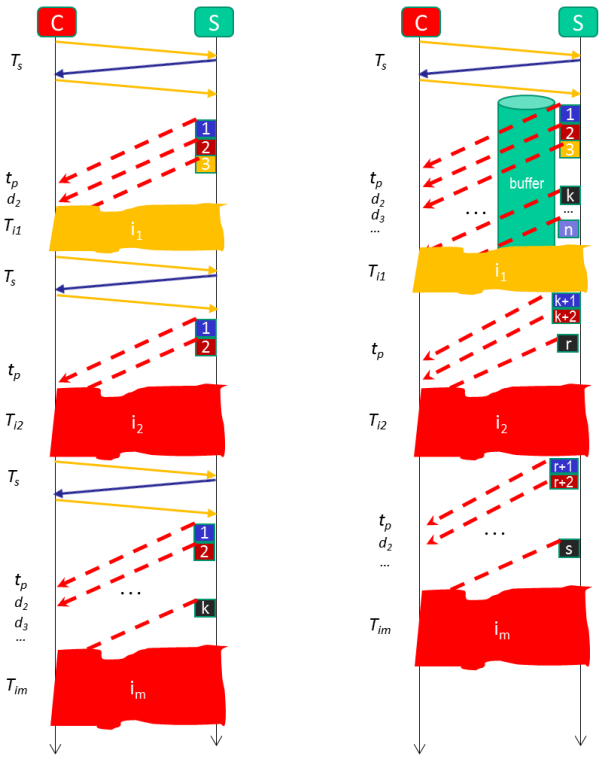


Fig. 4. Tiempos de comunicación con interrupciones de servicio de video streaming y con el nuevo método de mitigación de los efectos adversos de las interrupciones de servicio.

IV. Construcción del prototipo, implantación del software y resultados experimentales

A partir de la especificación basada en patrones de diseño software y ontologías se construyó un prototipo hardware y toda la programación del sistema completo: servidor de videostreaming, clientes inalámbricos y procesador de los datos sensados.

A. *Prototipo hardware*

Se construyó una maqueta física que emula un Hogar típico con varias habitaciones, garaje, escaleras… (Fig. 5). Dentro de esa maqueta se instalaron los sensores y la cámara de video para detectar alarmas de intrusión no deseadas y la grabación de las acciones correspondientes. En ella se instalaron 2 sensores de luz en la parte delantera y trasera del Hogar, 2 sensores de temperatura en la sala y cocina, 4 sensores infrarrojos para detección de movimiento (entrada delantera, trasera, cocina, baño) y 2 actuadores para enfriamiento (ventiladores). Las características técnicas de los sensores utilizados son:

- *Luz*: dimensiones 65x11x13 mm, serie Fotoresistor-BH1750, mediciones 1-65535 LX, muestreo 2s.
- *Temperatura digital* (*Ds28b20*): cada dispositivo tiene un código de serie 64-bit único almacenado en su *Read Only Memory* (*ROM*), capacidad multipunto que simplifica el diseño de las aplicaciones de detección de temperatura, puede ser alimentado desde la línea de datos. El rango de suministro de energía es de 3.0 V a 5.5 V. Mide temperaturas de -55 °C a + 125 °C (-67 °F a + 257 °F) ± 0,5 °C con precisión de -10 °C a + 85 °C. La resolución del termómetro es seleccionable por el usuario de 9 a 12 b y convierte la temperatura en códigos de 12 b en 750 ms (máximo).
- *Barrera Infrarroja*: fototransistor de 10.2 x 5.8 x 7 mm, distancia de operación pico: 2.5 mm, intervalo de operación para una variación de corriente de colector mayor al 20 %: 0.2mm a 15 mm, corriente de salida típica bajo prueba: Ic = 1 mA, filtro de bloqueo de luz ambiental, longitud de onda del emisor: 950 nm.

El procesado de los datos sensados se hizo mediante un microcontrolador arduino Atmega 328. Se utilizó una cámara de 20.7 Mpixels, flash LED, con sensor de ½.3”, geotagging, y estabilizador de imagen. Para implantar el servidor de videostreaming y otro software de control de almacenamiento… se usó una RaspBerry Pi B+ con un procesador de 900 MHz de cuatro núcleos de CPU ARM Cortex-A7, 500 GB de RAM, 4 puertos USB, 40 pines GPIO, interfaz de la cámara (CSI), interfaz de pantalla (DSI), ranura para tarjeta Micro SD y núcleo de gráficos VideoCore IV 3D.

B. *Implantación del software basado en componentes de software libre*

En la Fig. 6 se muestra un diagrama de los distintos componentes de software libre que se utilizaron para la implantación de los distintos componentes del sistema de videovigilancia.

Mediante *Raspduino* se hace interoperar a la plataforma arduino de sensores con la RaspBerry Pi B+ logrando que se pueda recoger datos de: el identificador del sensor, la medida que sensa y un sello de tiempo para ese valor sensado. Estos valores se envían a un *proxy* capaz de interactuar con agentes de JADE. En estos agentes se define el comportamiento determinado para decidir si existen condiciones no normales de los valores de los sensores en conjunto. Dado que JADE utiliza originalmente una ontología propia para el paso de mensajes, se decidió hacer esta comunicación de forma interoperable a través de servicios Web. Por eso se instaló el addon WISG que transforma esos mensajes en servicios Web. A través de *Web Services Description Language* (*WSDL*) y *Simple Object Access Protocol* (*SOAP*) se pasan los datos al servidor de aplicaciones Web *Tomcat*. Otro motivo por el que usamos *Tomcat* es que estamos interesados en utilizar un navegador Web compatible con la recepción de video en *Hiper Text Markup Language* (*HMTL*) 5.

Si se genera una alarma se dispara la *PiCam* para registrar lo que ocurre en el Hogar. Esta cámara se conecta al servidor de video Streaming propio de la RaspBerry Pi B+ llamado *RPiCam* que no registra audio de forma nativa pero si video en formato *Motion Joint Photographic Expert Group* (*MJPEG*) versión 4. El cual se comunica en formato H.264 para que se pueda recibir en HTML5 en el navegador Web (que puede ser cualquiera que acepte video en HTML5: prácticamene cualquier navegador actual).

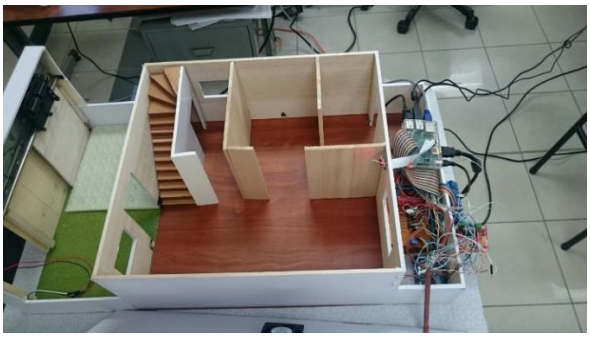


Fig. 5. Fotografía de la maqueta construida para emular situaciones reales de intrusiones no deseadas en un Hogar.

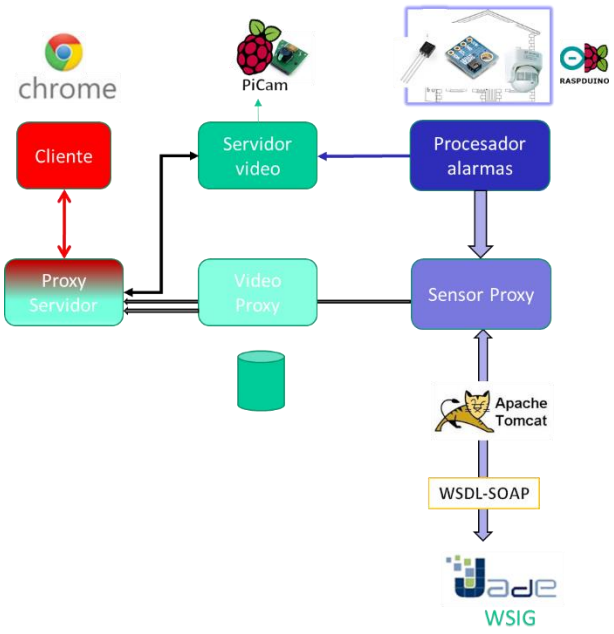


Fig. 6. Esquema del uso del software libre para la implantación de los componentes del sistema de videovigilancia.

C. Resultados experimentales de la mitigación de interrupciones de video streaming y la QoE

Para verificar el funcionamiento del sistema de videovigilancia se hicieron múltiples pruebas. En ellas dos aspectos importantes a observar es el comportamiento ante diferentes tipos de interrupciones de video streaming y la medida subjetiva de la QoE.

Se hicieron pruebas de tiempo de ejecución en presencia de interrupciones del video streaming que se provocaban moviendo un teléfono móvil de alta gama fuera y dentro del área de cobertura de un punto de acceso WiFi, al objeto de medir el tiempo total que se tardaba en emitir un video de 300 s. En los casos en los que las interrupciones duraban poco tiempo y eran sólo 3 interrupciones, era aceptable continuar con la visualización del video. Pero hubo casos en los que cerca del 50% del tiempo de visualización se invirtió en interrupciones, con lo cual era inviable estar cerca 400 s para ver el video entero.

Para medir la QoE se hicieron pruebas de MOS. Se determinó el tamaño de la muestra en consonancia con [24] (5% de error, con un nivel de confianza del 95% y una heterogenidad del 50%), obteniendo un total de 25 personas. Para la escala de opinión se utilizó la norma *International Telecommunication Union* (*ITU*) *P.800.1*, y se hicieron las siguientes preguntas (usando un formulario de *Google Forms*): *¿cómo calificaría los siguientes parámetros de la aplicación*? (*diseño de la interfaz, usabilidad, interactividad*) *¿tuvo problemas en el acceso a la aplicación*? *¿la aplicación tuvo algún error de ejecución*? *¿la aplicación reconectó el vídeo sin recargar la página*? *¿la aplicación envió las alertas de seguridad*? Las clases de respuestas para la primera pregunta fueron: *Excelente*, *Muy bueno*, *Bueno*, *Mejorable*, *Muy mejorable*. Y para el resto de preguntas *Si* y *No*. En la Fig. 7 se muestran las respuestas para la primera pregunta destacando el elevado grado de satisfacción del usuario. Las dos últimas preguntas tuviero un 100% de respuestas *Si*. El resto de preguntas tuvo un valor superior al 90% de respuestas *No*.

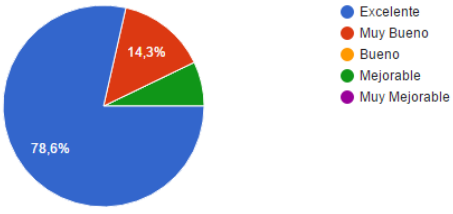


Fig. 7. Respuestas proporcionadas para la pregunta ¿Cómo calificaría los siguientes parámetros de la aplicación? (diseño dela interfaz)

V. CONCLUSIONES

Los sistemas de videovigilancia del Hogar son muy comunes a día de hoy. Con el auge de la IoT y los computadores embebidos en una placa de muy bajo coste económico, y el software libre es posible implantar sistemas de videovigilancia de muy bajo coste capaces de mitigar los efectos adversos de las interrumpciones de video en el cliente. Nosotros mostramos que es posible modelar el software con patrones de diseño y ontologías para el diseño de la inteligencia del sistema de control de alarmas. También mostramos que a través de figuras de medida como el MOS se satisface ampliamente la QoE del usuario en un porcentaje muy elevado.

Como trabajo futuro planteamos el diseño de un sistema proactivo de estimación inligente de las interrupciones de video que permita adelantar información al usuario de las interrupciones que podría experimentar.

AGRADECIMIENTOS

This work has been funded by the Spanish Ministry of Economy and Competitiveness/FEDER under project TEC2015-67387- C4-4-R. Agradecer a las personas que se ofrecieron a hacer las encuestas para el MOS.

REFERENCIAS

[1] Ministerio del Interior de España, *Infracciones Penales Registradas En Ccaas, Provincias, Islas, Capitales y Localidades con Población Superior A 50.000 Habitantes*. Disponible en: http://www.interior.gob.es/documents/10180/5791067/informe+balance+2016_ENE_MARZO.pdf/2a4b89ea-6ddc-448f-870a-428412cbf691.

[2] CEDATOS, *Los problemas que más preocupan a los ecuatorianos después del terremoto*. Disponible en: https://www.cedatos.com.ec/detalles_noticia.php?Id=259.

[3] Agencia española de protección de datos, *Videovigilancia*. Legislacción disponible en: http://www.agpd.es/portalwebAGPD/canaldocumentacion/informes_juridicos/videovigilancia/index-ides-idphp.php.

[4] Kumar Mandula; Ramu Parupalli; CH.A.S. Murty; E. Magesh; Rutul Lunagariya, *Mobile based home automation using Internet of Things (IoT)*, 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp: 340-343, 2015.

[5] Sagar R N, Sharmila S P, Suma B V, *Smart Home Intruder Detection System*, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 4, April 2017, ISSN: 2278 – 1323.

[6] Salamone, F.; Belussi, L.; Danza, L.; Galanos, T.; Ghellere, M.; Meroni, I. Design and Development of a Nearable Wireless System to Control Indoor Air Quality and Indoor Lighting Quality. Sensors 2017, 17, 1021.

[7] Sanjana Prasad, P.Mahalakshmi, A.John Clement Sunder,R.Swathi, *Smart Surveillance Monitoring System Using Raspberry PI and PIR Sensor*, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (6), pp: 7107-7109, 2014, 7107-7109, ISSN: 0975-9646.

[8] Dmitry Stepanov, Igor Tishchenko, *The Concept of Video Surveillance System Based on the Principles of Stereo Vision*, 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), 2016, ISSN: 2305-7254.

[9] Amal Ben Hamida, Mohamed Koubaa, Henri Nicolas, Chokri Ben Amar, *Video surveillance system based on a scalable application-oriented architecture*, Springer Verlag Multimedia Tools and Applications, Volume 75, Issue 24, pp 17187–17213, December 2016.

[10] Wang Zai-Ying, Chen Liu, *Design of Mobile Phone Video Surveillance System for Home Security Based on Embedded System*, 27th Chinese Control and Decision Conference (CCDC), pp: 5856-5859, 2015.

[11] Elsa Mª Macías and Alvaro Suarez, *Video Streaming based Services over 4G Networks: Challenges and Solutions*, Book Chapter: Fourth-Generation Wireless Networks: Applications and Innovations, Adibi, Sasan, Amin Mobasher and Mostafa Tofighbakhsh (editores), Chapter 22, pp: 494 - 523. Estados Unidos de América: IGI Global, 2010. ISBN 978-1-61520-674-2.

[12] Elsa Macías; Alvaro Suárez; F. Espino, *Multi-platform video streaming implementation on mobile terminals*. Alvaro Suarez and Elsa Macias Lopez (editores). "Multimedia Services and Streaming for Mobile Devices: Challenges and Innovations." 1-350 (2012), accessed December 18, 2012, doi:10.4018/978-1-61350-144-3, Capítulo 14, pp. 288 – 314, Estados Unidos de América: IGI Global, 2012, ISBN: 978-1-61350-144-3.

[13] F. Bellifemine, A. Poggi, and G. Rimassa, "JADE–A FIPA-compliant agent framework," in *Proceedings of PAAM*, 1999, vol. 99, no. 97–108, p. 33.

[14] D. Greenwood, P. Buhler, and A. Reitbauer, "Web service discovery and composition using the web service integration gateway," in *e-Technology, e-Commerce and e-Service, 2005. EEE'05. Proceedings. The 2005 IEEE International Conference on*, 2005, pp. 789–790.

[15] A. Aloman, A. I. Ispas, P. Ciotirnae, R. Sanchez-Iborra, and M. D. Cano, *Performance evaluation of video streaming using MPEG DASH, RTSP, and RTMP in mobile networks*, in 2015 8th IFIP Wireless and Mobile Networking Conference (WMNC), Oct 2015, pp. 144–151.

[16] Laksono, I., *Achieve End-To-End Qos for Wireless Video Streaming, 2004*. Disponible en: http://www.eetimes.com/document.asp?doc_id=1272006.

[17] Navin Sharma, Dilip Kumar Krishnappa, David Irwin, Michael Zink, and Prashant Shenoy, *GreenCache: Augmenting Off-the-Grid Cellular Towers with Multimedia Caches*, Proceedings of the 4th ACM Multimedia Systems Conference, 271-280, 2013.

[18] Mohammad Abu-Lebdeh, Fatna Belqasmi, and Roch Glitho, *An Architecture for QoS-Enabled Mobile Video Surveillance Applications in a 4G EPC and M2M Environment*, IEEE Access, Volume 4, pp: 4082-4093, August 2016, Digital Object Identifier 10.1109/ACCESS.2016.2592919.

[19] Tatiana Gualotuña, *Diseño de una Platalforma de Agentes para Control de Servicios de Video Streaming Móvil*, PhD Thesis, Supervisors: Alvaro Suárez y Elsa Macías, Universidad de Las Palmas de Gran Canaria (ULPGC), España, Febrero 2016.

[20] Andrés Rivandeneira, *Video Vigilancia Autonoma mediante Sistemas Empotrados-Hardware Libre*, Master Thesis, Supervisora: Tatiana Gualotuña, Universidad de Las Fuerzas Armadas (ESPE), Ecuador, 2016.

[21] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, *Design patterns: elements of reusable object-oriented software*, Addison-Wesley, 1995, ISBN: 0-201-63361-2.

[22] Brian Henderson-Sellers, Cesar Gonzalez-Perez, Tom McBride, Graham Low, *An ontology for ISO software engineering standards: 1) Creating the infrastructure*, Elsevier Computer Standards & Interfaces, 36 (2014) 563–576.

[23] C. Gonzalez-Perez, B. Henderson-Sellers, T. McBride, G.C. Low, X. Larrucea, *An Ontology for ISO software engineering standards: 2) Proof of concept and application*, Computer Standards & Interfaces 48 (2016) 112–123.

[24] Takuto Kimura, Masahiro Yokota, Arifumi Matsumoto, Kei Takeshita, Taichi Kawano, Kazumichi Sato, Hiroshi Yamamoto, Takanori Hayashi, Kohei Shiomoto, and Kenichi Miyazaki, *QUVE: QoE Maximizing Framework for Video-Streaming*, IEEE Journal of Selected Topics in Signal Processing, Vol. 11, No. 1, February 2017.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6492

Improving the energy efficiency of VoIP applications in IEEE 802.11 networks through control of the packetization period

Rafael Estepa, Antonio Estepa and German Madinabeitia.
Departmento de Ingeniería Telemática
Universidad de Sevilla
C/ Camino de los descubrimientos s/n. 41092 Sevilla.
{rafaestepa,aestepa,german}@us.es
Mark Davis
School of Electronic and Communications Engineering.
Dublin Institute of Technology.
Kevin Street, Dublin. Ireland.
mark.davis@dit.ie

***Abstract*—This paper presents an adaptive algorithm that improves the energy efficiency of VoIP applications over IEEE 802.11 networks. The algorithm seeks to achieve the largest energy savings subject to reaching a minimum speech quality under the prevailing network conditions. The control mechanism used is the dynamic selection of the packet size during the communication.**

This algorithm has been implemented in an experimental testbed and the results demonstrate that our packetization period control algorithm can provide energy savings in uncongested IEEE 802.11 networks (up to 30%). Furthermore, under poor network conditions the algorithm can prolong the duration of the call before it is dropped at the expense of a higher energy consumption.

***Palabras Clave*—Energy efficiency, VoIP, IEEE 802.11 jitel, telemática**

I. INTRODUCTION

Increasingly powerful smartphones and wider deployment of free Wi-Fi coverage have boosted the use of VoIP applications over IEEE 802.11 networks (VoWiFi). Preserving VoWiFi quality of service (QoS), along with saving battery in the portable device are two research fields with abundant literature predominantly focused on the link layer [1], [2], [3].

VoIP is a real-time application and as such, small end-to-end delays (typically less than 300 ms) and low packet loss (typically between 1 and 5% depending on the codec) must be maintained during communication in order to ensure acceptable speech quality [4]. The challenge of preserving quality in VoIP over Wi-Fi has been extensively addressed in literature. Research efforts to support quality in VoWiFi have been focused on two main approaches: (a) dimensioning works aimed at finding the maximum number of simultaneous VoIP flows that IEEE 802.11 networks can accommodate while satisfying QoS constraints (e.g. network delay, packet loss ratio) [5], [6], [7], [8], [1], [2]; and (b) link-layer proposals aimed at meeting QoS constraints in the IEEE 802.11 network by finding optimum values for MAC-layer variables such as contention window size, maximum retry limits, etc. [9], [10], [11], [12]. As shown in a recent survey [13], VoIP software can also deal with QoS by dynamically handling application-layer variables such as codec choice, or number of speech frames encapsulated into an IP datagram (N_f), which depends on the packetization period and the codec inter-frame period.

Since VoIP software frequently runs on battery-powered terminals, an emergent research topic concerned with VoIP energy efficiency in IEEE 802.11 networks has been developed over recent years. A number of proposals are based on minimizing the time spent in active states (i.e.TX,RX) through MAC-layer based strategies [3]. However, as found in [14], [3], a VoWiFi device typically spends less than 2% of its time in these active states, which justifies that most research efforts are focused on maximizing the sleep time during idle periods whilst avoiding quality impairments through fine tuning of the Power Saving Mode (PSM) parameters. In [12] the authors propose a sleep strategy that dynamically adjusts the sleep time and packetization interval according to the collision probability to achieve a trade-off between energy saving and VoIP

capacity in ideals IEEE 802.11 channels; in [15] the authors schedule sleep and wake-up intervals to save energy based on end-to-end network delay and packet loss; in [16] an adaptive U-APSD (unscheduled automatic power save delivery) is proposed to achieve a certain delay constrain for each access category in IEEE 802.11 PSM.

Although some effort has been made to achieve energy efficiency subject to a minimum QoS level through setting IEEE 802.11 link-layer parameters (e.g. delay in [17]), to the best of our knowledge no attempts have been made to simultaneously deal with energy efficiency and QoS in VoIP by exerting control of application-layer parameters (i.e. codec and/or packetization period). This work aims to be a first approach to address the problem of saving energy associated to the execution of VoIP software subject to QoS restrictions by controlling the packetization period of VoIP during the communication. We believe that our solution is compatible with and complementary to other approaches such as dynamic codec setting [18] or optimization of IEEE 802.11 MAC parameters [17].

The remainder of the paper is as follows. Section II overviews the dependencies between QoS and energy efficiency in VoWiFi. Section III states the problem addressed and the scope of our study. Section IV details the proposed control algorithm. Section V addresses the test-bed used and implementation aspects. Results are presented in Section VI. Finally, Section VII concludes the paper.

II. Relationship between packetization period, Quality and Energy Efficiency in VoWiFi

The foundation of our proposal is based on the interdependence of QoS and energy efficiency in VoWiFi. Thus, it is worth analyzing the relationships between the number of speech frames encapsulated on each packet (i.e. *packetization period* from now on), VoWiFi quality and IEEE 802.11 energy efficiency before stating our problem in more detail.

QoS in VoIP can be assessed during conversation time through the widely accepted E-Model [13]. The outcome of such model is a score from 0 to 100 termed the *R* factor which, in its simplest form, can be expressed as:

$$R = 92.8 - I_d(codec, delay) - I_{e,eff}(codec, loss, PLB) \quad (1)$$

where the factor I_d accounts for the effect of delay, and $I_{e,eff}$ is associated with codec compression, packet loss rate and packet loss behavior (PLB)(i.e. burst ratio as defined in[13]). The delay factor can be broken down into network delay, and terminal delay (e.g. codec look-ahead, packetization period or jitter buffer). The packet lost impairment is attributable to lost packets in the network, as well as discarded packets due to delays greater than the jitter buffer of the receiver.

Figure 1 illustrates basic relationships between the packetization period (N_f) and QoS (R factor) in VoIP. High values of N_f reduce the packet generation frequency (i.e. transmission opportunities in the IEEE 802.11 network) and hence the traffic load. However, it also increases the overall delay (due to the frame generation period at the terminal) and, as packet size is increased, the probability of frame error increases. Large packets also have a negative effect in IEEE 802.11 networks if low Signal to Noise Ratio (SNR) or hidden node problems are present. Therefore, we can conclude that when the Wi-Fi network operates under poor performance conditions (due to MAC layer congestion or transmission problems - i.e. low SNR or hidden node problem), there is no clear predictable outcome from increasing or decreasing N_f as high values can help reduce the MAC traffic load, but may result in lower performance. Observe that both impairments (congestion and transmission problems) can often exhibit high temporal variability.

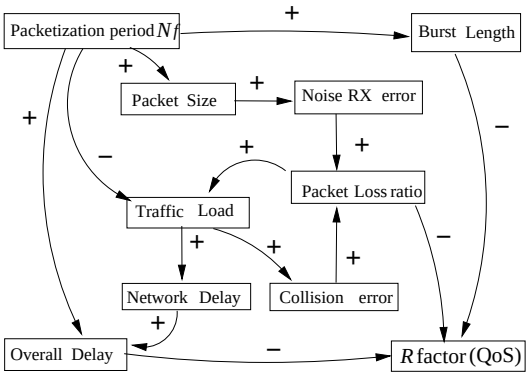


Fig. 1. Relationship between N_f and QoS in VoWiFi.

The energy consumption at the IEEE 802.11 interface can be expressed as:

$$E = P_{tx} \cdot t_{TX} + P_{rx} \cdot t_{RX} + P_{idle} \cdot t_{idle} + P_{sleep} \cdot t_{sleep} \quad (2)$$

Where $P_{tx} > P_{rx} > P_{idle} > P_{sleep}$ represent the power coefficients at the radio interface for transmission, reception, idle and sleep respectively; and $t_{TX}, t_{RX}, t_{idle}, t_{sleep}$ the respective time spent in each state. Nominal values of P_{tx} and P_{rx} are similar in most Wi-Fi cards, and P_{idle} is about 75% of P_{tx} [3]. VoWiFi terminals typically spend less than 2% of their time [14] in active states (i.e. TX or RX). Consequently, the best strategy to achieve significant energy savings is to use PSM and go into the sleep state for as long as possible (as P_{sleep} is significantly smaller than all the others).

In general, increasing N_f (i.e. reducing the packet generation frequency) results in less time spent in active states, reducing the energy consumption. However, in lossy or saturated IEEE 802.11 channels, there exists no such straightforward relationship since larger packets are more prone to suffer transmission errors which will lead to retransmission attempts.

III. Scope and Problem Statement

Our target scenario includes VoWiFi terminals connected through a IEEE 802.11 Access Point to the Internet. The respective communication partners are VoIP terminals operating in a wired environment such as POTS or switched Ethernet (e.g. a call center). This is the scope of

the study which has also been reproduced in the test-bed. In this context, the VoWiFi terminals are assumed to periodically receive QoS-related information from their wired counterparts through various RTCP reports as defined in RFC 3550: Extended(XR), Sender(SR) and Receiver(RR). Upon reception of a report VoWiFi terminals trigger a procedure to determine the value of N_f to be set on both sides of the call for the next period through a SIP RE-INVITE message.

In such a scenario, our goal is to allow VoWiFi users minimize their energy consumption at the IEEE 802.11 interface as long as a minimum target speech quality (R_{min}) is reached in the VoIP communication. Otherwise, the main goal is to reach R_{min}.

IV. Three-step Optimization Procedure

Upon the reception of each RTCP report VoWiFi terminals execute the following procedure.

- **Step 1: Assessing VoIP quality**.
  Counters from RTCP reports along with some locally-obtained information (e.g. packetization and jitter buffer delay) allow the estimation of the network delay, end-to-end delay, packet lost rate and loss burst size. Using this information, each VoWiFi station estimates the QoS of the past period using Eq. 1 as done in [19] from the data obtained from RTCP reports. The output of this stage is the value of R .
- **Step 2: Estimation of Wi-Fi network conditions**
  We assume that in our scenario packet loss and delay come mainly from the IEEE 802.11 network [7], so when the network performance exceeds certain thresholds (e.g. a packet loss rate $> 5\%$ or a delay $> 300ms$) it can be inferred that stations are saturated due to lack of transmission opportunities, which will be indicated by setting flag SAT (stations saturated). The network delay can be obtained by measuring the time difference between Sender and the corresponding Receiver RTCP reports. Network losses can be estimated from the Ratio Loss counter used in RTCP XR reports.
- **Step 3: Finding N_f for the next interval**
  The results from steps 1 (R) and 2 (SAT) are used as input for an algorithm that determines the optimal value of N_f for the next period. A stateful behaviour is needed to keep the values of R, SAT and N_f between two consecutive executions $i-1$ and i.
  Figure 2 shows the proposed algorithm[1]. The flow diagram follows the discussion in Section II. If $R \geq R_{min}$ we increase N_f in order to get higher energy saving at the cost of reducing the QoS (R factor) due to the overall delay increment. However, when $R < R_{min}$ the priority will be to restore R to acceptable values. If flag SAT was off then we could conjecture that the poor quality is caused by excessive packetization delay, so we will decrease N_f. Conversely, if we observe poor network performance

[1]A fixed step size has been used here for simplicity. However, adjustable step size could be used for faster adaptation.

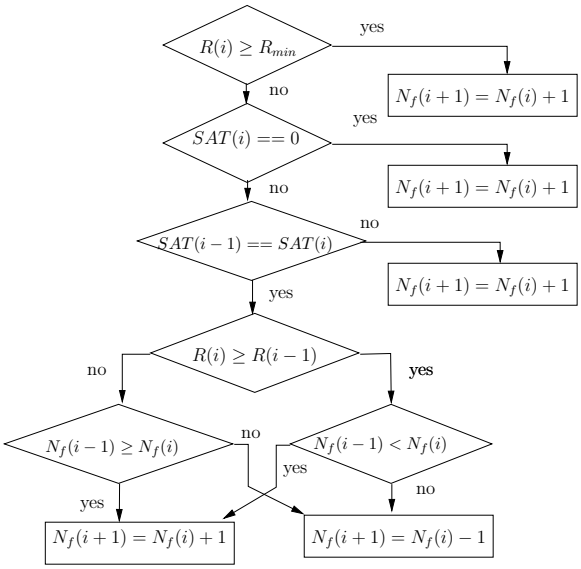


Fig. 2. Flow diagram of step 3

(flag SAT is on) then we perform a heuristic search to find out the N_f value that improves the R factor (recall that no a-priori outcome can be foreseen as discussed in Section II). In this case, if increasing (or decreasing) N_f improves the R factor, the algorithm will continue increasing (or decreasing) N_f until R_{min} is achieved.

V. Testbed and Implementation

A. Scenario set-up

The test scenario is shown in Figure 3. Stations 1 to 4 are equipped with a D-link DWA-131 Wi-Fi card managed by the *ndiswrapper* driver v1.59-6 under Linux Ubuntu distribution v14.04.2. These wireless stations communicate with their wired peers connected to an Ethernet 100Mbps switch. The wireless stations are associated with an Access Point (Airlink DWL-3500AP) set to operate in the IEEE 802.11b with a PHY rate set at 1 Mbps so we can easily experiment with saturation conditions. Finally, a PC acting as router between the Access Point and the Switch is running the *Network Emulator for Windows Toolkit* (NEWT) software to emulate adverse network conditions (i.e. to insert delay or packet losses into the network).

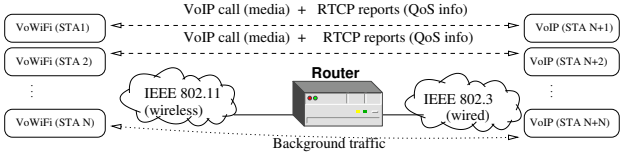


Fig. 3. Test scenario implemented.

Stations 1 to 3 and its peers 5 to 7 are running the VoIP software *pjsip* (the *pjsua* command-line client) [20] containing the modifications necessary to accommodate our packetization period control algorithm from Section IV. Each communication pair uses a 30 min conversation

playback from recorded telephony conversations (LDC data bank, CallHome recordset) [21], each station playing its respective side of the conversation. In our tests, we use the codec G.711 with VAD. Stations 4 and 8 are dedicated to inserting background traffic. Both use *rtptools* [22] as traffic generator to send and receive UDP datagrams to each other.

B. Algorithm implementation

An optimization module that implements the algorithm from Section IV has been written in C language. As illustrated in Figure 4, this module communicates with the VoIP client *pjsua* rather than being integrated into it. Thus, changes in the original *pjsua* VoIP client consist of reporting QoS-related information to our module, reading the algorithm output (N_f) and sending a SIP RE-INVITE message to its peer so that the new packetization period can take effect. VoIP clients can be configured to locally apply the new value of N_f or to apply the value received from the other side of the call.

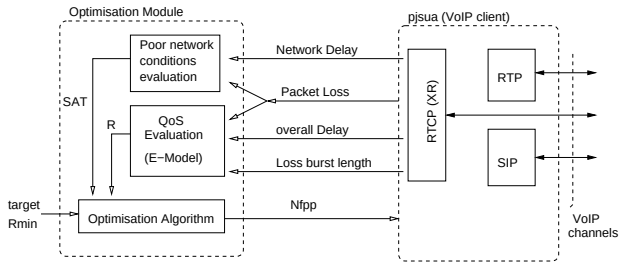


Fig. 4. Module architecture implemented.

QoS-related information is periodically evaluated by *pjsua* and transported in various RTCP reports such as Extended (XR), Sender (SR) and Receiver (RR). By taking advantage of this functionality already implemented, we can simply use these counters and complement them with some additional locally-obtained information to get the input variables for our algorithm. In contrast to the information included in the RTCP reports, these counters are reset between successive reports which allows for an evaluation of the quality of service for during an interval. The QoS evaluation period configured in our tests is 7.64 secs. Small intervals (i.e. < 5 secs.) provide uneven and coarse-grained loss rates due to insufficient number of packets generated (especially with large N_f values), whereas large intervals (e.g. $\geq$ 10secs.) provide a poor reaction time. Therefore, we experimentally found this value to be a good balance.

The QoS variables used by the algorithm are as follows:

- Network Delay. When a SR is transmitted we measure the time until we get the corresponding RR. This round-trip time is corrected by subtracting the processing time at the far end (as indicated in the RR). The network delay is obtained by halving this corrected round-trip time.
- End-to-End delay (d). This is the delay resulting from adding network delay, jitter buffer and packetization period. The packetization period and jitter buffer delay are locally measured by the *pjsua* application every time it reads or writes a new speech frame to the buffer.
- Packet loss rate. This is the overall packet loss resulting from network losses and discarded packets at the jitter buffer. Network losses are calculated from the *Ratio Loss* counter used in the RTCP XR reports. Discarded packets at the jitter buffer are already computed by *pjsua*.
- Loss burst size. This is measured by default in *pjsua* and is included in RTCP XR reports.

Using the previous variables, the optimization module evaluates the following:

- Network conditions flag (SAT) which is set to 1 when either the Network Delay or the Packet loss rate exceeds a certain threshold (300 ms or 5% respectively in our tests).
- R factor. Calculated using Eq. 1 where I_d is calculated from the end-to-end delay d as $0.0024 \cdot d + 0.11(d - 177.3)H(d - 177.3)$ being $H(x)$ the step function (Heavyside function); $I_{e,eff}$ is calculated as $I_e + (95 - I_e) \cdot P_{pl}/(P_{pl}/\text{BurstR} + B_{pl})$, where I_e is a measure of the codec intrinsic speech quality degradation, P_{pl} is the loss rate percentage, BurstR is the ratio of the registered packet loss burst size over random losses, and B_{pl} measures the loss concealment of each codec. Tables with values for the constants B_{pl} and I_d can be found in [19].

After evaluating SAT and R the rest of the algorithm as described in Section IV is executed. In our results we have used a maximum value of 16 and a minimum value of 2 (i.e. 20 ms) for the packetization period. The output of the algorithm is the new value of N_f. The wired VoIP client executes the algorithm, applies the resulting N_f and sends this same value to its wireless counterpart which also applies it.

C. Energy Measurement procedure

In order to measure the energy consumption of each WiFi station 1-3 during the VoIP calls we have used a dedicated device running Wireshark to capture all the physical frames (including management frames) over the air interface during each experiment. Captured frames have been processed afterwards with *awk* to calculate for each station the time spent in every energy state (i.e. TX, RX, IDLE, SLEEP) over intervals of 10 secs. The energy consumption over such interval is then calculated using Eq. 2 with the Wi-Fi adapter power coefficients $1.65, 1.2, 0.9$ and 0.1 Watts for the respective states. Although Wi-Fi adapter power coefficients exhibit high variability, those values can be found in Wireless Adapter from vendors like Intel or D-Link ([23])

VI. Results

Two experiments have been carried out to test the goodness of the proposed algorithm in both uncongested and congested IEEE 802.11 network conditions. Results represent average values from stations 1 to 3.

A. Unloaded Network Scenario (no background traffic)

Figure 5 shows the variation of QoS, delay and energy over the first 700 secs of the conversation for different quality targets (R_{min}). Fig. 5(a) shows the value of R and N_f (as calculated by the algorithm) during the test. Note that we have reduced R_{min} from 90 to 85 and 80 at 300 secs and 550 secs respectively, which resulted in increments of the packetization period up to its maximum value of 16. As observed in Fig 5(c) the increment in N_f results in energy savings at the IEEE 802.11 interface with respect to the default constant value of 2 (20ms) recommended in RFC5761. The bottom part of Fig 5(c) shows savings as large as 30% when PSM is used (approximately 1 Joule of difference), whereas without PSM only 1% is saved which can be attributed to the fact that for 98% of the time Wi-Fi cards are in idle mode as observed in [14]. Fig. 5(b) shows the end-to-end and network delays. The fluctuations in the former can be mostly attributed to the playout buffer management performed by *pjsip*, which dynamically sets the buffer size according to network conditions or discarded packets. For large values of N_f (e.g. 16) a small increment in the buffer size (e.g. 18 speech frames) causes waiting for various packets to arrive (i.e. a packetization delay of 16x2 speech frames) while the next time only one packet is necessary.

B. Congested Network Scenario (background traffic)

In this test background traffic is inserted by sending 100B packets with an increasing frequency resulting in a bit rate from 450 kbps to 650 kbps in steps of 20 kbps and back to its initial value. Packet loss and network delay peak at 20% and 600 ms respectively. Such conditions would prevent any form of acceptable communication and in practice such a VoIP call would be torn down after a few seconds.

We have run the test three times using our algorithm with $R_{min} = 90$, $R_{min} = 75$, and the RFC5761 default constant packetization (N_f=2). Figure 6 shows the results. As expected, while the network conditions allow us to achieve R_{min}, the packetization period increases gradually to save energy. However, as traffic increases and the network conditions deteriorate (the SAT flag is on), the algorithm adjusts N_f to preserve the target quality. But as the background traffic continues to increase, there is a point when it is no longer possible to maintain the quality and it quickly drops to zero. This point is first reached for the case of the default packetization period at around t = 450 secs and 490 kbps of background traffic, next for the algorithm-controlled with $R_{min} = 90$ at around t = 600 secs and 510 kbps of background traffic, and finally for the algorithm-controlled with $R_{min} = 75$ at around t = 1,150 secs and 570 kbps of background traffic. When the background traffic returns to lower values and channel conditions improve, we can observe R returning to acceptable values. Therefore, the use of our proposed algorithm allows longer periods of acceptable quality, typically 48% of the time for the more relaxed target quality and 16% for the more demanding target. The packetization period explains the difference in the quality obtained for each R_{min}. To achieve a higher *R* value, N_f is reduced to its minimum value in order to realize the lowest packetization delay possible. This increases the traffic inserted by the VoIP application and consequently causes stations to saturate faster. This is even worse with the default packetization value of 2.

Regarding the energy consumption, we can observe in Figure 6(b) that the algorithm-controlled packetization schemes achieve higher energy efficiency than the default packetization period. Moreover, the lower the minimum target quality that one is prepared to accept the greater the energy savings that can be realized. From Figure 6(b) it can be observed that there is a difference of 11% in the R value between 600 and 1,150 secs. This difference is greater with respect to the default packetization period as expected (i.e. around 20%). During prolonged saturated conditions there is no significant differences in the energy consumption between both targets as both seek to maximize the packetization period. At any rate, the energy consumption during this period is not relevant as in practice the call would be torn down by users after few seconds due to the lack of a minimum quality.

VII. Conclusions and further work

From our experiments we have demonstrated that the application of our packetization period control algorithm can provide energy savings in VoWiFi terminals at the cost of reducing the VoIP quality down to a minimum acceptable level. In uncongested IEEE 802.11 networks with low delay and loss conditions, we continue to experience acceptable QoS levels and also obtain significant energy savings of up to 30% using PSM depending of the terminal characteristics.

Under poor network conditions, our algorithm allows the extension of experiencing the target quality at the cost of a higher energy consumption when compare with the uncongested scenario. In the worse case we can get no energy saving, but on these network conditions the VoIP call should be torn down. Therefore by adaptatively controlling the packetization period we can tolerate saturated channel conditions for longer periods.

Further works includes dynamic codec selection and variable N_f increments as part of the optimization strategy.

References

[1] S. Harsha, A. Kumar, and V. Sharma, "An analytical model for performance evaluation of multimedia applications over edca in an ieee 802.11e wlan," *Wirel. Netw.*, vol. 16, no. 2, pp. 367–385, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1007/s11276-008-0137-y

[2] K.-W. Chin, "On maximizing voip capacity and energy conservation in multi-rate wlans," *Communications Letters, IEEE*, vol. 14, no. 7, pp. 611–613, 2010.

[3] T. Shiao-Li and H. Chung-Huei, "A survey of energy efficient mac protocols for ieee 802.11 wlan," *Computer Communications*, vol. 34, no. 1, pp. 54 – 67, 2011.

[4] R. G. Cole and J. H. Rosenbluth, "Voice over ip performance monitoring," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 2, pp. 9–24, Apr. 2001.

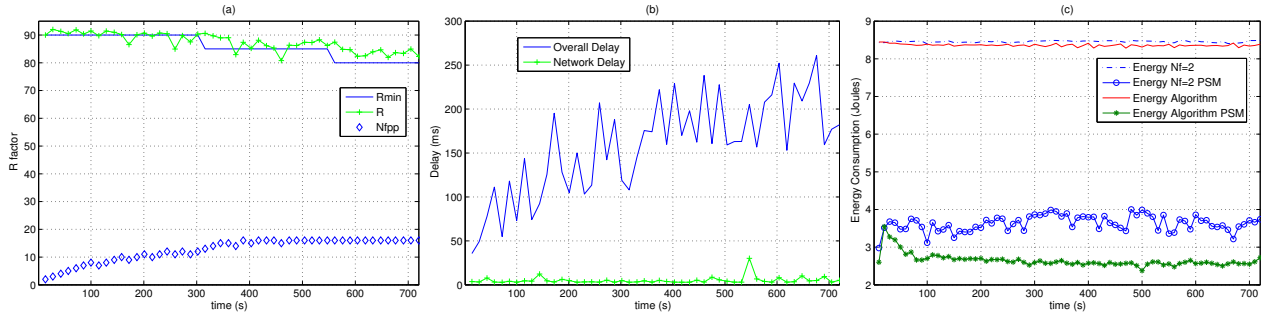


Fig. 5. Scenario without background traffic: evolution of (a) R and N_f, (b) delay and (c) energy over the first 700 secs.

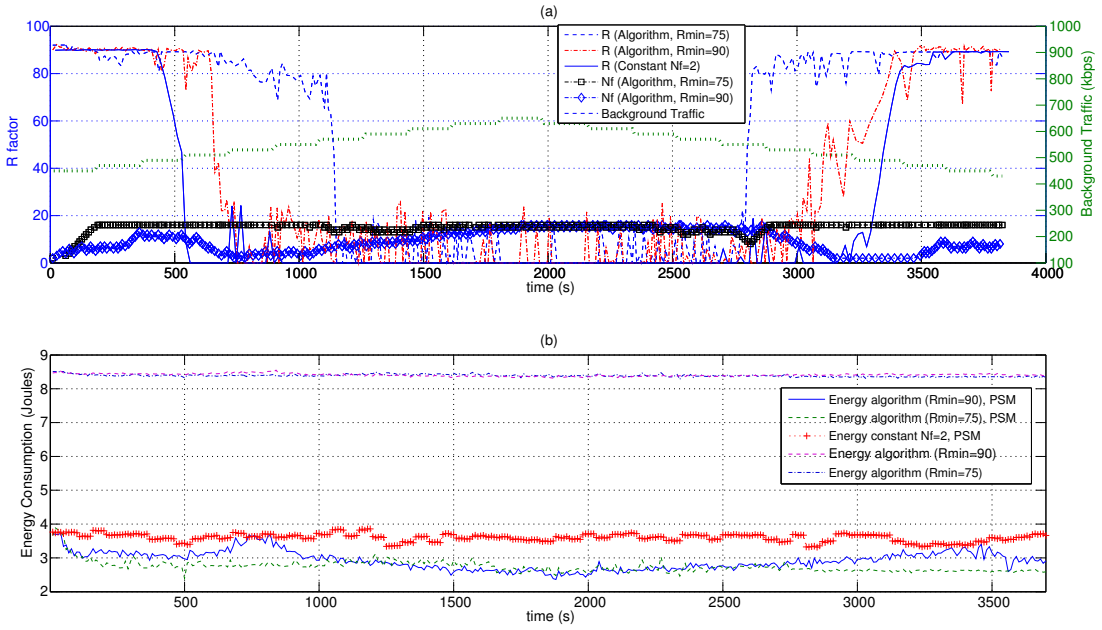


Fig. 6. Scenario with background traffic (a) quality and packetization period for different packetization schemes; (b) energy consumption with different packetization schemes.

[5] A. Trad, F. Munir, and H. Afifi, "Capacity evaluation of voip in ieee 802.11e wlan environment," in *Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE*, vol. 2, 2006, pp. 828–832.

[6] L. Cai, X. Shen, J. W. Mark, and Y. Xiao, "Voice capacity analysis of wlan with unbalanced traffic," in *Quality of Service in Heterogeneous Wired/Wireless Networks, 2005. Second International Conference on*, 2005, pp. 8 pp.–9.

[7] S. Shin and H. Schulzrinne, "Measurement and analysis of the voip capacity in ieee 802.11 wlan," *Mobile Computing, IEEE Transactions on*, vol. 8, no. 9, pp. 1265–1279, 2009.

[8] G. Kuriakose, S. Harsha, A. Kumar, and V. Sharma, "Analytical models for capacity estimation of ieee 802.11 wlans using dcf for internet applications," *Wirel. Netw.*, vol. 15, no. 2, pp. 259–277, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1007/s11276-007-0051-8

[9] G. Boggia, P. Camarda, L. Grieco, and S. Mascolo, "Feedback-based control for providing real-time services with the 802.11e mac," *Networking, IEEE/ACM Transactions on*, vol. 15, no. 2, pp. 323–333, 2007.

[10] G. Nikolakopoulos, A. Panousopoulou, and A. Tzes, "Experimental controller tuning and qos optimization of a wireless transmission scheme for real-time remote control applications," *Control Engineering Practice*, vol. 16, no. 3, pp. 333 – 346, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066107001074

[11] K. Stoeckigt and H. Vu, "Voip capacity analysis in ieee 802.11 wlan," in *Local Computer Networks, 2009. LCN 2009. IEEE 34th Conference on*, 2009, pp. 116–123.

[12] S.-L. Tsao and C.-H. Huang, "An energy-efficient transmission mechanism for voip over ieee 802.11 wlan," *Wireless Communications and Mobile Computing*, vol. 9, no. 12, pp. 1629–1644, 2009. [Online]. Available: http://dx.doi.org/10.1002/wcm.747

[13] L. S. G. D. Carvalho and E. D. S. Mota, "Survey on application-layer mechanisms for speech quality adaptation in voip," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, p. 36, 2013.

[14] A. J. Estepa, J. M. Vozmediano, J. López, and R. M. Estepa, "Impact of voip codecs on the energy consumption of portable devices," in *Proceedings of the 6th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks*. ACM, 2011, pp. 123–130.

[15] V. Namboodiri and L. Gao, "Energy-efficient voip over wireless lans," *Mobile Computing, IEEE Transactions on*, vol. 9, no. 4, pp. 566–581, 2010.

[16] X. Pérez-Costa, D. Camps-Mur, and A. Vidal, "On distributed power saving mechanisms of wireless lans 802.11 e u-apsd vs 802.11 power save mode," *Computer Networks*, vol. 51, no. 9, pp. 2326–2344, 2007.

[17] P. Serrano, A. Banchs, P. Patras, and A. Azcorra, "Optimal configuration of 802.11 e edca for real-time and data traffic," *Vehicular*

Technology, IEEE Transactions on, vol. 59, no. 5, pp. 2511–2528, 2010.

[18] A. J. Estepa, R. Estepa, J. Vozmediano, and P. Carrillo, "Dynamic voip codec selection on smartphones," *Network Protocols and Algorithms*, vol. 6, no. 2, pp. 22–37, 2014.

[19] A. Kovac, M. Halas, M. Orgon, and M. Voznak, "E-model mos estimate improvement through jitter buffer packet loss modelling," *Advances in Electrical and Electronic Engineering*, vol. 9, no. 5, pp. 233–242, 2011.

[20] S. PJSIP-Open Source, "Stack and media stack for presence, im/instant messaging, and multimedia communication, 2008."

[21] C. Alexandra, D. Graff, and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium, Philadelphia*, 1996.

[22] H. Schulzrinne, P. Pan, A. Tsukamoto, D. Sisalem, and S. Casner, "Rtp tools," *URL: ftp://ftp. cs. columbia. edu/pub/schulzrinne/rtptools/rtptools. html*, 1996.

[23] S. Chiaravalloti, F. Idzikowski, and L. Budzisz, "Power consumption of wlan network elements," no. TKN-11-002, 2011.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6499

Web of Energy: hacia la integración inteligente para las redes de sensores en Smart Grids

Víctor Caballero, David Vernet, Agustín Zaballos, Guiomar Corral
Grupo de Investigación en Internet Technologies & Storage (GRITS)
La Salle - Universitat Ramon Llull (URL),
Barcelona, España.
{vcaballero,dave,zaballos,guiomar}@salleurl.edu

***Resumen*- Las Smart Grids se presentan como una solución a la demanda de la gestión distribuida e inteligente de la energía, mejorando así las funciones de automatización, recolección y procesamiento de datos. Para dar soporte a estas funciones inteligentes, es necesario el despliegue de novedosas infraestructuras TIC diseñadas para tal propósito. Una de las principales problemáticas de las Smart Grids es la heterogeneidad de protocolos de comunicación utilizados por los dispositivos inteligentes que las integran. Para dar solución a este problema, se propone el uso del concepto de la *Web of Things*. Sin embargo, somos conscientes de que las entidades que se encargan de la generación y gestión de la energía eléctrica pueden encontrar dificultades para adaptar su infraestructura a este novedoso entorno. Este artículo propone la introducción del *Actor Model* como modelo de diseño para una infraestructura que soporte las demandadas funciones inteligentes y sea capaz de agrupar y convertir la heterogeneidad de las infraestructuras tradicionales en la homogeneidad característica de la Web of Things.**

***Palabras Clave*- Smart Grid, Redes de Sensores, Internet of Things, Web of Things, Actor Model**

I. Introducción

El Internet of Things (IoT) se presenta como la evolución natural de Internet, pues se refiere a la aglomeración heterogénea de dispositivos conectados a la red. Esta evolución ha sido posible gracias a los avances tanto del silicio, permitiendo incrustar unidades de computación cada vez más y más pequeñas en los dispositivos cotidianos, como de los avances en protocolos *wireless* de bajo consumo para dichos dispositivos. Las Tecnologías de la Información y las Comunicaciones (TIC) de las Smart Grids representan un buen ejemplo de esta heterogeneidad, donde diferentes dispositivos, tanto sensores como actuadores, de diferentes vendedores y usando diferentes protocolos, se combinan para alcanzar un objetivo: la integración de energía y servicios inteligentes. Ante la problemática de la heterogeneidad, la Web of Things (WoT) [1] se presenta como una capa de abstracción que posibilita la interacción homogénea entre los dispositivos de diferente índole utilizando tecnologías web. En este sentido, se propone agrupar la aplicación de las metodologías proporcionadas por la WoT a la gestión de las Smart Grids bajo el término Web of Energy (WoE) [2]. En trabajos anteriores [3] se introdujeron algunas propuestas en este sentido donde se concluía que se necesitaba otro acercamiento para el desarrollo de una arquitectura para la Web of Energy.

Este artículo se estructura como sigue: en primer lugar se presenta el concepto de Web of Energy junto con los dos conceptos de los que se nutre: (1) las Smart Grids y sus propuestas de valor y retos asociados y (2) la Web of Things y sus propuestas de valor. A continuación, se plantea el problema de la integración directa de las Smart Grids, la Web of Things y los protocolos de comunicación. Y, finalmente, se plantea una propuesta de arquitectura híbrida que cubra las necesidades de una infraestructura de Smart Grids y aporte las ventajas que supone integrar las TIC de las Smart Grids con la Web of Things.

A. Smart Grids y Web of Energy

Las Smart Grids promueven la gestión de la energía eléctrica de forma distribuida y flexible. Sin embargo, los sistemas de gestión actuales son (i) centralizados en su

control, (ii) ubicados en silos independientes entre ellos y (iii) gestionados por aplicaciones fragmentadas, sin integración entre ellas y solo intercomunicadas gracias a canales de comunicación específicos, generalmente propietarios.

El objetivo de las Smart Grids es el de proveer mejores servicios y características (también conocidas como funciones inteligentes), tanto para los consumidores como para los productores y prosumers. Además, la mayor utilización de generación distribuida y renovable de energía demanda cambios en el sistema de gestión de la electricidad. Se hace necesaria la mejora de los sistemas de automatización, inteligencia distribuida, minería de datos en tiempo real y gestión para mejorar las funciones de control de la red, reducir la configuración y reducir los tiempos de recuperación y auto cicatrización de los sistemas.

Las TIC de las Smart Grids se pueden considerar un caso particular del IoT. Uno de los objetivos y retos más relevantes de las Smart Grids es la integración de datos y comunicaciones entre una red de diferentes tipos de sensores y actuadores que las integran (sensores con y sin cables, medidores inteligentes o *smart meters*, generadores distribuidos, etc.) que, a su vez, deben ser capaces de cooperar y coordinarse entre ellos para que puedan ejecutar cualquier función deseada. Esto es, el objetivo de las Smart Grids es el de crear un único sistema de integración para que las diferentes aplicaciones puedan aprovecharse de las mismas ventajas [4]. En [5] se exponen esta y otras problemáticas inherentes al IoT y, por ende, a las Smart Grids.

Para dar solución a la integración de los dispositivos heterogéneos se ha propuesto el uso del concepto de la Web of Things [1] para acceder a los múltiples dispositivos usando una misma interfaz facilitada por las tecnologías web, implicando uniformidad en los protocolos de comunicación (HTTP y WebSockets), en el modelo de representación de los datos y en el descubrimiento de dispositivos (Web Thing Model [6] y Web Semántica [7]). Este concepto es, hasta el momento, de difícil aplicación en escenarios reales que involucren la gestión de la energía debido a la posibilidad de crear nuevas vulnerabilidades de seguridad, la falta de dispositivos que implementen estándares de la Web of Things y la oposición de la industria de la energía a incluir módulos o dispositivos externos en sus sistemas propietarios.

El objetivo del trabajo que se presenta en este artículo es el de crear una arquitectura basada en el paradigma del IoT para gestionar las necesidades de almacenamiento y de comunicación de las Smart Grids y a la vez enlazar las Smart Grids con el usuario final a través de las metodologías de la Web of Things. Para ello, se establece una interfaz bidireccional H2M (*human-to-machine*) inspirada por la WoT que permite un control ubicuo de los sistemas de energía, la Web of Energy [2].

De este modo, la WoE podría representar una oportunidad para que las eléctricas puedan disponer de representaciones virtuales de sus dispositivos más flexibles (basadas en software, actualizable, configurable y con posibilidad de desplegar nuevas aplicaciones encima de ellos), posibilitando una la distribución y gestión de bajo coste de la red eléctrica [8]. La WoE facilita la compartición de datos de diferentes dispositivos (puntos de recarga de vehículos eléctricos, *smart metering* o monitorización de subestaciones) con terceros. Además, el uso de tecnologías web permite la creación de herramientas de visualización multiplataforma y sin coste de instalación, pudiendo ofrecer interfaces gráficas simples y usables para una mayor adopción para los DSO (*Distribution Systems Operator*) o cualquier usuario interesado en la consumición o producción de energía (*prosumer*). Así pues, podemos identificar distintos campos relacionados con la gestión de la energía en los que la WoE sería de gran utilidad:

- Acceso remoto desde subestaciones a los servidores centrales
- Monitorización y control de DER (*Distributed Energy Resources*)
- Distribución de SCADA a subestaciones secundarias
- Control de EVSE (*Electrical Vehicle Supply Equipment*)

II. Web of Things

Aunque las predicciones iniciales de 1 billón de dispositivos conectados a Internet en 2015 [9] fueron rápidamente rebajadas a 26.000 millones en 2020 [10] y 20.800 millones en 2020 [11], es evidente que el número de dispositivos conectados a Internet incrementa día a día. Cómo acceder a todos estos sensores y actuadores a través de una interfaz uniforme es indudablemente uno de los mayores retos del IoT. Actualmente, el IoT está dividido en silos, esto es, existen soluciones propietarias que ayudan a la integración de un conjunto determinado de dispositivos, pero se alejan de la integración global prevista para el IoT. Es por este motivo se propone el uso de tecnologías web ya existentes para la integración global de los dispositivos. Los prerrequisitos básicos para la habilitación de los dispositivos del IoT en la web son dos: (1) capacidad mínima de procesamiento de datos y (2) conectividad a la red (no es necesario que se conecten directamente a Internet).

A. De la heterogeneidad a la homogeneidad

El modelo que propone la Web of Things [1] pretende dar solución al reto de la heterogeneidad de los dispositivos que componen el IoT. Esto lo consigue, en gran parte, generando traductores o *mappings* entre el lenguaje hablado por cada dispositivo (protocolo de comunicación y formato de los datos) y el lenguaje que podemos considerar "universal" por su uso extendido: las tecnologías web. En concreto disponemos del protocolo HTTP y las APIs RESTful [12]. Estas traducciones habilitan a los dispositivos a hablar un mismo lenguaje, lo que permite que puedan ser accesibles de forma homogénea por actores humanos o

computerizados, como otros dispositivos. La acción a realizar sobre estos dispositivos puede ser tanto para actuar como para captar información.

Si nos centramos en el flujo de datos, este se compone de dos estados que se pueden entender como un ciclo: de heterogeneidad a homogeneidad y de homogeneidad a heterogeneidad.

- De heterogeneidad a homogeneidad. Un dispositivo envía datos que ha captado con formato y protocolo específicos a través de un traductor WoT, que traduce los datos a un formato común y el protocolo a HTTP.
- De homogeneidad a heterogeneidad. El actor recibe estos datos y decide actuar sobre el dispositivo, por ejemplo, cambiando su configuración. Entonces envía una instrucción mediante el protocolo HTTP y un formato común al traductor WoT que, a su vez, traduce el protocolo y el formato de los datos al protocolo y formato específicos.

III. La arquitectura de la Web of Things

El concepto de Web of Things se desarrolla principalmente en los trabajos de Dominique Guinard [1] y Vlad Trifa [13], donde se presentan la WoT y los métodos de habilitación de dispositivos del IoT a la WoT respectivamente. En [1] se propone organizar la WoT en cuatro capas, cada una con una función específica. Sin embargo, estas capas no siguen un modelo de aislamiento y encapsulación entre capas no contiguas como ocurre con el modelo OSI o TCP/IP, sino que, por el contrario, las aplicaciones se pueden construir encima de cada una de ellas (Fig. 1), ya que todas ellas forman parte del nivel de aplicación de los dos modelos de red mencionados anteriormente. A continuación, se presentan las diferentes capas de la WoT propuestas en [1].

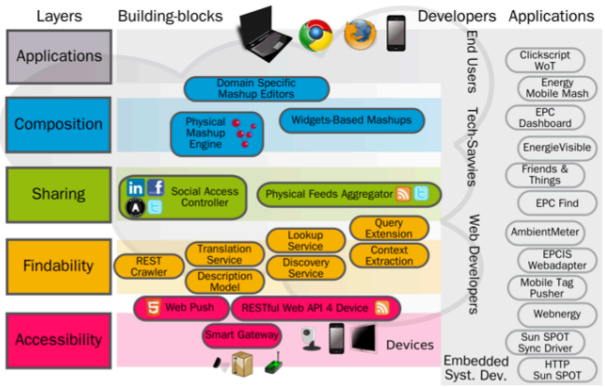


Fig. 1 Modelo de capas propuesto en [1]. Imagen extraída de [1]

Las funciones de cada una de estas capas son:

(i) Accesibilidad: Habilita un acceso consistente a todo tipo de dispositivos del IoT exponiendo sus funcionalidades mediante una API RESTful HTTP.

(ii) Localización: Facilita el descubrimiento de las representaciones de los diferentes dispositivos, modelando de manera uniforme el método de acceso a estas (a través ya del protocolo HTTP) y estableciendo relaciones entre ellas en el momento de su representación.

(iii) Participación: Se encarga de preservar la privacidad entre las representaciones de los dispositivos y de gestionar la autenticación y autorización del acceso a las representaciones por parte de otros actores distintos al propietario del dispositivo.

(iv) Composición: Su función es la de habilitar la integración entre las distintas representaciones de los dispositivos que, en última instancia, permite la integración entre las diferentes funcionalidades de los dispositivos físicos.

A. Capa de Accesibilidad

La capa de accesibilidad actúa de interfaz entre el IoT y las tecnologías web. Por lo tanto, es la más próxima a la heterogeneidad de protocolos del IoT, tales como MQTT [14] o CoAP [15], entre otros. En esta capa se encuentran soluciones en forma de proxy o puente entre los protocolos del IoT y HTTP.

Como metodologías básicas se pueden considerar la incrustación de servidores web a los dispositivos o la creación de *gateways* [16] con las funciones de agregación y/o proxy entre protocolos del IoT y protocolos web. Se encuentran en esta capa también soluciones cloud, generalmente transversales a más capas de la WoT, como ThingWorx [17], Watson IoT Platform de IBM [18], Octoblu [19] o EVRYTHNG [20], entre otros. Estas soluciones *cloud* ofrecen *gateways* de traducción IoT – WoT.

B. Capa de Localización

Esta capa la componen aquellas tecnologías que permiten explorar y encontrar los distintos dispositivos expuestos a la WoT. Agrupa aquellas tecnologías que permiten la publicación de datos estructurados e interconectados. En la publicación de las representaciones de los dispositivos se aplica el concepto de REST [12] y Linked-Data [21] para interconectar distintas representaciones (generalmente en forma de URLs – *Uniform Resource Locators*) .Las tecnologías de la Web Semántica [7] tales como RDF [22], RDFa [23] o OWL [24] intervienen para dotar de significado semántico tanto a las representaciones como a las conexiones con otras representaciones, permitiendo, por ejemplo, su exposición a los motores de búsqueda como Google.

C. Capa de Participación

Esta capa agrupa todos aquellos métodos que permiten autenticar y autorizar a diferentes actores a realizar una acción sobre la representación virtual del dispositivo y, en última instancia, sobre el dispositivo físico. Agrupa métodos de autenticación y autorización

simples desde la autenticación con usuario y contraseña hasta el uso de protocolos más avanzados como claves API u OAuth.

	HTTP	WebSocket	CoAP	MQTT	XMPP	AMQP	DDS
Topología	*Req/Resp*	*Two-way, realtime*	*Req/Resp*	*Pub/Sub*	*Pub/Sub yReq/Resp*	*Pub/Sub*	*Pub/Sub*
Arquitectura	*P2P*	*P2P*	*P2P*	*Broker*	*P2P*	*P2P o Broker*	*Global Data Space*
Nivel de Transporte	*TCP*	*TCP*	*UDP*	*TCP*	*TCP*	*TCP*	*TCP/UDP*
Encriptación	*TLS*	*TLS*	*DTLS*	*TLS*	*TLS*	*TLS*	*TLS/DTLS*
Autenticación	*TLS*	*TLS*	*DTLS*	*User/Pass*	*TLS*	*SASL*	*TLS/DTLS*
QoS	-	-	*Confirmable*	*3*	-	*3*	*23*

Tabla 1. Protocolos del WoT/IoT

En [25] se propone un middleware, el *Social Access Controller* que, combinando las APIs OAuth de las diferentes redes sociales y el acceso por credenciales simples a los dispositivos (usuario y contraseña p.ej.) permite (1) preservar la privacidad de los dispositivos físicos, (2) aprovechar la estructura y la función de las redes sociales para autenticar a los diferentes actores potencialmente interesados en participar en las acciones que se puedan realizar sobre la representación del dispositivo, (3) integrar las representaciones de los dispositivos en las redes sociales, creando un *Social Web of Things* y (4) posibilitar la publicación de agregaciones de datos mediante protocolos de sindicación de datos como ATOM [26].

D. Capa de Composición

Esta capa permite la composición de las diferentes funcionalidades expuestas por las representaciones de los dispositivos. Teniendo en cuenta que cada representación es accesible mediante el protocolo HTTP, es sencillo elaborar un *script* para que los diferentes dispositivos actúen de forma coordinada. La elaboración de esta idea consiste en proporcionar al usuario web (humano) una interfaz visual para componer mediante relaciones diferentes dispositivos. Soluciones como ThingWorx Composer [27], NODE-Red [28] de IBM o Octoblu [19] ejemplifican la composición de dispositivos físicos (*physical mashups).*

IV. Protocolos de la WoT y la WoE

El elemento principal de la propuesta del Web of Things es el uso de tecnologías web para la transmisión de información. En lo que se refiere a protocolos web, se encuentran el protocolo HTTP(S) y el protocolo WebSocket (WS) o WebSocket Secure (WSS). HTTP es un protocolo de petición/respuesta mientras que WebSocket permite una conexión bidireccional entre cliente y servidor.

No obstante, muchos dispositivos no disponen de las características necesarias para conectarse directamente a la web mediante HTTP o WebSocket. Para poder conectarse a la web, necesitan hacer uso de *gateways* que se encarguen de traducir la heterogeneidad de protocolos del IoT a la homogeneidad de la WoT. Para ello se necesitan establecer puentes de traducción o mapeo de protocolos IoT a WoT y viceversa.

Por ejemplo, tanto CoAP como MQTT son dos protocolos en auge[1] para el IoT, entre otros. CoAP, estándar abierto, fue diseñado específicamente para el IoT y para ser directamente compatible con HTTP [29]. Es un protocolo basado en petición/respuesta mediante paquetes UDP y sigue los mismos esquemas que HTTP, permitiendo crear recursos REST. El protocolo MQTT, diseñado por IBM, es ahora también de estándar abierto, pero sigue un paradigma publicador/subscriptor sobre TCP, por lo que se hace más complicado establecer un puente de traducción entre HTTP-REST y MQTT [30].

En la Tabla 1 se pueden apreciar las diferencias y similitudes entre los protocolos CoAP, HTTP, MQTT y WebSocket. Como se ha comentado, la traducción entre CoAP y HTTP es directa pues la topología de ambos es petición/respuesta (Req/Resp). En cambio, la topología entre HTTP y MQTT es distinta por lo que la traducción entre estos dos protocolos es más difícil de conseguir. Otro ejemplo de traducción (casi) directa se da entre WebSocket y MQTT, ya que la comunicación bidireccional de WebSocket se puede considerar una particularidad del protocolo publicador/subscriptor (Pub/Sub) de MQTT. Así pues, HTTP y WebSocket son las dos tecnologías web con topologías distintas que permiten la traducción con protocolos del IoT con topologías petición/respuesta y bidireccionales respectivamente. En la Tabla 1 se presentan también otros protocolos que pueden formar parte del IoT, los cuales presentan más similitudes con HTTP o WebSocket según su topología del mismo modo que ocurre con CoAP y MQTT.

Además de los dos protocolos ya presentados en la Tabla 1 y de muchos otros tantos propietarios como no propietarios para el IoT, a medida que nos adentramos en organizaciones cuya infraestructura es más antigua, surgen otros protocolos mucho más adaptados a las necesidades específicas de estas y, a su vez, más difíciles de adaptar para la WoT debido a que (1) su especificidad

[1]https://trends.google.es/trends/explore?q=MQTT,CoAP,XMPP,AMQP. A fecha de abril 2017, las búsquedas de MQTT superan las búsquedas de CoAP, con 92 y 15 puntos sobre 100 respectivamente. Se puede observar como a partir de la segunda mitad de octubre de 2015, las búsquedas relacionadas con MQTT superan a las búsquedas relacionadas con XMPP.

reduce el interés de terceras personas a contribuir a la exposición de los dispositivos (o conjuntos de ellos) a la WoT y, por lo tanto, las empresas deben invertir más capital en la traducción y (2) en el caso de que sea de interés para la empresa exponer algunos de sus dispositivos, la WoT añade un *stack* de nuevas tecnologías, abriendo nuevos agujeros de seguridad a su sistema. Estos retos se agravan en la Web of Energy, pues los sistemas implicados abarcan una gran cantidad de dispositivos que usan protocolos muy específicos y adaptados a las necesidades de estos.

V. Complementando la Arquitectura WoT

El objetivo de esta sección es el de presentar los parámetros clave que creemos que debe cumplir una arquitectura WoT. Estos son:

(i) Aunque ciertos dispositivos del IoT pueden soportar *stacks* HTTP, existen muchos de ellos que por el hecho de disponer de recursos limitados solo pueden soportar protocolos más ligeros. Aunque idealmente un estrecho abanico de protocolos del IoT (p. ej. CoAP y MQTT) facilitaría la integración de los dispositivos a Internet y a la Web, un requisito básico e indispensable para que los dispositivos puedan formar parte de la Web es que se puedan conectar a Internet.

(ii) La arquitectura debe proveer abstracciones para que los desarrolladores puedan interactuar con los dispositivos independientemente del protocolo de comunicación, ya sea en dirección hacia la heterogeneidad de protocolos del IoT como en la dirección de la homogeneidad de protocolos de la WoT.

(iii) La arquitectura debe ser capaz de escalar horizontalmente y proporcionar auto cicatrización para sus sistemas. Debe permitir el despliegue bajo demanda de recursos.

(iv) Para la arquitectura, los dispositivos se convertirán en objetos virtuales o *Things*, aunque también se podrán crear objetos virtuales a partir de agregaciones de otros objetos virtuales.

(v) Cada objeto virtual será accesible desde una interfaz REST HTTP y, en caso de que las características del objeto virtual lo permitan, desde un enlace WebSocket.

(vi) Debe permitir ejecutar protocolos de autenticación y autorización hacia los distintos dispositivos implicados.

En nuestra propuesta, identificamos 5 capas para la creación de una arquitectura de la Web of Energy (Fig. 2) que, aunque tienen una profunda relación con las capas propuestas en [1], también se pueden identificar algunas diferencias, principalmente debido a que las capas que se presentan a continuación están más enfocadas al desarrollo de la arquitectura:

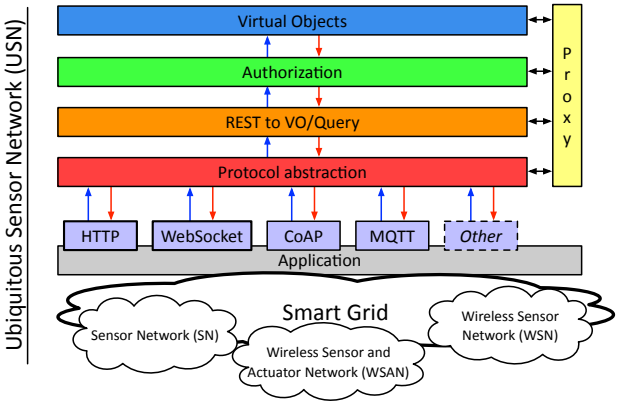


Fig. 2 Arquitectura para el Web of Energy

(i) Capa de Abstracción de Protocolos: El objetivo de esta capa es el de proveer una capa de abstracción para los desarrolladores para que puedan interactuar con los dispositivos físicos, tanto en el sentido de desarrolladores de funcionalidades de la arquitectura a nivel de capas más internas como de los desarrolladores de aplicaciones para la Smart Grid que, como ya se ha especificado, agrupa protocolos distintos. La idea no es solo exponer los dispositivos del IoT como recursos HTTP REST, sino proveer a los desarrolladores de mecanismos de abstracción tanto del protocolo HTTP como de protocolos del IoT. En este sentido, se mantiene la premisa expuesta en [1] de la habilitación de acceso a los dispositivos físicos a través de una interfaz HTTP.

(ii) REST a VO/Query: El propósito de esta capa es el de traducir las diferentes URIs generadas a partir de la información de los dispositivos físicos a acciones sobre los objetos virtuales o *Things*. De este modo, no se actúa directamente sobre los dispositivos físicos sino a través de los objetos virtuales. Esta capa también incluye una interfaz para realizar consultas más complejas sobre un *Thing* o las relaciones de diferentes *Things*.

(iii) Autorización: Esta capa es responsable de pedir y conceder acceso entre objetos virtuales o *Things*.

(iv) Espacio de Objetos Virtuales: Cuando los dispositivos físicos deban realizar acciones sobre otros dispositivos físicos enviarán la instrucción contra una representación virtual (objeto virtual o *Thing*). El objetivo de esta representación virtual es el de aumentar los recursos y funcionalidades de los dispositivos físicos. No podemos aportar una lista detallada de las funcionalidades añadidas, pues son específicas de cada solución. Aun así, nuestro objetivo es proporcionar un método (inyección de código) para que estas funcionalidades se puedan añadir de forma dinámica. Algunas funcionalidades

comunes son la de razonamiento (inteligencia artificial) o la de caché.

(v) Capa de Proxy: La arquitectura de la Web of Things presentada en [1] propone que todas las representaciones de los dispositivos utilicen tecnologías web (HTTP, p.ej.) para comunicar y exponer sus características y conseguir el acceso homogéneo a los dispositivos, tanto entre dispositivos como entre dispositivos y humanos. Sin embargo, usar tecnologías web para todo tipo de comunicación entre servidores específicos de la Web of Things no siempre será lo más óptimo. La motivación de esta capa es pues explicitar que la comunicación eficiente entre servidores se puede llevar a cabo mediante otro protocolo que no esté agrupado dentro de las tecnologías web. No obstante, es importante mencionar que incluso los servicios a los que se hace referencia podrían disponer de una interfaz HTTP para la integración en la WoT.

Finalmente, la capa de aplicación conceptualiza todas aquellas aplicaciones que son susceptibles de interactuar con la arquitectura. Dentro de todo el rango de aplicaciones soportadas por la Web of Things y, por lo tanto, del Internet of Things, se agrupan aquellas aplicaciones propias de las Smart Cities y las Smart Grids. En concreto la generación de una Red de Sensores Ubicua [4] es una aplicación que comprende la aglomeración de Redes de Sensores y Actuadores por cable y sin cable.

VI. Implementación

A. Primera propuesta, implementación en PHP

El objetivo que nos propusimos en [3] era el de crear una arquitectura de la Web of Things de alto rendimiento, a la vez que experimentábamos con el lenguaje de programación web más usado hasta ese momento, PHP, para comprobar si se podía facilitar a los muchos programadores web en PHP [31, 31] las interficies de programación para la Web of Things consiguiendo así una mayor adopción en menor tiempo. Los resultados no fueron positivos por las siguientes razones:

- Los servidores web de PHP siguen un modelo de ejecución muy estricto y dificultan el uso de tecnologías web como WebSocket. Las conexiones WebSocket son permanentes y los servidores PHP acotan la conexión con el cliente hasta un cierto límite de tiempo o hasta que se ha finalizado de ejecutar el script ya que el modelo de ejecución principal es *"load, execute, die"*. Existen alternativas a este modelo de ejecución como el que se implementa en React PHP [33] (patrón reactor), pero esta librería no está disponible para entornos de producción a gran escala como la WoT.
- PHP, un lenguaje interpretado, es más lento que los lenguajes compilados. Además, con el modelo de ejecución comentado anteriormente, los desarrolladores principales nunca se han preocupado de optimizar las instrucciones generadas ni de eliminar los múltiples *"memory leaks"*. Ahora, con la llegada de PHP 7 sí que se ha conseguido una mejora del rendimiento [34], aunque las otras razones que se comentan en esta lista siguen siendo válidas.
- Existen muy pocas librerías enfocadas a funcionalidades que involucren cálculos matemáticos intrínsecos al *Machine Learning* o modelos de computación distribuida. Esto es debido, una vez más, a la idiosincrasia de este tipo de lenguajes, los cuales se emplean principalmente para servir páginas web dinámicas.

B. Segunda propuesta, aproximación a través del Actor Model

En [3] se concluye que PHP no es suficientemente eficiente ni dispone de las herramientas necesarias para una implementación a gran escala de una infraestructura para la Web of Things. Una vez descartado el uso de PHP para desarrollar la infraestructura de la Web of Things, buscamos un lenguaje de programación que tuviera el rendimiento, el ecosistema y las herramientas necesarias para desarrollar tal infraestructura. La JVM (Java Virtual Machine) provee un entorno de ejecución estable (ya que está apoyada por Oracle) y muchas herramientas y librerías (protocolos, frameworks web, librerías de *Machine Learning*) están disponibles para esta plataforma.

Por otro lado, nos interesaba encontrar un modelo de programación que facilitase la creación de sistemas distribuidos, flexibles y auto cicatrizantes, características obligatorias para la WoE. Es por este motivo que el *Actor Model* [35] parece el modelo ideal. Este modelo o paradigma de programación facilita la creación de sistemas con estos requerimientos y, además, gracias a que permite concurrencia, es capaz de aprovechar los diferentes núcleos de computación de una máquina [36].

Los principios básicos de funcionamiento de este modelo son que cuando un Actor recibe un mensaje, éste puede:

- Enviar mensajes a otros Actores
- Crear nuevos Actores
- Designar como gestionar el siguiente mensaje que reciba

Aunque el *Actor Model* puede ser usado para crear agentes y, por lo tanto, Sistemas Multi Agente (MAS), este artículo se centra en la exposición y utilización del modelo *per se* en el IoT o la WoT.

En [37] los autores se benefician de este modelo para habilitar características como el *multi-cloud* y la *multi-tenencia* para dispositivos del IoT. Se aprovechan de los pocos recursos que necesita un actor para operar, compartimentando los dispositivos físicos en diferentes módulos de *software* (actores) aislados y conectados con diferentes infraestructuras *cloud* y diferentes propietarios.

En [38] se propone el uso de CAF (C++ Actor Framework) para proveer a los desarrolladores de un sistema de abstracción del Sistema Operativo de alto nivel para desarrollar aplicaciones del IoT. Destacan también las propiedades de abstracción, distribución y flexibilidad de este modelo de computación.

Gracias a los principios sobre el que está construido el *Actor Model* podemos conseguir propiedades interesantes para la infraestructura WoE.

- Distribución y flexibilidad: Uno de los principios básicos del modelo es la creación de nuevos actores. Esto facilita la utilización de los recursos de computación no solo en un mismo nodo sino en diversos nodos de forma distribuida ya que la comunicación es asíncrona y transparente entre actores de nodos locales y remotos. Además de poder crear nuevos actores, estos también se pueden destruir y, por lo tanto, permite la gestión de la utilización de recursos.
- Auto cicatrización: Esta propiedad no se basa en ningún principio del *Actor Model* teórico, pero todas las librerías que implementan este modelo la codifican desde la creación de *Erlang* debido a su gran uso práctico. Por definición, el estado de un actor está aislado en ese actor y solo se puede comunicar con el exterior a través de mensajes. La auto cicatrización se aprovecha de este principio de aislamiento para gestionar el fallo de un actor determinado. Por ejemplo, Akka [39] implementa la supervisión parental. La creación de actores por parte de otro actor resulta en la creación de una jerarquía con un actor "padre". En el caso de que un actor "hijo" falle, el actor "padre" puede decidir cómo gestionar este fallo: reiniciarlo o no hacerlo, por ejemplo. Gracias al aislamiento entre actores y a la supervisión entre actores se pueden aislar y gestionar los errores de ejecución de forma controlada.

Gracias a los principios básicos de un actor y a las propiedades que se derivan de estos, proponemos además que cada actor (o grupo de ellos) represente la virtualización de un dispositivo físico, aumentando los recursos y las funcionalidades de tal dispositivo y aislando así los fallos de ejecución de los demás actores, es decir, del sistema. Realmente, esta propuesta no es novedosa pues existen diferentes referencias [38, 39] que proponen este mismo uso para el IoT o incluso implementaciones específicas. Nuestra intención es proveer de un marco de trabajo y de una implementación para la WoE que cumpla con sus expectativas. También es nuestro objetivo promover el uso del *Actor Model* pues provee de importantes abstracciones para desarrollar sistemas distribuidos, flexibles, auto cicatrizantes y diseñados para alcanzar un uso máximo y óptimo de recursos gracias a la computación paralela.

C. Prueba de concepto

Se han implementado prototipos de algunas de las capas descritas anteriormente. En concreto se ha implementado el Espacio de Objetos Virtuales, la capa de Abstracción de Protocolos y, para comunicar dos servidores remotos, se ha considerado usar un protocolo publicador/suscriptor para la capa de proxy. A continuación se detallan cada uno de ellos:

a) Espacio de Objetos Virtuales: Cada dispositivo físico es representado por un objeto virtual y éste a su vez por uno o más actores. Esta capa define la lógica y las abstracciones necesarias para que cada dispositivo pueda ser representado por uno o más actores, de tal forma que el flujo de datos, tomando como punto de partida la recepción de datos de un protocolo de los que consideramos heterogéneos o del IoT (MQTT, p.ej.), es transformado a un protocolo homogéneo (capa de proxy) para que sea entendible por las demás partes de la arquitectura.

b) Capa de Proxy: Según las necesidades de esta prueba de concepto, se ha implementado la capa de proxy mediante un protocolo publicador/subscritor. De esta forma, los objetos virtuales que representan a los dispositivos pueden publicar aquellos datos captados y pueden suscribirse a mensajes también recogidos por otros dispositivos o a mensajes de actuación enviados por otros dispositivos.

c) Capa de Abstracción de Protocolos: Contiene las implementaciones que hacen de interfaz entre diferentes protocolos como MQTT o HTTP.

En la Fig. 3 se puede observar la interacción entre las distintas capas, encontramos, desde arriba abajo, (i) la capa de aplicaciones con las funciones inteligentes de las Smart Grids, (ii) el middleware que se propone en este documento, (iii) la capa de accesibilidad al USN con los agregadores de múltiples sensores y, finalmente, (iv) las redes de sensores. Intencionadamente, este esquemático tiene un gran parecido a la Fig. 2 que se muestra en [4]; en este caso nos hemos centrado en el desarrollo del Middleware para una USN (Ubiquituous Sensor Network).

Existen tres secciones de la implementación bien diferenciadas:

(i) Sección MQTT: La conforman aquellos dispositivos que se comunican mediante el protocolo MQTT y los servidores o servicios encargados de traducir el protocolo MQTT hacia un protocolo "entendible" por las capas internas de la arquitectura.

protocolo "entendible" por las capas internas de la arquitectura.

Como servidor MQTT se ha usado Mosquitto [41]. Como implementación del Actor Model se ha usado

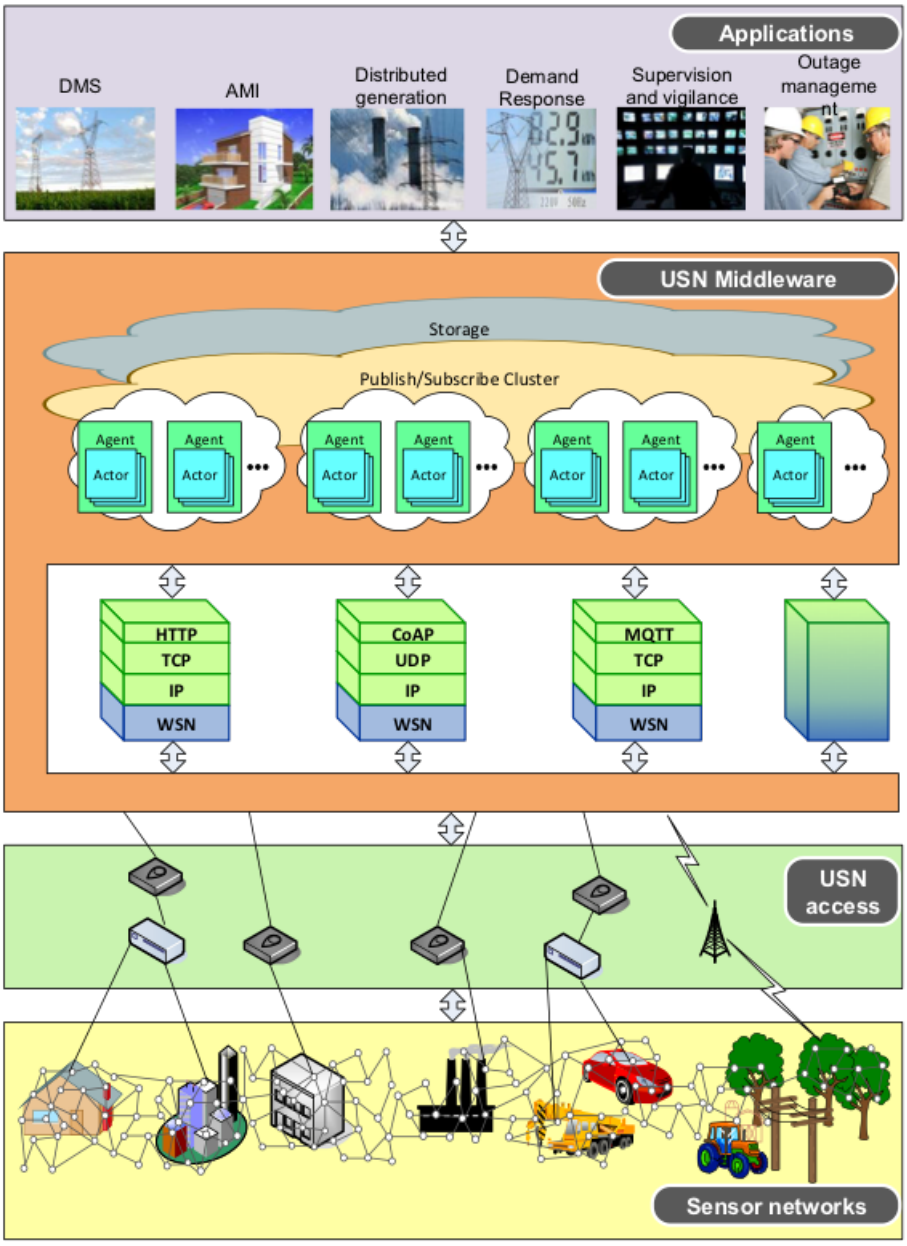


Fig. 3 Esquemático de las capas para un USN aplicado a las Smart Grids

(ii) Espacio de objetos virtuales y capa de proxy: Esta sección la conforman aquellos módulos encargados de representar cada dispositivo físico mediante un objeto virtual y el protocolo publicador/subscriptor.

(iii) HTTP/WebSocket: Análoga a la sección MQTT, la conforman aquellos dispositivos que se comunican con la arquitectura mediante protocolos web y los servicios encargados de traducir estos protocolos hacia un

Akka [39] a través de la API de Scala. Para la capa de proxy con la función de publicación/subscripción se ha usado una implementación con Akka-Cluster. Como servidor HTTP y WebSocket se ha usado Play Framework [42].

VII. CONCLUSIONES

En este artículo se usa el término Web of Energy para referirse a las características críticas que debe cumplir una arquitectura de la Web of Things para ser aplicada

al dominio de las Smart Grids. Se destacan sobre todo las necesidades de poder llevar a cabo funciones inteligentes, como son la distribución, resiliencia y auto cicatrización del sistema y la dificultad de renovar los sistemas tradicionales desplegados para la generación y gestión de energía, tanto a nivel de los dispositivos como a nivel de software, generalmente interrelacionados. Dados estos retos, se propone el uso del *Actor Model* para afrontarlos. Este paradigma está diseñado específicamente para realizar el modelado de sistemas concurrentes y distribuidos, aplicando así una mejora a las inherentes características de los sistemas distribuidos. En este sentido, se ha presentado en este artículo una propuesta de arquitectura *middleware* usando este paradigma para la creación de una red de sensores ubicua (USN). Esta propuesta se ha implementado mediante una prueba de concepto capaz de representar, de forma individualizada y virtual (y, por ende, con capacidad ilimitada de recursos) hasta 1000 dispositivos físicos simulados, con una representación visual web y a tiempo real.

AGRADECIMIENTOS

Parte de esta investigación ha estado financiada por la SUR de la DEC de la Generalitat de Cataluña y por Fondos Sociales Europeos 2017 FI_B 00583.

REFERENCIAS

[1] D. Guinard, "A Web of things application architecture: Integrating the real-world into the Web", Tesis doct., ETH Zurich, 2011.

[2] J. Navarro, A. Sancho-Asensio, A. Zaballos, V. Jiménez-Ruano, D. Vernet y J. E. Armendáriz-Iñigo, "The Management System of INTEGRIS", en Proceedings of the 4th International Conference on Cloud Computing and Services Science, SCITEPRESS-Science and Technology Publications, Lda, 2014, págs. 329-336.

[3] D. Vernet, A. Zaballos, R. Martin de Pozuelo y V. Caballero, "High performance web of things architecture for the smart grid domain", International Journal of Distributed Sensor Networks, vol. 2015, 2015.

[4] A. Zaballos, A. Vallejo y J. M. Selga, "Heterogeneous communication architecture for the smart grid", IEEE Network, vol. 25, n.º 5, 2011.

[5] D. Miorandi, S. Sicari, F. De Pellegrini e I. Chlamtac, "Internet of things: Vision, applications and research challenges", Ad Hoc Networks, vol. 10, n.º 7, págs. 1497-1516, 2012.

[6] V. Trifa, D. Guinard y D. Carrera, "Web Thing Model", W3C Member Submission, 2015.

[7] N. Shadbolt, T. Berners-Lee y W. Hall, "The semantic web revisited", IEEE intelligent systems, vol. 21, n.o 3, págs. 96-101, 2006.

[8] R. Martín de Pozuelo, A. Zaballos, J. Navarro y G. Corral, "Prototyping a Software Defined Utility," Energies, vol. 10, no. 6, p. 818, 2017.

[9] J. Iwata. (2012). "Making Markets: Smarter Planet", [Online]. Disponible: https://www.ibm.com/investor/events/investor0512/presentation/05_Smarter_Planet.pdf (visitado el 05/04/2017).

[10] "Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020", Gartner.com, 2017. [Online]. Disponible: http://www.gartner.com/newsroom/id/2636073. (visitado el: 05/04/2017).

[11] "Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015", Gartner.com, 2017. [Online]. Disponible: http://www.gartner.com/newsroom/id/3165317. (visitado el: 05/04/2017).

[12] R. T. Fielding, "Architectural styles and the design of network-based software architectures", Tesis doct., University of California, Irvine, 2000.

[13] M. V. Trifa, "Building blocks for a participatory web of things: devices, infrastructures, and programming frameworks", Tesis doct., ETH Zurich, 2011.

[14] Banks y R. Gupta, "MQTT Version 3.1.1", OASIS standard, 2014.

[15] Z. Shelby, K. Hartke y C. Bormann, "The constrained application protocol (CoAP)", 2014.

[16] V. Trifa, S. Wieland, D. Guinard y T. M. Bohnert, "Design and implementation of a gateway for web-based interaction and management of embedded devices", Submitted to DCOSS, págs. 1-14, 2009.

[17] "Enterprise IoT Solutions and Platform Technology", ThingWorx, 2017. [Online]. Disponible: https://www.thingworx.com/. (visitado el 05/04/2017). [17]

[18] "IBM Watson IoT - IoT Platform", Ibm.com, 2017. [Online]. Disponible: https://www.ibm.com/internet-of-things/platform/watson-iot-platform/. (visitado el 05/04/2017).

[19] "Octoblu | Integration of Everything", Octoblu.com, 2017. [Online]. Disponible: https://www.octoblu.com/. (visitado el: 05/04/2017).

[20] "EVRYTHNG IoT Smart Products Platform", EVRYTHNG IoT Smart Products Platform, 2017. [Online]. Disponible: https://evrythng.com/. (visitado el 05/04/2017).

[21] C. Bizer, T. Heath, K. Idehen y T. Berners-Lee, "Linked data on the web (LDOW2008)", en Proceedings of the 17th international conference on World Wide Web, ACM, 2008, págs. 1265-1266.

[22] W. W. W. Consortium y col., "RDF 1.1 concepts and abstract syntax", 2014.

[23] Adida, M. Birbeck, S. McCarron y S. Pemberton, "RDFa in XHTML: Syntax and processing", Recommendation, W3C, vol. 7, 2008.

[24] S. Bechhofer, "OWL: Web ontology language", en Encyclopedia of Database Systems, Springer, 2009, págs. 2008-2009.

[25] D. Guinard, M. Fischer y V. Trifa, "Sharing using social networks in a composable web of things", en Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on, IEEE, 2010, págs. 702-707.

[26] J. Gregorio y col., "The atom publishing protocol", inf. téc., 2007.

[27] "ThingWorxComposer™ - 1Worx", 1Worx, 2017. [Online]. Disponible: http://www.1worx.co/the-thingworx-platform/thingworx-composer/. (visitado el 18/04/2017).

[28] "Node-RED", Nodered.org, 2017. [Online]. Disponible: https://nodered.org/. (visitado el 05/04/2017).

[29] E. Dijk, A. Rahman, T. Fossati, S. Loreto y A. Castellani, "Guidelines for HTTP-to-CoAP Mapping Implementations", 2016.

[30] M. Collina, G. E. Corazza y A. Vanelli-Coralli, "Introducing the QEST broker: Scaling the IoT by bridging MQTT and REST", en Personal indoor and mobi- le radio communications (pimrc), 2012 ieee 23rd international symposium on, IEEE, 2012, págs. 36-41.

[31] "The 2016 Top Programming Languages", IEEE Spectrum: Technology, Engineering, and Science News, 2017. [Online]. Disponible: http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages. (visitado el 06/04/2017).

[32] "Usage Statistics and Market Share of Server-side Programming Languages for Websites, April 2017", W3techs.com, 2017. [Online]. Disponible: https://w3techs.com/technologies/overview/programming_language/all. (visitado el 21/04/2017)

[33] "reactphp/react", GitHub, 2017. [Online]. Disponible: https://github.com/reactphp/react. (visitado el 05/04/2017).

[34] "Get performance insight into the upcoming release of PHP 7", Zend.com, 2017. [Online]. Disponible: http://www.zend.com/en/resources/php7_infographic. (visitado el 05/04/2017).

[35] C. Hewitt, P. Bishop y R. Steiger, "Session 8 Formalisms for Artificial Intelligence A Universal Modular ACTOR Formalism for Artificial Intelligence", en Advance Papers of the Conference, Stanford Research Institute, vol. 3, 1973, pág. 235.

[36] H. Sutter, "The free lunch is over: A fundamental turn toward concurrency in software", Dr. Dobb's journal, vol. 30, n.o 3, págs. 202-210, 2005.

[37] D. D. Sanchez, R. S. Sherratt, P. Arias, F. Almenarez y A. Marin, “Enabling actor model for crowd sensing and IoT”, en Consumer Electronics (ISCE), 2015 IEEE International Symposium on, IEEE, 2015.
[38] R. Hiesgen, D. Charousset, T. C. Schmidt y M. Wa¨hlisch, “Programming Actors for the Internet of Things”, ERCIM NEWS, pág. 25, 2015.
[39] "Akka", Akka.io, 2017. [Online]. Disponible: http://akka.io/. (visitado el 05/04/2017).
[40] M. Asay, "How One Developer Set Out To Make The Internet Of Things Manageable - ReadWrite", ReadWrite, 2017. [Online]. Disponible: http://readwrite.com/2014/07/10/akka-jonas-boner-concurrency-distributed-computing-internet-of-things/. (visitado el 05/04/2017).
[41] "An Open Source MQTT v3.1 Broker", Mosquitto.org, 2017. [Online]. Disponible: https://mosquitto.org/. (visitado el 18/04/2017).
[42] "Play Framework - Build Modern & Scalable Web Apps with Java and Scala", Playframework.com, 2017. [Online]. Disponible: https://www.playframework.com/. (visitado el 18/04/2017).

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6570

A passive, non-intrusive, cheap method to identify behaviours and habits in the Campus

Javier Andión, José M. Navarro, Manuel Álvarez-Campana, Juan C. Dueñas
Departamento de Ingeniería de Sistemas Telemáticos,
Universidad Politécnica de Madrid
ETSI Telecomunicación, av. Complutense 30, 28040 Madrid
{j.andion,josemanuel.navarro,manuel.alvarez-campana,juancarlos.duenas}@upm.es

***Abstract*- Infrastructure usage discovery, positioning and analysis of behaviours of its users usually requires a collection of accurate and frequent positioning data. This paper shows how a network of inexpensive and non-intrusive sensors can serve to perform this kind of analysis by detecting devices with Wi-Fi connectivity.**

By this analysis, we show that, although individual tracking is not possible because of limitations of sensors, we can obtain the hours of use of the infrastructures, the occupation of the different areas at each moment and some of the most common users' behaviours.

***Keywords*- Wi-Fi sensors, MAC address, semantic locations, behaviour patterns discovery**

I. INTRODUCTION

Knowing the way users of a public infrastructure behave is a key to allocate the resources that must be assigned, ensure its safety, and control the usage. In this paper we propose a method, based on a minimal communication and computing infrastructure, to discover users individual and collective behaviour as regards usage of buildings.

Indoor location or pedestrian location has been a key research topic in last years [1]. Most works aim to discover the fine grain movements of people inside buildings, by using the mobile network, or personal area networks; these systems try to help the users to discover the path in a building, measure the length of stay in a mall for commercial purposes, or simply to ease people movements by removing obstacles, aside allowing for automatic movements of objects [2]. These works are based on location methods that make it possible to discover detailed paths in buildings, but they require either installing Bluetooth beacons or the cooperation of the mobile network antennas [3].

Based on the description of the region of interest, the concept of semantic trajectory has emerged as a key element to relate people trajectory with the activities they perform on it. In order to reason about people's habits this concept conveys more information than the pure trajectory, since the trajectory (series of points and times a people moves) is enhanced with labels marked as points of interests; even though the trajectory may not be so granular, thus getting to the concept of regions of interest. So, the semantic trajectory evolves to become the series of regions of interests visited with the time elapsed in each. We state that it is this information what counts, for example, to identify behaviours [4], and not much information is required for such purposes. This information can be obtained easily in a passive, non-intrusive way by using Wi-Fi probes emitted by autonomic antennas, apart of the data network giving service to the users, and providing most of users carry a smartphone with Wi-Fi capabilities.

We have applied this method to the analysis of user behaviour in campus; this is a public installation that anybody can use -up to a certain extent-, on which it is difficult to cover all places using cameras, expensive to cover with PAN beacons, and at the same time it is very interesting to know the movements of groups of students, the behaviour of the staff, the usage of shared areas (library, rest rooms, cantina, meeting rooms), the average number of people using the facilities and their non-regular usage. Key elements in this approach are: the minimal non-intrusive infrastructure required, the

small amount of information handled and its inaccuracy, and the simplicity of algorithms used to discover patterns of behaviour -mostly queries on the dataset.

II. APPLICATION SCENARIO

Universidad Politécnica de Madrid (UPM), as part of its City of the Future initiative, deployed a platform for experimentation composed by more than one hundred sensors able to perform real-time monitoring of 20 buildings in its Campus de Excelencia Internacional de Moncloa, gathering information on power consumption, environmental parameters, light, and buildings occupancy. The platform also includes real-time storing, processing, analysis and visualization of data. During a normal day, there are usually around 4000 people in the school: around 3000 students, at most 500 teachers and researchers and less than 500 administrators and maintenance staff.

Understanding how space and the installations were used by the students and the staff at the university was soon proposed as one of the key insights extracted from the data, and therefore the need to detect presence of people was of paramount importance. The research group in charge opted to building sensors able to detect Wi-Fi devices, also known as Wi-Fi tracking [5]; this solution has already used to analyze usage of public transportation (London underground nov 2016) or movements in public spaces such as airports. Then, they decided to develop cheap sensors for Wi-Fi tracking, based on Raspberry Pi boards with an external Wi-Fi module able to perform passive monitoring (Fig. 1. shows a photograph). The sensors are connected to power avoiding the usage of batteries. They read the header of radio IEEE 802.11 packages in its region of reach, and extract the MAC addresses of devices. As these MAC addresses are unique per device, counting them is a good indicator of the number of devices available on the sensor's surroundings, and they allow for temporal correlation analysis, thus obtaining useful information such as stay time, availability patterns, temporal patterns. All in all, and as MAC addresses are considered personal data under Spanish law (Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal), the system anonymizes data for analysis by an irreversible hash MD-5 function with salt applied by the sensors, which avoids brute-force attacks with pre-computed tables. Once anonymized, data is moved to a central server in the University by means of Ethernet connection -all buildings are cabled this way.

We have been provided data from one set of buildings equipped with 9 sensors, strategically allocated in certain points of interest. By means of Wi-Fi probes, each sensor scans its surroundings each minute in all Wi-Fi channels, so most of the Wi-Fi devices are captured, anonymized and stored in the sensor. For devices, we have checked that probe requests are sent between 20 seconds and 60 seconds period, depending on the type of smartphone. So, we were provided with log text files per sensor containing in each line: anonymized MAC address seen and its timestamp.

Fig. 1. Raspberry-pi Wi-Fi sensors

Our analysis is based on aggregating all the information in the log files of the sensors for a certain time span, counting identifiers in each sensor, identifiers in each time period, and then trajectories of each identifier, frequent trajectories, and behaviour profiles based on trajectories. One month renders: $n * 30 * 24 * 60 * ids$, where n is the number of sensors and ids is the average number of identifiers in our data. We processed them using the Spark processing libraries, whose execution speed eased the execution of many tests in this exploration exercise.

Specifically, we have used all the data collected during the whole month of 2016 May. This dataset is composed by 8.3M samples, where each sample is one register of one user seen by one sensor in one certain minute. Throughout the month 18K different devices are detected, this number reduces to 10K for devices seen more than 30 minutes in the month. On average 3.5K different devices are seen daily. These data were collected by 9 sensors:

1. Building A entrance ("Entr A"): this sensor is placed in the main entrance to the installations
2. Building A secondary entrance ("Entr A Sec"): this sensor is located in one of the secondary access to building A, it covers most of the classrooms of this building.
3. Work and study tables ("Std Tables"): this sensor is close to "Entr A", just above an area of tables where students gather to work and study in groups.
4. Library ("Library"): the sensor is inside the library, which is open every day from 9:00 to 21:00.
5. Building B entrance ("Entr B"): building B contains student's laboratories and offices.
6. Building B laboratories entrance ("Entr B Lab"): this sensor is placed in the secondary

entrance of this building closest to the laboratories.

7. Building B secondary entrance ("Entr B Sec"): this sensor is located in one of the secondary access to building B covering the most used classrooms of this building.
8. Building C entrance ("Entr C"): building C contains research laboratories and offices. Students do not have classes in this building.
9. Building D entrance ("Entr D"): this last sensor covers the entrance of building D and the library back.

Further on, Fig. 5 shows the positions of these sensors represented on the school map.

It is clear that results can only be approximations to the usage of space and infrastructure in the Campus, as there are many sources of error: not all people carry a mobile phone with its Wi-Fi capability turned on (students are suggested to switch off their mobile phones while at lectures or at library); on the contrary, some users carry more than one device; coverage area of sensors is severely conditioned by physical disposition of buildings; there are errors in capturing a mobile phone by the sensor; coverage areas overlap; sensors are put on a 2D map while devices move in a 2.5D space (at least two floors). But even under these limited conditions and applying simple algorithms we have been able to get a hint on people movements and behaviour, and to identify user types.

The main concepts in our study are defined now:

- Device position: we are not using power measurements, so the only valid approximation for the position of a single device in a given time is the point where the sensor is located. For devices seen by more than one sensor at once we have allocated to the sensor that saw for more timeslices.
- Regions or zones: we defined a zone per sensor, as the places where a device is detected by this sensor. Ultimately, if all sensors would get power enough the regions would define a Voronoi map, but as we do not know the effective reach of sensors we can only speculate on this.
- Timeslice: we are using 1-minute timeslices, as sensors are able to launch Wi-Fi probes at that pace. Later, for temporal analysis, we aggregate timeslices to create 1-hour sensing windows.
- Semantic location: sensors are located close to places where people actually do something (studying at the library, attending lectures at classrooms, performing experiments at labs, having lunch at the cantina, etc.). Particularly useful are sensors located in entry/exit places. So, once we know the sensor zone a device is in, we can infer the most likely location and annotate this with the most likely activity. In fact, sensors names reflect this: Library, Entr B Lab, Entr A Sec, Entr D, Entr A, Std Tables, Entr B Sec, Entr C, Entr B.
- Stay: if a device is seen by the same sensor, over a certain threshold (5 minutes, just enough to distinguish stays from transits) in a given time window, we conclude the user stays at the sensor zone, doing the activity in that semantic location.
- Path: for a device that is seen by a set of sensors without large interruptions, the path is defined as its sequence of stays.
- A frequent path is a path followed by many devices/users, in relation to the whole number of paths in the dataset.
- A device/user behaviour pattern is the set of frequent paths followed by many devices/users.

III. TEMPORAL ANALYSIS

First, we performed temporal analysis of the data. The purpose of this analysis is to find behaviours related to class schedules, work days, or hours of activity. To carry out this analysis, the first step is to aggregate the data so that it is easy to analyse its temporal behaviour. We did so, using the key (time, sensor), counting the number of users (Different) that were seen by said sensor during that hour and the total number of minutes in which a user is detected. This aggregation gives us a new set of data from which several conclusions can be drawn, observable both analytically and graphically.

Fig. 2 shows the number of different users seen by any sensor at a given time. From this figure, we can draw some obvious conclusions like that the activity in the school is much greater in the working days than in the weekends, or that at night there is no activity at all, but also, other less obvious conclusions can be extracted:

- The number of people in school is greater in the mornings than in the afternoons. This can be extracted by observing any particular day, the number of different users is composed of two peaks, with a valley in the middle that coincides with the lunch time. The first of these peaks reflects the number of people in the morning (whose maximum is around 11:00 am, half the morning) which is approximately 20% higher than the second peak, which reflects the number of people in the afternoon.
- Holidays, or days without lessons: On Monday 16th, the graph's behaviour is similar to a weekend, this is because that day was festive and there were no lessons, but the library remained open to students. Lower level activity days to 1st, 2nd, 16th (holidays), and 7, 8, 14, 15, 16, 21, 22, 28 and 29th for

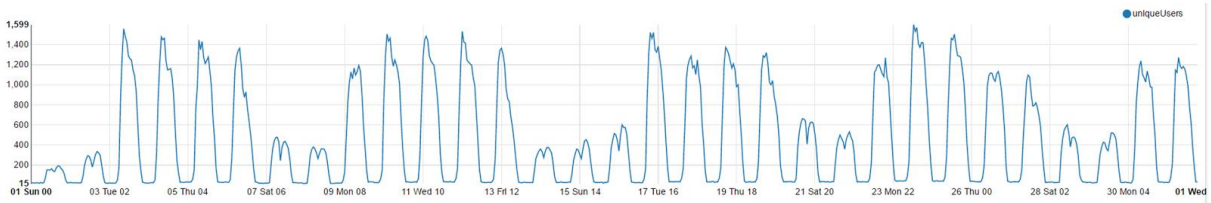


Fig. 2. Number of unique users per hour seen by any sensor

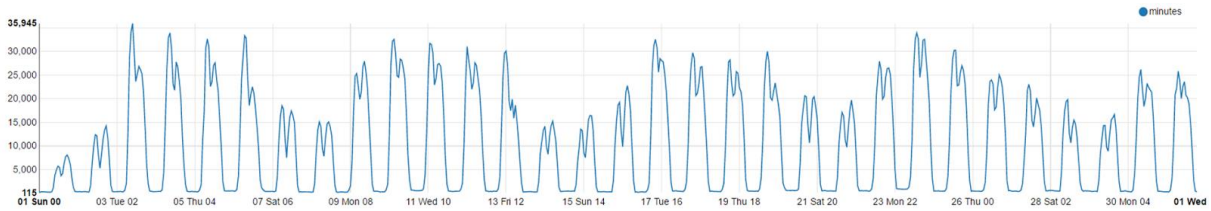


Fig. 3. Number of minutes accumulated per hour

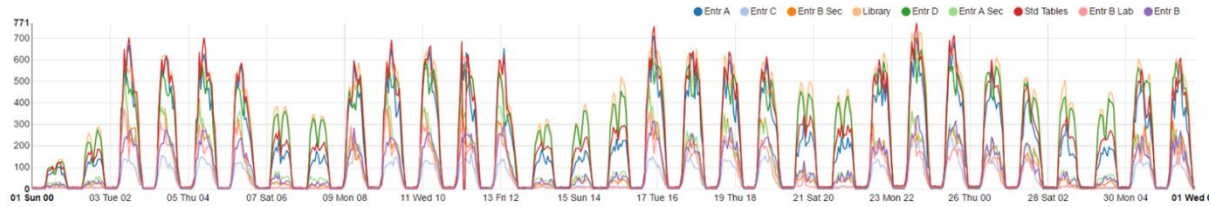


Fig. 4. Number of unique users per hour seen by any sensor

weekends. As an example of usage of library, we got 347, 336, 361, 377, 380, 354, 375, 284, 399, 349, 415 users for these days, rendering on average 361 users for a whole amount of 480 seats, which accounts for 75% of occupancy on weekends. It is also easy to see that whole activity on Mondays is lower than the rest of labour days but higher than weekends -which is due to examinations are done on Mondays.

Fig. 3 shows the number of minutes of activity accumulated in all sensors per hour. This chart allows us to distinguish between periods of transit and periods of stay. An example of this is the comparison between a weekend day and a weekday in the morning: while on weekends the number of users is much less than between daily, the relative difference between the number of accumulated minutes is not so big, this is because most of the weekend users are students in the library, who are standing for long periods of time, while on weekdays both teachers and students move through the school changing between classrooms. The same thing happens when comparing a weekday in the morning and in the afternoon, from which it can be deduced that students usually go to the library in the afternoon. Another behaviour that is observed is lunch time: at this point the graph drops to a trough, as people start to move around the school to go to eat at the cantina or outside the school.

Finally, in this analysis, Fig. 4 shows the same data as Fig. 2 but separating it by sensor. This figure, allows us to observe how users move throughout the day on an hourly basis. Observing the lines of the sensors "Entr A Sec" and "Entr B Sec", corresponding to the sensors located near the classrooms, it is observed as the number of users grows during lesson hours. In contrast, the "Std Tables" and "Entr A" sensors are maintained at the same level during all hours of daytime activity.

IV. SPATIAL ANALYSIS

The second approach is a spatial analysis. This time, we will try to find patterns related to how people are grouped in the studied area, using the positions of the sensors as an indication of the location of the users seen. It is not intended to perform an accurate trajectory analysis, but, an analysis of buildings and areas average occupation. In this case, the data are transformed by counting the number of occurrences of the key (time, sensor, user), where, now, time is just the hour corresponding to the timestamp of the sample, regardless of the day. With this transformation, we obtain a data set in which the number of minutes that, during the month studied, a user has been seen at a certain time in a certain place can be observed; e.g. how many times the user has been seen between 10 and 11 AM by the sensor in the library.

TABLE I
UNIQUE USERS PER SENSOR

Sensor	Unique users
Std Tables	9348
Entr A	8870
Library	8483
Entr D	7810
Entr A Sec	6770
Entr B	6329
Entr B Lab	3798
Entr C	2524

The first result allows to get an estimate of the profiles of the people detected by the sensors. Table I shows the total number of (different) users seen by each sensor throughout the month. By using the semantic location of the sensors, users can be classified by the activity that is performed in the coverage area of each sensor:

- Entr A: this is the main entrance; any kind of user can be observed here.
- Entr C: building C is composed by offices and investigation laboratories, so, most of the user seen here will be teachers or researchers.
- Entr B sec: the coverage of this sensor overlays the classrooms of building B, so users seen here will be both students and teachers.
- Entr A sec: this sensor is placed near the classrooms where the first-year courses take place, so first year students will be detected by this sensor.
- Library: the library is public access, students from this school and others will be seen here.

The second part of this analysis is based on the users' centroid analysis. This centroid is calculated as the average position of each sensor which saw the user weighted by the number of minutes seen during a defined time interval.

$$(1) \qquad C_u = \frac{\sum_{i=1}^{N} p_i * m_{u,i}}{\sum_{i=1}^{N} m_{u,i}}$$

Where C_u is the centroid of a single user, N is the number of sensors, p_i is the position of the i-th sensor and $m_{u,i}$ is the number of minutes which the i-th sensor dectected the user in the studied period of time. Then, these centroids are painted in a heat map over the school map to show the users concentration in each zone.

Fig. 5. Heatmaps of centroids since 8:00 to 10:00

(1) 8:00-9:00 (2) 9:00-10:00

Fig. 5 shows the heatmaps at 8 and 9 am, showing the transition between hourly intervals. This interval coincides with the start of daily activity, since the school is closed at night. The map on the left shows the users centroids between 8:00 and 9:00, "Entr A" is the sensor that registers more activity, since it is the main entrance and most of users access the area studied through this point. In the map to the right, which corresponds to the period between 9:00 and 10:00, it is clear that the number of users has been rapidly increasing. Four hot zones stand out:

1) "Entr A": users are still entering the school.
2) "Library": from the first hour, the library is full of students.
3) "Entr A Sec": this part of building A contains most of the classrooms. The morning schedule of classes is from 9 to 13, so the users that appear around this sensor will be the students who are there.
4) "Entr B Sec" and "Entr B Lab": like the above, these two sensors cover an area of classrooms, therefore, the heating of this area is also related to the time when lectures start.

Fig. 6 presents the heatmaps corresponding to the four hourly intervals which cover lunchtime, from 12:00 to 16:00. Lessons usually end at 13:00 (some end at 14:00), consequently, lunchtime is from 13:00 to 15:00 (the two central maps). The school cantina is located outside the area enclosed by the sensors. But, the path to it crosses the coverage of sensors "Entr A" and "Std Tables", therefore, people going toward or staying at the cantina will be seen by these two sensors.

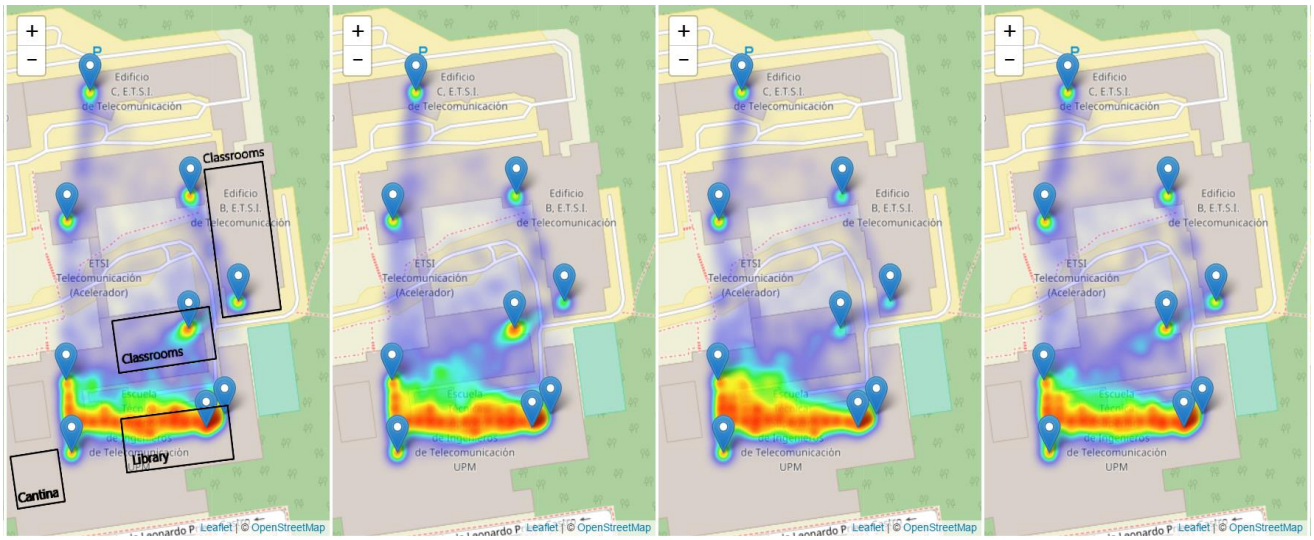


Fig. 6. Heatmaps of centroids since 12:00 to 16:00.
(1) 12:00-13:00 (2) 13:00-14:00 (3) 14:00-15:00 (4) 15:00-16:00

Observing the transition between the maps, it is evident how the user's centroids move towards the zone of the cantina (southwest part of the map), and between the 13:00 and the 15:00 the zone gets a higher activity than the normal one. It can also be seen how the sensor that covers the classrooms of building A, "Entr A Sec", has activity on the first two maps, on the third map, 15 to 16, it is practically turned off since there are no lectures at that time; starting at 16:00 it warms again, coinciding with the beginning of the afternoon class schedule.

V. BEHAVIOUR ANALYSIS

Finally, the third analysis performed is a behaviour analysis. On this occasion, the goal is to find behaviour patterns that describe the activities that are performed in the school and which are an indication to identify the different user profiles that can appear in the system. Final exams period begins at the end of May, so it can be expected (as can be inferred from previous analyses) that the most common behaviours will be students who stay long periods of time studying in the library.

To perform this analysis, we made a joining of sensors considering their semantic positions. The sensor "Entr D" covers the back of the library, and most of the users it detects are students in the library, being able to verify that because users detected by this sensor, in general, are also detected by the sensor "Library"; With this premise, the data of these two sensors are joined. The second join is made between the sensors "A" and "Std Tables", due to their proximity and that both cover part of the main hall, the cantina and the stairs (transit areas).

To discover each user's behaviour, data is partitioned by day and user. For each key (day, user) and each hour of the day along the activity hours, we extracted the sensor that has seen that user most of time. This renders the desired behaviour for that certain day and user. A total of 58311 behaviours are obtained, of which there are 18326 unique values, but analysing the most frequent it is detected that many of them are quite similar. Of the behaviours found, 42% of them (24488) only have activity in the morning, 35% (20717) only have activity in the afternoons, and 83% (48,608) are interrupted for at least one hour during mealtime.

Table II presents the 10 most frequent behaviours (the ones that are most repeated). Each row represtents one frequent behaviour over the observation time (working hours), columns show, in one-hour intervals, where users that follow this behaviour are. At first sight, almost all of them can be grouped group into two kinds: morning stays in the library and afternoon stays in the library.

Table II

TOP 10 MOST FREQUENT BEHAVIOURS

Index	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00
1								Library	Library	Library	Library	
2	Library	Library	Library	Library	Library							
3									Library	Library	Library	
4	Library											
5							Library	Library	Library	Library	Library	
6			Entr A									
7					Entr A							
8	Library	Library	Library	Library								
9						Entr A						
10		Library	Library	Library	Library							

Table III

TOP 10 MOST FREQUENT BEHAVIOURS FILTERING THE MOST COMMON SENSORS

Index	**9:00**	**10:00**	**11:00**	**12:00**	**13:00**	**14:00**	**15:00**	**16:00**
1	Entr Sec A	Entr Sec A						
2	Entr Sec A							
3	Entr Sec B							
4			Entr Sec A					
5			Entr Sec B					
6		Entr Sec A						
7					Entr Sec A			
8		Entr Sec B						
9							Entr Sec B	Entr Sec B
10	Entr Sec A	Entr Sec A	Entr Sec A	Entr Sec A				

The difference in activity registered by the sensors "Entr A", "Std Tables", "Library" and "Entr D", may hide the typical behaviours seen by other sensors. In order to prevent this, we filter previous result by searching only behaviours which not contains any of the most active sensors. The result of this filter is presented in Table III. The table, shows again the ten most frequent behaviours, this time with the named filter. It shows behaviours expected in a school: lecture attendance. Most subjects are imparted in lessons of two hours, starting at 9:00, 11:00, 15:00 and 17:00.

VI. CONCLUSIONS

Indoor positioning, path discovery and resource allocation have been a research area with great interest on recent years. Most of the current techniques require a precise and frequently updated position of the users. This paper proposes a simple approach to approximate to these problems using a very inexpensive infrastructure.

Semantic trajectories provide a method to get answers to some of the questions raised by these problems without the need of a precise user tracking. A simple and small network of low cost Wi-Fi sensors is enough to perform the analysis. The sensors scan every minute the MAC address of any device which has an active Wi-Fi sensor inside their coverage area. Accumulating these data for a medium period of time, one month in this analysis, results in a dataset with sufficient information to obtain some interesting conclusions.

Along our analysis' description we have proved that is possible to extract useful information about the operation of the school and about its users' behaviours, having a limited prior knowledge. Studying the temporal distribution of the number of people in the school we could distinguish between a weekday and a weekend or holiday, it is also quite easy to find out the installation working hours. We show how people from all around the area concentrate in the cantina during lunch time. Comparing the number of unique users seen by each sensor, and taking into account their semantic position, we could estimate the library occupancy during weekends. Finally, we discovered the massive use, during all day, of the library during the month before the final exams.

In future works, we will include data from a longer observation period and increase our scope including sensors located in another schools. Increasing our dataset both temporal and spatially will allow us to discover new behaviour patterns, confirm our observations and apply our methodology to new scenarios.

REFERENCES

[1] Czogalla, O., & Naumann, S. (2016, October). Pedestrian indoor navigation for complex public facilities. In Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on (pp. 1-8). IEEE.

[2] Zheng, Z., Chen, Y., Chen, S., Sun, L., & Chen, D. Location-aware POI Recommendation for Indoor Space by Exploiting WiFi Logs.

[3] A. Kurkcu and K. Ozbay, "Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2644, pp. 72–82, Jan. 2017.

[4] Y. Xu, I. D. G. Groeneveld, R. Sulzer, E. Theocharous, O. T. Willems, and M. S. Tryfona, "Determine activity based on the classified identity of users by using Wi-Fi monitoring," *Geomatics Synth. Gr. Proj. Rep.*, 2016

[5] Kopytoff, V. Stores Sniff Out Smartphones to Follow Shoppers. [Online] MIT Technology Review, November 12, 2013.

[6] Duque Domingo, J., Cerrada, C., Valero, E., & Cerrada, J. A. (2016). Indoor Positioning System Using Depth Maps and Wireless Networks. Journal of Sensors, 2016.

[7] Zhang, X., Kim, G. B., Xia, Y., & Bae, H. Y. (2012, August). Human activity recognition with trajectory data in multi-floor indoor environment. In International Conference on Rough Sets and Knowledge Technology (pp. 257-266). Springer Berlin Heidelberg.

[8] Schauer, L., Marcus, P., & Linnhoff-Popien, C. (2016, October). Towards feasible Wi-Fi based indoor tracking systems using probabilistic methods. In Indoor Positioning and Indoor

Navigation (IPIN), 2016 International Conference on (pp. 1-8). IEEE.

[9] Bowman, G., & Jaebker, K. (2013). Using commodity hardware as an affordable means to track on-site visitor flow. In Museums and the Web.

[10] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., ... & Theodoridis, Y. (2013). Semantic trajectories modeling and analysis. ACM Computing Surveys (CSUR), 45(4), 42.

[11] Guo, S., Xiong, H., Zheng, X., & Zhou, Y. (2017). Activity Recognition and Semantic Description for Indoor Mobile Localization. Sensors, 17(3), 649.

[12] Ying, J. J. C., Lee, W. C., Weng, T. C., & Tseng, V. S. (2011, November). Semantic trajectory mining for location prediction. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 34-43). ACM.

[13] Deb, B., & Basu, P. (2015, February). Discovering latent semantic structure in human mobility traces. In European Conference on Wireless Sensor Networks (pp. 84-103). Springer International Publishing.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6623

Diseño de una red de sensores para monitorizar una instalación acuícola

Javier Rocher, Lorena Parra, Miran Taha, Jaime Lloret

Instituto de Investigación para la Gestión Integrada de zonas Costeras.

Universidad Politécnica de Valencia, Spain

jarocmo@alumni.upv.es, loparbo@doctor.upv.es, miab2@doctor.upv.es, jlloret@dcom.upv.es

***Resumen*- En las instalaciones acuícolas la monitorización de la calidad del agua es fundamental para la automatización de los procesos. En este artículo presentamos una red de sensores que realizan medidas de la turbidez y la temperatura en todos los tanques. Como nodo se ha empleado un Flyport que manda los datos a un servidor y cuenta con una serie de alarmas programadas. Se ha diseñado la topología de red y física atendiendo a la estructura típica de estas instalaciones. Se ha estudiado el rendimiento de la red en distintos escenarios. Se ha establecido que el número máximo de Flyports por punto de acceso antes de devaluar la calidad de la conexión es de 5 Flyports, con una tasa de paquetes perdidos cercana al 0.5% y una tasa de paquetes por segundo media de 86.47.**

***Palabras Clave*- red inalámbrica; nodo sensor; calidad del agua; acuicultura; Flyport**

I. INTRODUCCIÓN

En la actualidad se está produciendo una sobreexplotación de los recursos pesqueros a nivel mundial [1]. Debido a esto, en los últimos años se ha producido una disminución de las capturas de pescado. No se están reduciendo los medios para realizar dichas capturas, sino que se están aumentando los medios para el aumento de las capturas [2]. Con el futuro aumento de la población y por ende el aumento de consumo de alimentos entre ellos el pescado, se debe asegurar una mayor cantidad de alimentos. Como no se puede extraer más peces del mar, es necesario aumentar la cantidad de peces criados mediante la acuicultura [3]. Está se puede desarrollar en mar abierto o en instalaciones en tierra. En las instalaciones en tierra el agua recibida pasa inicialmente a un tanque de recepción. En ese tanque de recepción el agua permanece cierto periodo de tiempo tras el cual es distribuida a los tanques de producción.

La ventaja de las instalaciones en tierra, es poder controlar el agua de los tanques de producción. Esto es importante, pues la calidad del agua puede afectar negativamente al rendimiento bioenergético de los peces. Los sólidos suspendidos tienen efectos abrasivos, reducen la visión y producen problemas en las agallas [4, 5]. La temperatura también se debe controlar debido a que una temperatura alta supone una bajada del oxígeno disuelto en el agua [6]. Además, la cantidad de alimento necesario para peces depende de la temperatura [7, 8]. El estudio de la calidad del agua es un problema que ha sido abordado por muchos autores [9 - 17]. Una de las principales limitaciones de las redes de sensores es el consumo energético [18].

Para monitorizar las variables físico-químicas que puedan causar efectos negativos en los peces, se están utilizando redes de sensores inalámbricos (WSN). Los sensores pueden colocarse en los tanques. Cuando algún parámetro del agua no es adecuado se realizan acciones correctivas. Sin embargo, puede resultar más interesante detectar de forma anticipada las entradas de agua con problemas. Si monitorizamos la calidad del agua en el tanque de recepción podremos aislar determinados tanques de producción. Puede ser interesante aislar tanques con peces más sensibles o por estar aplicando algún tratamiento especial.

En este artículo, presentamos el diseño de una red de sensores para monitorizar la calidad de agua de una piscifactoría en tierra firme. El sistema se basa en sensores que se ponen en una caja estanca. Los sensores empleados son de temperatura y de turbidez. La caja está atravesada por un tubo de vidrio por donde pasa el agua. Además, se instala un sensor de humedad dentro de la caja estanca para detectar una posible entrada de agua dentro de la caja. Cada grupo de 3 sensores estará conectado a un Flyport que se conectará a un punto de

acceso (PA) mediante una conexión WiFi. Los distintos PA de la planta se conectarán a un Switch con una salida a internet mediante un Router. Los datos recogidos por los Flyport son enviados a un servidor y serán accesibles tanto en local como en remoto.

El resto de este artículo se estructura de la siguiente forma. En la sección 2, se explican algunos trabajos relacionados con las redes de sensores WiFi y su uso para la monitorización de parámetros del agua. En la sección 3, se explican las partes de nuestro sistema, tanto los sensores utilizados como la topología. El rendimiento de la WSN se muestra en la sección 4. Por último, en la sección 5 mostraremos las conclusiones de nuestro trabajo.

II. Estado del arte

En esta sección se muestra el estado del arte. La utilización de redes de sensores Wireless es muy utilizada en la actualidad. Bri et al. [19] presentó los usos de las redes de sensores WiFi como son el monitoreo de la salud humana, la industria el medio ambiente, aplicaciones militares etc.

Un ejemplo de la utilización de redes de sensores para la agricultura de precisión lo presenta Sendra et al. [10] en su artículo de sensores utilizados para la protección de ovejas y cabras de ataques de lobos. Este sistema funciona midiendo la temperatura corporal y la frecuencia cardiaca de las ovejas (o cabras) y en caso de que se produzca un ataque, estos parámetros cambiaran. Lo que enviará una alarma y se tomaran las medidas oportunas para detener el ataque. Respecto a las redes de sensores para monitorizar el agua O'Flynn et al. [11] nos explica el proyecto Smartcoast un proyecto que tiene como objetivo el desarrollo redes de sensores Wireless que permitan observar los datos de los sensores de forma remota. El sistema se basa en sensores "Plug and play" que permite la integración de sensores con interfaz "Transducer Electronic datasheet". Estos sensores utilizaran un sistema de comunicación basado en Zigbee. Los resultados del estudio indicaron que era viable la utilización de estas redes con un bajo consumo eléctrico.

Estas redes de sensores se pueden instalar sobre boyas para el monitoreo en el mar como hizo Sendra et al. [12]. Los autores desarrollaron un sistema con sensores low cost para controlar las áreas con praderas de Posidonia. Estos sensores fueron montados sobre una boya y controlaban diferentes parámetros del agua que son la salinidad, turbidez, la presencia de hidrocarburos y temperatura además de medir parámetros meteorológicos. La boya cuenta con un sistema de paneles solares con una batería para alimentar todo el circuito eléctrico. Los datos obtenidos son procesados por un microcontrolador y enviados mediante WiFi al servidor central. Desde el servidor central son enviados mediante internet al móvil o pc de la persona encargada de la vigilancia.

Estas redes de sensores se pueden utilizar para el monitoreo de las mareas Parra et al. [13] diseñaron una red de sensores WiFi para el monitoreo del movimiento del agua de las mareas en los estuarios. Esta información era útil para entender los cambios de flora y fauna en un estuario. El sistema propuesto se basa en un nodo con un sensor de salinidad basado en dos bobinas de cobre. Cada nodo envía la información a un servidor donde se almacena la información.

Otros ejemplos de la monitorización de agua son los presentados por, Simbeye et al. [14] y por Rasin y Abdullah [15]. Ambos, propusieron en sus artículos una red de sensores para la monitorización de distintos parámetros del agua. Estos dos sistemas constan de nodos de sensores, nodos de coordinación y un pc. Los nodos sensores controlan los distintos parámetros y transmiten la información al nodo coordinador utilizando los protocolos Zigbee y esta información es enviada al pc donde se mostrará de forma visual.

La utilización de software y hardware Open Souce se puede utilizar para estas redes como hizo Rao et al. [16], que presentaron un sistema de sensores inalámbrico de detección de parámetros físico-químicos de bajo coste que funcionan de forma autónoma. Los diferentes sensores se conectaron a un Arduino Mega 2530. Estos sensores incluían la temperatura, el pH y el contenido en oxígeno. Según el autor, con la correcta calibración, se puede establecer un sistema de monitoreo confiable. Con una mayor resolución espacial que los actuales sistemas. Lo que ayudará a comprender el comportamiento de los seres acuáticos.

Por último, Santoshkumar and Hiremath [17] propuso un sistema basado en la medición de pH, salinidad y temperatura para la monitorización de la acuicultura utilizando un microcontrolador Arduino. Zigbee fue utilizado como protocolo de comunicación.

El sistema que proponemos se diferencia del resto en que no solo evaluamos la calidad del agua en los tanques de producción, sino también en la entrada del agua. En caso de que se produjera una entrada de contaminantes la red sería capaz de detectar esta entrada y se podrían aislar ciertos tanques de producción del circuito general.

III. Propuesta

En esta sección mostramos la RDI empleada en la monitorización. Inicialmente, mostramos el diseño, desarrollo y calibrado del sensor de turbidez. Más adelante, se presenta el despliegue y localización del sensor en el tanque de acuicultura. También se muestra en esta sección el nodo inalámbrico utilizado. Finalmente mostramos la topología empleada.

A. Desarrollo del sensor de turbidez

A continuación detallamos el diseño y calibrado del sensor óptico empleado para monitorizar la turbidez. Para crear este sensor se ha usado un emisor de luz infrarrojo (IR). El sensor está basado en el diseño presentado por Sendra et al. [12]. Como emisor de IR se ha empleado un LED IR con una longitud de onda pico de 850nm. Como detector se ha utilizado un fotodiodo sensible a IR con un rango de sensibilidad desde 790 a 1050nm. El emisor se ha establecido a

6.5cm del receptor en un Angulo de 180º. De los 6.5cm que los separan 2.7 son ocupados por un canal de agua.

Con el fin de calibrar el sensor, se han utilizado muestras con distinta turbidez. Estas muestras están compuestas por agua y sedimento de tamaño limo. El motivo por el que hemos utilizado sedimento tan fino es debido a que en condiciones normales las arenas sedimentan en el tanque de recepción en la instalación acuícola. Solo el material más fino puede llegar a los tanques de producción. Para calibrar el sensor 5 muestras de distinta turbidez se han generado La muestra con menor turbidez contenía 0mg/L de limos. La muestra con mayor turbidez contenía 378.55mg/L de limos. Todas las muestras fueron homogenizadas antes de realizar las medidas. El modelo matemático que relaciona la respuesta del fotodiodo (MΩ) con la turbidez (mg/L) y los datos del calibrado se ven en la Fig. 1 se detalla en la ecuación (1). Este modelo presenta un coeficiente de correlación de 0.999 con un error absoluto medio de 3.87mg/L. En la ecuación (1) *Turb* representa la turbidez en mg/L, mientras que, *IR* hace referencia a la respuesta del fotodiodo en (MΩ).

Tras la calibración, se ha llevado a cabo un proceso de verificación. Para este proceso, dos nuevas muestras fueron generadas de forma aleatoria dentro de los valores máximo y mínimo del calibrado. Se ha sustituido el valor de IR y se han obtenido los valores de turbidez teóricos según la ecuación (1). En la Tabla 1 podemos ver los resultados del proceso de verificación. El error absoluto máximo es de 4.76mg/l y el error relativo máximo de 4.37%. Por último se ha utilizado un circuito de acondicionamiento de la señal del fotodiodo que es alimentado a 9V para recibir la señal en el Flyport. Tras aplicar el nuevo circuito a equivalencia entre la Señal recibido por el Flyport y la turbidez del agua se muestra en la Ecuación (2)

$$Turb.\left(\frac{mg}{L}\right) = 377.36 + \frac{81.14}{\mathrm{IR\,(M\Omega)} - 3.90} - 46.66 * \mathrm{IR(M\Omega)} \quad (1)$$

$$Turb.\left(\frac{mg}{L}\right) = 408 + \frac{30.9}{Señal^2(V) - Señal(V)} - 214V \quad (2)$$

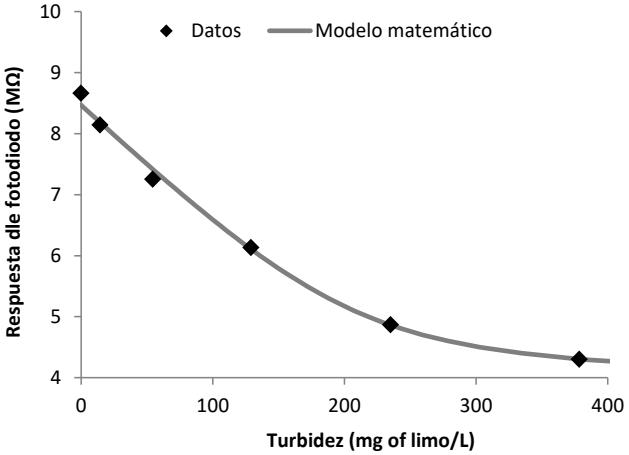


Fig. 1. Calibrado del sensor de turbidez

B. Nodo utilizado

En esta subsección, el nodo utilizado en la WSN se presenta. El nodo elegido es el módulo Flyport con es USB Nest (Ver Fig. 2). Está basado en la plataforma openPicus con código abierto.

Incluye un procesador de 16 Bits del PIC24FJ256, con 256K de flash y 16K of RAM. El Flyport es capaz de trabajar con el estándar 802.11 b/g/n. El tamaño del nodo es de 35x48x7mm y un peso de 11g. Puede proveer de alimentación para los sensores a 5 y a 3.3V. El principal motivo por el que elegimos este nodo es su gran flexibilidad y la gran cantidad de entradas y salidas disponibles. La posibilidad de combinar entradas analógicas y digitales es importante para futuras aplicaciones. Por último, la oportunidad para programar diferentes aplicaciones es crucial.

C. Implementación de la WSN

En esta subsección, el despliegue, localización y aislamiento de los sensores se va a mostrar. En primer lugar, se detallarán las labores realizadas para asegurar el aislamiento de los sensores y el módulo Flyport del agua. Posteriormente se presentará la localización de los nodos en la instalación acuícola.

Teniendo en cuenta que nuestro objetivo es monitorizar la turbidez del agua, será necesario depositar los sensores en el agua. Para ello debemos asegurarnos de lograr la estanqueidad del receptáculo para los sensores y el Flyport. Como la profundidad a las que estarán los sensores será inferior al metro de agua, se utilizarán juntas tóricas. Hemos empleado una caja estanca de termoplástico.

La caja tiene unas dimensiones de 17,5x11,5x7cm. Se ha modificado el cierre de la caja y hemos logrado que siga siendo estanca a 1.5m de profundidad en un tanque de acuicultura. Antes de esta comprobación se ha realizado dos aberturas para pasar una tubería de vidrio de 2.7cm de diámetro que permitirá el paso de agua para tomar las medidas de turbidez. Para sellar la unión entre la caja y la tubería de vidrio se emplea una silicona especial para soportar la humedad y la presión.

Tabla I. VERIFICACIÓN DEL CALIBRADO

Turbidez (mg/L)	Respuesta del fotodiodo (MΩ)	Turbidez calculada (mg/L)	Error absoluto (mg/L)	Error relativo (%)
108.86	6.36	113.62	4.76	4.372
373.14	4.32	369.78	-3.36	-0.90

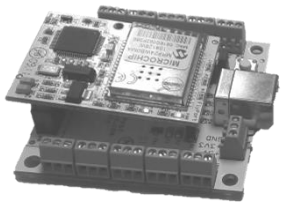


Fig. 2. Flyport module

Tres sensores se depositan en la caja protectora. En primer lugar, el sensor de turbidez que se sitúa en la parte donde se ha colocado la tubería de vidrio. El segundo sensor que utilizamos, es un sensor de temperatura basado en un diodo Zener de 2 terminales. Este sensor está previamente calibrado y tiene un error de 1ºC. El sensor de temperatura se sitúa pegado a la tubería de vidrio. Se ha elegido esta ubicación ya que el vidrio transmite la temperatura mejor que el plástico. El motivo por el que hemos usado este sensor es debido a que la temperatura es también un parámetro crucial para el rendimiento bioenergético de los peces.

El último sensor utilizado, es un sensor de humedad. Este sensor se pondrá en la parte más baja de la caja. La función de este sensor será detectar de forma temprana una posible entrada de agua en la caja estanca. Para alimentar los tres sensores se utiliza una batería de 9V mientras que para alimentar el Flyport se utiliza una batería externa de 5V.

Los nodos se distribuirán en distintos puntos de la instalación, tanto en los tanques de producción como en el de recepción. En los tanques de producción se ubicará un nodo por tanque. En el de recepción se situarán dos nodos, a la entrada y a la salida de agua.

D. Arquitectura

En esta sección se muestra la topología utilizada en el desarrollo propuesto.

La topología de red utilizada en este caso está basada en la topología de estrella extendida. Los dispositivos Flyport se conectan a un punto de acceso (PA) con una conexión WiFi. Los PA se conectan a su vez a un Switch de capa 2 que tiene salida a internet a través de un Router. La topología de red puede verse en la Figura 3. Existen dos PA a los que se conectan los Flyports y un tercer PA al que se conectan otro tipo de dispositivos. Entre esos dispositivos se encuentran ordenadores, dispositivos móviles y un servidor. En el servidor se recoge toda la información de los sensores. Los diferentes dispositivos podrán conectarse al servidor para visualizar la información de los sensores. Está información podrá ser consultada tanto en local como en remoto.

La topología física está condicionada por la estructura de la instalación acuícola. En este artículo, se ha utilizado la estructura habitual de la mayoría de instalaciones acuícolas en tierra sin recirculación. Estas infraestructuras cuentan con un gran tanque de recepción de agua que habitualmente está en un nivel superior al que están los tanques de producción. Además tienen uno o más salas utilizadas como oficinas. Las condiciones habituales de estas instalaciones incluyen una alta humedad y grandes cantidades de agua fluyendo, que en algunos casos es agua salada. Estas condiciones se dan en la zona del tanque de recepción y en la zona de los tanques de producción. En ambas zonas es importante minimizar la cantidad de equipos expuestos. Por este motivo, los equipos como el Switch y el Router se localizarán en las oficinas. En las zonas del tanque de recepción y los tanques de producción solamente se dejarán los PA necesarios. Esta topología puede verse en la Figura 4.

En la entrada y salida de cada tanque existen llaves de paso para cortar el suministro de agua. Nuestra propuesta consiste en monitorizar los niveles de turbidez en el agua de los tanques y en caso de detectarse una subida de turbidez en el tanque de recepción cerrar la entrada de agua en los tanques en los que sea necesario. No todos los taques contienen los mismos peces, ni todos los peces tienen las mismas necesidades.

Por otro lado, si el sensor de humedad registra un valor mayor al calor establecido como umbral (50% de humedad relativa), accionará un sistema para cortar la alimentación de todos los sistemas de esa caja. De esta forma se minimizarán los posibles daños causados por la humedad sobre los sistemas electrónicos.

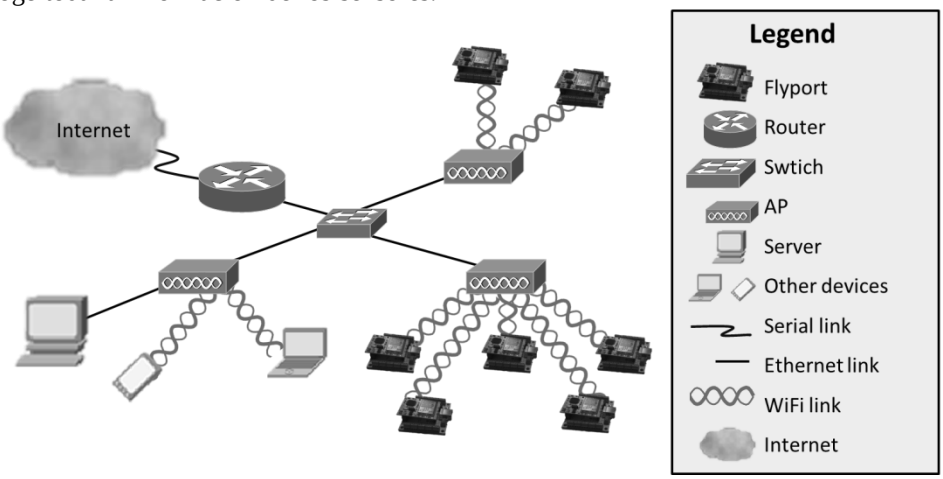


Fig. 3. Topología de red basada en estrella extendida

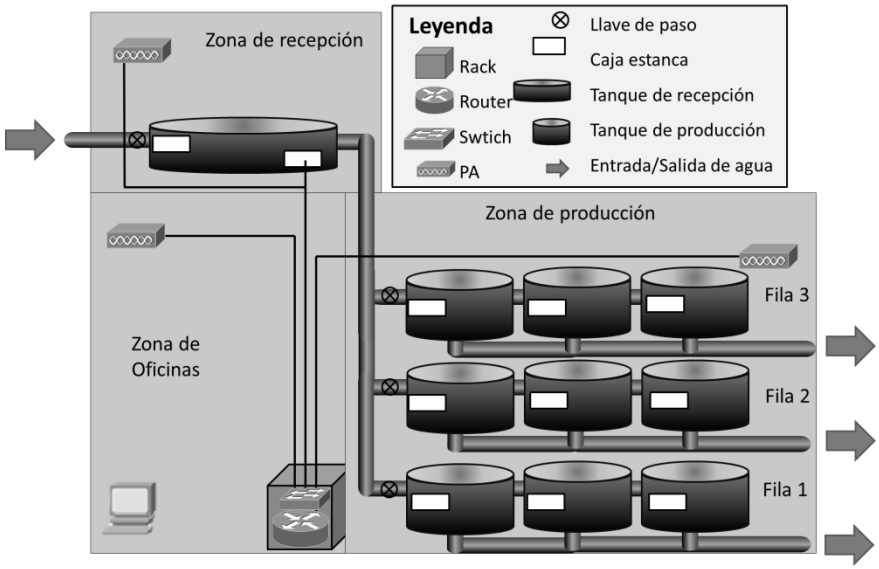


Fig. 4. Topología física adaptada a la infraestructura típica de las instalaciones acuícolas en tierra sin recirculación de agua

Finalmente, se detalla el algoritmo que regula el funcionamiento del sistema (Ver Fig. 5). En un primer momento, se definen las filas de tanques (*TFi*) como Fila 1 (*TF1*), Fila 2 (*TF2*) y Fila 3 (*TF3*). Atendiendo a que cada fila tiene unos requerimientos mínimos de calidad de agua.

Una vez el Flyport recibe las señales de los sensores identifica la entrada como *Turbidez*, *Humedad* o *Temperatura*. Tras lo cual, los convierte en valores digitales y transmite los datos al servidor y a los usuarios por HTTP. Solo los valores de *Turbidez* y *Temperatura* son enviados. El dato de *Humedad* se utiliza solo como un mecanismo de seguridad para el Flyport. Si el dato de *Humedad* es mayor a un valor umbral (50% de humedad relativa) se accionara un proceso. En este proceso se enviará una alerta al personal de mantenimiento y se iniciará el apagado de emergencia para evitar daños electrónicos. En esa alarma se indica el tanque en el que se ha detectado la avería y un operario sustituye la caja estanca con todos los componentes por otra nueva. La caja que ha provocado la alarma se lleva a mantenimiento para comprobar si ha habido daños o no y volver a poner la caja en funcionamiento.

Por otro lado, si el nivel de *Turbidez* supera los umbrales establecidos de calidad de agua se procederá a detener el flujo de agua a los tanques. De forma previa, el responsable de producción establecerá para cada fila o grupo de filas cual es el valor umbral. Estos valores dependen de la especie cultivada y el estadio de desarrollo de los peces. En el momento en que una fila de tanques se quede aislada del tanque de recepción se pone en marcha un reloj. Este reloj maracá el tiempo que lleva el tanque aislado, la información servirá para el responsable de producción. El responsable podrá tomar la decisión de volver a iniciar el flujo de agua aunque la calidad siga siendo inferior a la deseada con el fin de evitar situaciones de baja concentración de oxígeno en los tanques. Así mismo el aviso de que una fila de tanques ha quedado aislada llega también a los operarios que deberán prestar mayor atención a los tanques aislados.

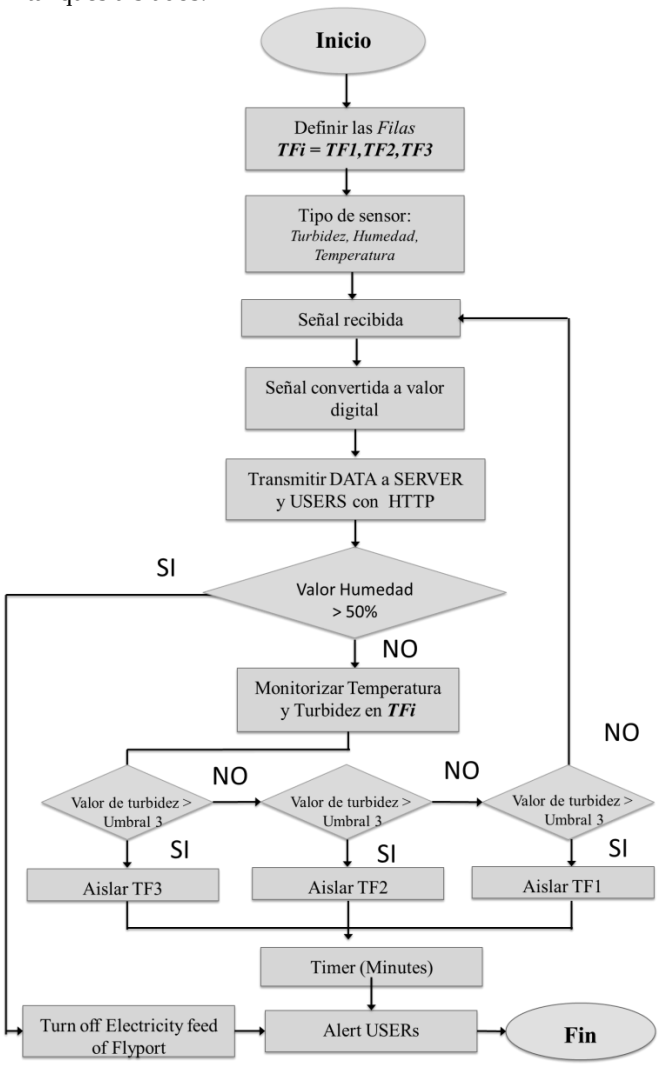


Fig. 5. Funcionamiento del sistema

IV. Rendimiento de la WSN

En esta sección, vamos a analizar el rendimiento de la WSN. Nuestra mayor preocupación es asegurar un buen rendimiento de la red, proporcionando una baja tasa de paquetes perdidos. Sin embargo, necesitamos minimizar el número de PA desplegados en las zonas de los tanques. Para ello, vamos a realizar una serie de pruebas conectando distintas cantidades de Flyports a un único PA y analizaremos los parámetros de la red

Los parámetros de red estudiados incluyen el número de paquetes por segundo, la tasa de paquetes perdidos y los paquetes reenviados. Se han estudiado escenarios con distintas cantidades de Flyports conectados a un mismo PA. Los escenarios incluyen desde un solo Flyport por PA hasta 10 Flyports en cada PA. Primero, se analizan los resultados de los paquetes por segundo transmitidos en los distintos escenarios (Ver Fig. 6). Podemos observar como a medida que aumenta el número de Flyports conectados aumenta la cantidad de paquetes por segundo que son recibidos en el PA. La media de paquetes enviados por segundo con 3, 5 y 10 Flyports en un periodo de 60s es de 70.33, 86.47 y 95.42 pps respectivamente. La desviación típica para 3, 5 y 10 nodos es de 6.28, 4.55 y 3.43. Esto muestra como a menor cantidad de Flyports mayores son las fluctuaciones de tráfico, como podemos observar en la Fig. 6. Para tener más información de estas fluctuaciones analizamos los valores máximos y mínimos de paquetes por segundo trasmitidos en cada escenario. Con una configuración de 3 Flyports el valor máximos alcanzado es de 80 pps mientras que el mínimo es de 60 pps. Los datos para 5 Flyports y 10 Flyports son de 80 y 95 pps y de 90 y 100 pps respectivamente.

A continuación, se procede a analizar la información obtenida de la tasa de paquetes perdidos en un intervalo de 60s. En este caso, las configuraciones utilizadas han ido desde solo 1 Flyport por PA hasta 10 Flyports por PA. El resultado aparece en la Fig. 7. Podemos observar que con 2 Flyports o menos no hay paquetes perdidos, entre 3 y 7 Flyports la tasa de paquetes perdidos es inferior o igual al 1%. A partir de 7 Flyports la tasa de paquetes perdidos se dispara llegando al 3.5% con 10 Flyports.

Por último, como el Flyport utiliza TCP para las transmisiones es capaz de reenviar los paquetes que han sido perdidos. La información sobre los paquetes reenviados se muestra en la Fig. 8. En este caso mostramos la información para 3, 5 y 10 Flyports. Se muestran los paquetes retransmitidos para intervalos de tiempo de 10 s. Se puede observar que para 3 Flyports solo en 2 intervalos es necesario reenviar paquetes y en ambos casos el número de paquetes retransmitidos es 1. Para 5 Flyports en 4 de los 6 intervalos de tiempo es necesario reenviar paquetes. La cantidad de paquetes retransmitidos va de 1 a 3. Finalmente, para la configuración con 10 Flyports, en 4 intervalos se hace necesario el reenvío de paquetes. A diferencia de la configuración de 5 Flyports en este caso el número de paquetes reenviados varía entre 2 y 5. En este caso, podemos afirmar que a mayor cantidad de Flyports mayor es la tasa de paquetes reenviados y mayor es la variación en el número de paquetes reenviados por unidad de tiempo. En términos globales, el escenario con 3 Flyports ha producido el reenvío de 2 paquetes en un intervalo de 60 s, mientras que para 5 y 10 Flyports los valores son de 7 y 14 paquetes.

Podemos concluir que un máximo de 5 Flyports podrían ser conectados a un único PA sin que ello suponga una elevada tasa de paquetes perdidos, cercana al 0.5%. Tendremos una tasa de paquetes por segundo media de 86.47 pps con una desviación estándar de 4.55.

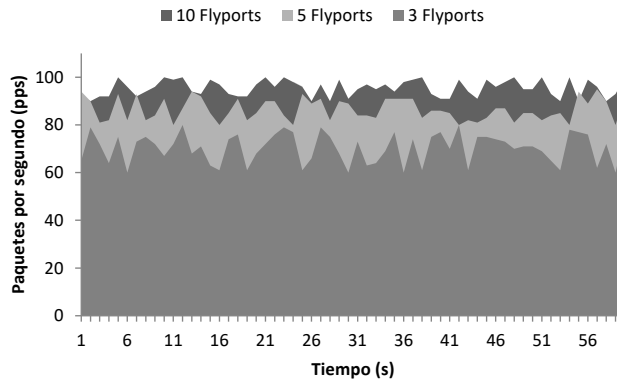


Fig. 6. Transmisión de paquetes por segundo en distintos escenarios

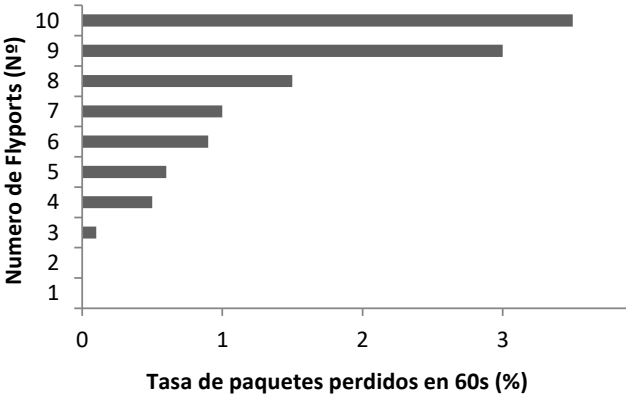


Fig. 7. Número de paquetes perdidos en distintos escenarios

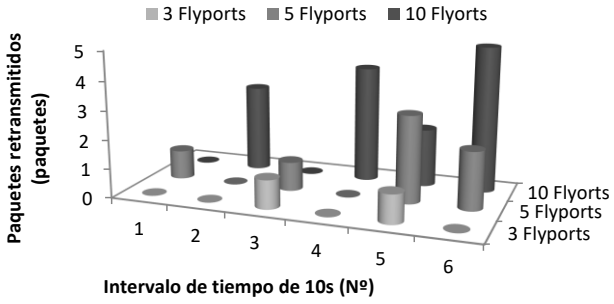


Fig. 8. Número de paquetes retransmitidos en distintos escenarios

V. Conclusiones

En este artículo hemos mostrado el diseño de una red de sensores para monitorizar una instalación acuícola con un sistema abierto. Hemos mostrado el sensor que se ha desarrollado para monitorizar turbidez

como parámetro indicador de la calidad del agua. Hemos testeado una caja estanca que contendrá el Flyport y los tres sensores que utiliza nuestro sistema. Se ha diseñado la topología física y de red atendiendo a la estructura típica de las instalaciones acuícolas. Los nodos mandarán la información al servidor y generará alarmas cuando sea necesario. Y hemos evaluado el rendimiento de la red WiFi entre el PA y os Flyports con el fin de decidir cuantos Flyports podemos conectar a un mismo PA sin perder calidad en el rendimiento. Se ha tratado de minimizar el número de PA necesarios debido a las condiciones de humedad y corrosión que se da en las zonas de los tanques.

Como trabajos futuros, diseñaremos el servidor que recibirá y mostrará los datos recogidos por los sensores. Además, pretendemos dotar al servidor de seguridad, requiriendo de autenticación para acceder a la información. Además, se incluirá un cuarto sensor que obtendrá información de la salinidad. Se utilizará un sensor basado en el que se desarrolló en [20].

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado con un contrato pre-doctoral del programa Ayudas para contratos predoctorales de Formación del Profesorado Universitario FPU (Convocatoria 2014), con referencia FPU14/02953 del Ministerio de Educación, Cultura y Deporte.

REFERENCIAS

[1] R. Watson and U. R. Sumaila, “Food security implications of global marine catch losses due to overfishing,” *Journal of Bioeconomics*, vol. 12, no 3, pp. 183–200, 2010.

[2] G. Pontecorvo and W. E. Schrank, “The continued decline in the world catch of marine fish,” *Mar. Policy*, vol. 44, pp. 117–119, 2014.

[3] J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, and C. Toulmin, “Food security: the challenge of feeding 9 billion people,” *Science* vol. 327, no. February, 2010.

[4] M. . Bruton, “The Effects of Fishing on Fish Habitat,” *Hydrobiologia*, vol. 125(1), pp. 221–240, 1985.

[5] D. H. Wilber and D. G. Clarke, “Biological Effects of Suspended Sediments: A Review of Suspended Sediment Impacts on Fish and Shellfish with Relation to Dredging Activities in Estuaries,” *North Am. J. Fish. Manag.*, vol. 21, no. October 2012, pp. 855–875, 2001.

[6] S. M. Greig, D. A. Sear, and P. A. Carling, “A review of factors influencing the availability of dissolved oxygen to incubating salmonid embryos,” vol. 334, no. May 2006, pp. 323–334, 2007.

[7] F. J. Sa, “Self-feeding of European sea bass (Dicentrarchus labrax , L .) under laboratory and farming conditions using a string sensor,” *Hydrological processes* vol. 233, pp. 393–403, 2004.

[8] S. O. Handeland, A. K. Imsland, and S. O. Stefansson, “The effect of temperature and fish size on growth, feed intake, food conversion efficiency and stomach evacuation rate of Atlantic salmon post-smolts,” *Aquaculture* vol. 283, pp. 36–42, 2008.

[9] H. Khaleeq, A. Abou-elnour, and M. Tarique, “A Reliable Wireless System for Water Quality Monitoring and Level Control,” *Network Protocols and Algorithms* vol. 8, no. 3, pp. 1–14, 2016.

[10] S. Sendra, F. Llario, L. Parra, and J. Lloret, “Smart Wireless Sensor Network to Detect and Protect Sheep and Goats to Wolf Attacks,” *Recent Adv. Commun. Netw. Technol.*, vol. 2, no. 2, pp. 91–101, 2013.

[11] B. O’Flynn *et al.*, “SmartCoast A Wireless Sensor Network for Water Quality Monitoring B,” *32nd IEEE Conf. Local Comput. Networks SmartCoast*, pp. 815–816, 2007.

[12] S. Sendra, L. Parra, J. Lloret, and J. M. Jiménez, “Oceanographic multisensor buoy based on low cost sensors for posidonia meadows monitoring in mediterranean sea,” *J. Sensors*, vol. 2015, 2015.

[13] L. Parra, E. Karampelas, S. Sendra, J. Lloret, and J. J. P. C. Rodrigues, “Design and deployment of a smart system for data gathering in estuaries using wireless sensor networks,” *2015 Int. Conf. Comput. Inf. Telecommun. Syst.*, pp. 1–5, 2015.

[14] D. S. Simbeye, J. Zhao, and S. Yang, “Design and deployment of wireless sensor networks for aquaculture monitoring and control based on virtual instruments,” *Comput. Electron. Agric.*, vol. 102, pp. 31–42, 2014.

[15] Z. Rasin and M. R. Abdullah, “Water Quality Monitoring System Using Zigbee Based Wireless Sensor Network,” *Int. J. Eng. Technol. IJET*, vol. 9, pp. 24–28, 2009.

[16] A. S. Rao, S. Marshall, J. Gubbi, M. Palaniswami, R. Sinnott, and V. Pettigrovet, “Design of low-cost autonomous water quality monitoring system,” *Proc. 2013 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2013*, pp. 14–19, 2013.

[17] Santoshkumar and V. Hiremath, “Design and Development of Wireless Sensor Network System to Monitor Parameters Influencing Freshwater Fishes,” *Int. J. Comput. Sci. Eng.*, vol. 4, no. 6, pp. 1096–1103, 2012.

[18] R. Azizi, “Consumption of Energy and Routing Protocols in Wireless Sensor Network,” *Network Protocols and Algorithms* vol. 8, no. 3, pp. 76–87, 2016.

[19] D. Bri, M. Garcia, J. Lloret, and P. Dini, “Real deployments of wireless sensor networks,” *Proc. - 2009 3rd Int. Conf. Sens. Technol. Appl. SENSORCOMM 2009*, pp. 415–423, 2009.

[20] L. Parra, S. Sendra, J. Lloret, I. Bosh, " Development of a conductivity sensor for monitoring groundwater resources to optimize water management in smart city environments," *Sensors*, vol.15, no. 9, pp.20990-21015, 2015

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6631

Red de Sensores Inalámbricos de Bajo Consumo Energético en Agricultura Hidropónica

Carlos Cambra[1], Sandra Sendra[1,2], José Miguel Jiménez[2], Jaime Lloret[2]

[1] Instituto de Investigación para la Gestión Integrada de zonas Costeras. Universitat Politècnica de València, Valencia

C/Paraninf, 1. 46730. Grao de Gandia (Valencia)

[2] Departamento de teoria de la señal, telemática y comunicaciones, ETS Ingenierías Informática y de Telecomunicación. Universidad de Granada

C/ Periodista Daniel Saucedo Aranda, s/n. E-18071 Granada

carcamb1@doctor.upv.es, ssendra@ugr.es, jojiher@dcom.upv.es, jlloret@dcom.upv.es

***Resumen*- Las nuevas tecnologías son la clave para mejorar la sostenibilidad en el sector agrícola y producir alimentos de calidad. Además, el uso de nuevas técnicas de cultivo, como la hidroponía, basada en la producción de alimentos en agua sin necesidad de cultivarlos en tierra, hace que se puedan analizar mejor las necesidades de la planta y proporcionarle las mejores condiciones de crecimiento. En este artículo presentamos el desarrollo de una red de sensores móviles orientada a monitorizar los patrones de necesidades de las plantas y tomar decisiones inteligentes según la captación de datos ambientales obtenida. La red está compuesta por nodos sensores comunicados con transceptores de radio distribuidos en una red mallada, que podría ser fácilmente adaptada a cualquier tipo de uso a petición del profesional. Esta red ha sido probada en un entorno de agricultura hidropónica. Finalmente el artículo muestra los resultados obtenidos en cuanto a tráfico generado, lo que nos permitirá en un futuro, hacer la red escalable.**

***Palabras Clave*- Smart farming, invernaderos inteligentes, red de sensores en horticultura, agricultura de precisión, sensores hidropónicos.**

I. INTRODUCCIÓN

La incorporación de la tecnología al sector agroalimentario está tomando un papel muy importante en la competitividad de los productores agrícolas. Hoy en día, el mercado demanda productos de mayor calidad y buena presentación por lo que los sistemas de producción se renuevan constantemente.

La cosecha de cultivos en invernadero requiere una captación de datos en diferentes puntos, debido a la naturaleza espacial distribuida, donde las condiciones climáticas pueden variar dependiendo del punto de localización.

El uso de redes de sensores inalámbricas (del inglés, *Wireless Sensor Networks*, *WSN*) se está extendiendo de forma exponencial en el sector agrícola [1], ya que se basan en dispositivos inalámbricos, de bajo consumo energético y económico, necesarios en entornos donde se tienen distancias muy extensas y sin posibilidad de acceder a la red eléctrica para suplir al sistema.

Para desarrollar una red de monitorización en instalaciones agrícolas, como por ejemplo, tipo invernadero, podemos decidir entre emplear una infraestructura de comunicaciones inalámbricas, donde la cualificación técnica del personal encargado de la implementación de los dispositivos y su mantenimiento es elevada [2]. No obstante, podemos decantarnos por el uso de redes sin infraestructura que permita la movilidad de dispositivos sin perjudicar las comunicaciones entre ellos y de fácil instalación.

En nuestra propuesta se ha optado por el uso de una tecnología de bajo coste que permite multiplicar los nodos en un sistema, o simplemente desecharlos ante cualquier anomalía de funcionamiento. Se ha optado por el use de los módulos Nordic RF24 que operan en la banda libre en Europa de 2,4 GHz. El bajo coste económico en la tecnología seleccionada de Nordic, nos

permite aumentar el número de nodos con sensores. Esto supone un aumento de la cantidad de información recogida y la obtención de una mayor cantidad de resultados a partir de los cuales se pueden tomar esas decisiones inteligentes que precisamos para gestionar los entornos monitorizados. Un ejemplo claro de las mejoras de esta es que hace 5 años, únicamente se usaban un conjunto de sensores unitarios de varios tipos (de costes económicos elevados) para todo el invernadero [4], obteniendo un solo dato cada variable (temperatura, humedad, radiación solar, etc.) y se generalizaba para toda la infraestructura cultivable. Con el paso de los años, la tecnología se ha abaratado y han surgido nuevos protocolos de comunicación más eficientes.

Por último, esta información hay que procesarla y mostrarla para darle difusión entre las redes sociales de profesionales. Este es uno de los principales motivos por los que los desarrollos con sensores de este trabajo tengan como finalidad, el envío de datos e integración con la plataforma multimedia PLATEM PA (Plataforma Tecnologica Multimedia en agricultura de precisión) [5], diseñada para crear una interfaz de usuario con reglas inteligentes de decisión en la que pueden interactuar diferentes profesionales del sector agrícola con la finalidad de intercambiar opiniones y resolver problemas con una inmediata actuación en remoto.

PLATEM PA permite la interacción usuario – sistema inteligente (riego, invernadero, actuadores, etc.) y facilita el manejo de los cultivos gracias a su motor y reglas de inteligencia, por otro lado se genera un feedback donde usuarios de la plataforma intercambian opiniones y experiencias.

En este artículo se presenta una red de sensores de bajo coste energético para el uso en el sector de la agricultura hidropónica. El sistema se basa en una red de sensores que monitoriza de forma continua un invernadero. Los datos ambientales, calidad de agua y estado de la planta se almacenan en un servidor (la nueve)donde los agricultores y profesionales del sector pueden ver y usar. Al mismo tiempo, también son usados por un nodo coordinador para controlar los diferentes actuadores como dosificadores de abono, luces, válvulas de agua, etc., que actuarán sobre los propios cultivos. La interfaz gráfica y de control que nos permite acceder a estos datos es la propia plataforma PLATEM PA. La red se ha probado en términos de tráfico generado por los nodos en diferentes condiciones.

El resto del trabajo queda estructurado del siguiente modo. En la Sección II se presentan los trabajos realizados en materia de redes de sensores y tecnología de comunicaciones inalámbricas en varios sectores, especialmente en agricultura. La sección III describe el problema del control del clima del invernadero en cultivos hidropónicos. Posteriormente, se analiza el desarrollo de un sistema inteligente de control basado en eventos y la WSN para el control de fertirrigación y condiciones ambientales correctas en relación a temperatura de efecto invernadero, calidad del agua y nutrientes y radiación solar. Los dispositivos usados y desarrollo del frameworks de comunicación de nodos son descritos en la sección IV. En la sección V se detalla los resultados de los test del sistema. Finalmente, las conclusiones y trabajos futuros se exponen en la sección VI.

II. TRABAJOS RELACIONADOS

Las redes de sensores se caracterizan por su fácil adaptación a diferentes ámbitos de trabajo, como los sectores de la industria, agricultura, entornos forestales, marinos, militares, etc. Pero existen sectores donde este tipo de sistemas de sensores deben ser instalados en áreas donde no existe corriente eléctrica para alimentarlos. En este caso surge la necesidad de proponer nuevas posibilidades en el diseño de dispositivos capaces de realizar trasmisiones de datos más polivalentes en este medio y con menor consumo energético. Como nos muestra S. Sendra et al. [6] se ha realizado un estudio de red de sensores dedicada a la detección y verificación de incendios en los bosques. Esta red se caracteriza por el uso de cámaras IP creando una red de monitorización conectada entre sí, que es capaz de orientar las cámaras hacia el punto de detección de anomalías mediante sensores con el objetivo de verificar la presencia del incendio.

C. Cambra et al. [7] desarrolló una red de comunicaciones bajo el protocolo 800.11ab destinada a la detección de alpinistas en zonas rurales y alta montaña que pudieran necesitar algún tipo de ayuda. La propuesta se basa en un sistema de aeronave no tripulado capaz de crear una red de comunicaciones en zonas de alta montaña dedicada a la búsqueda y control de senderistas y montañeros.

Existen trabajos más focalizados en el campo de la agricultura de precisión, donde se desarrollaron redes de sensores y tele-gestión de programadores de riego. Este es el caso del trabajo presentado por C. Cambra et al. [5] donde podemos ver una red de comunicaciones inalámbricas de bajo consumo basada en el uso de transmisores de radio en la banda ISM de 868MHz para crear un sistema inteligente de riego a través de comunicaciones, programadores y sensores de bajo consumo energético en agricultura extensiva al aire libre.

Existen otros proyectos de monitorización de variables agronómicas en invernaderos a través de redes de sensores y comparativas de diferentes tecnologías dentro de las redes 802.15.4 que han tenido buena aceptación como la presentada por H. Ibayashi [8] donde expone el uso de redes funcionando a una frecuencia de 429 MHz y a 2,4GHz en Japón mostrando menores tasas de error el uso de frecuencias en 429MHz o Sensorscope en el que G. Barrenetxea nos presenta un sistema de monitorización ambiental usando nodos de comunicación en la banda 868 MHz [9], son proyectos que emplean los estándares IEEE 802.15.4, para la creación de WSN.

Tabla I
PROTOCOLOS DE COMUNICACIONES INALAMBRICAS

Protocolo	RF24	WiFi 802.11	LoRa WAN
Frecuencia	2.4 GHz	2.4GHz	868/433MHz
Tamño Paq.	32 Bytes	1480Bytes	256Bytes
Cobertura	50m-1 km	200m-10km	1km-20km
Consumo Idle/ Rx/Tx	20uA/8mA/ 8,4mA	10uA/60mA/ 215mA	2,8mA/38,9mA/ 50mA
Ancho de Banda	256Kbps/ 2Mbps	150/300Mbps	250bps/5Kbps
Precio	0,75$	7$	15$

Como muestran las diferentes tecnologías inalámbricas (Ver Tabla I), RF24 es la unica que nos oferce una solución compacta con muy poco consumo energético y con un coste claramente inferior a los demás, se observa una diferencia entre los consumos energéticos del uso de un transceptor bajo WiFi al uso de un RF24, con el que podemos asegurar las comunicaciones durante un tiempo de vida de batería bastante extenso. No es tanta la diferencia entre RF24 y un transceptor LoRa, pero si podemos observar que el coste económico de un dispositivo LoRa es superior al coste de RF24, donde en muchos proyectos con un gran número de nodos, puede suponer un incremento inasumible. La tecnología RF24 es la que mejor se adapta al planteamiento inicial y por ello las comparativas realizadas con tecnologías estudiadas en anteriores trabajos, muestran un gran potencial de la red RF24.

El trabajo presentado se centra en el uso del módulo RF basado en el chip Nordic nRF24L01, es ultra compacto y de muy bajo consumo. Trabaja en la banda ISM a una frecuencias de 2.4GHz (frecuencia libre) con modulación GFSK (Modulación por desplazamiento de frecuencia Gausiana) y es ideal para proyectos de telemetría, control de periféricos, industria y afines. Incorpora un transceptor RF de 2.4GHz, un sintetizador RF, algoritmos de control de errores y un acelerador para trabajar con interfaz SPI. En un futuro se verá el avance de LoRa y la disminución de costes de sus módulos, pero actualmente Nordic RF24 puede ser una alternativa muy buena y barata en redes de sensores agrícolas, demandadas por los profesionales del campo.

III. Sistema Inteligente de Manejo de Cultivos Hidropónicos

Esta sección expone los problemas actuales que los cultivos hidropónicos presentan y muestra cómo nuestra propuesta los puede solucionar. El cultivo hidropónico es un método de cultivar plantas usando agua y minerales en vez de tierra. La principal ventaja es el control total del alimento de la planta y la no necesidad de disposición de superficie cultivable, ya que la hidroponía suele expandirse en vertical.

A. Problema de control en los cultivos hidropónicos

El crecimiento del cultivo está influenciado principalmente por las variables climáticas ambientales que rodean a los cultivos y por la cantidad de agua y fertilizantes suministrados por el riego. Esta es la razón principal por la que un invernadero es ideal para el cultivo, ya que constituye un entorno cerrado en el que las variables climáticas y fertirrigación pueden ser controladas para permitir un óptimo crecimiento y desarrollo del cultivo [10]. El clima y la fertirrigación son dos sistemas independientes con diferentes problemas de control. A priori, se conocen los requisitos de agua y nutrientes de diferentes especies de cultivos y, de hecho, los primeros sistemas automatizados fueron los que controlaban estas variables estáticas. Una simplificación a través de la automatización inteligente consiste en suponer que las plantas reciban la cantidad de agua y fertilizantes que requieren en cada momento, según los datos que retornan.

B. Nuestra Propuesta

Aunque el sistema se compone de varias partes, en este trabajo nos centramos en la implementación del sistema de recogida de parámetros, que conlleva la configuración de la red, monitorización de parámetros, algoritmos de decisión y control de dispositivos actuadores. En la Fig. 1 se describe la parte del sistema relacionada a nuestro estudio.

Tanto los actuadores como la trasmisión de datos de los dispositivos con sensores se realizan de manera inalámbrica. En este caso se ha optado por el uso del transceptor NRF24L01 de la compañía Nordic semiconductor [11]. El chip de comunicaciones elegido es un Atmel de 8 bits, que se caracteriza por su bajo consumo y podemos encontrarlo en un formato comercial ya ensamblado sobre una placa electrónica con entradas para sensores, sockets de conexión del transceptor de radio y porta pilas [12].

Podemos encontrar algunas propuestas y recomendaciones de uso del transceptor NRF24L01 pero en todas ellas, únicamente se habla de la parte de la trasmisión de datos desde un nodo a su coordinador. En trabajos anteriores con el transceptor NRF24L01, el patrón del protocolo de comunicaciones se basaba en envíos tx desde el nodo a la red mesh para que llegara finalmente al coordinador, estas redes se desarrollan principalmente en redes de sensores para él envió de variables eventuales. En este trabajo, la red incluye una serie de actuadores móviles que necesitan recibir una serie de parámetros para su correcto funcionamiento. En la implementación de una red mesh se tiene una tabla de enrutamiento en ambos sentidos que puede ralentizar el sistema de comunicaciones. Como novedad en nuestra propuesta de red mesh, se ha diseñado un sistema de envío de mensajes basado en los identificadores de los nodos, en lugar de emplear en una tabla de enrutamiento prefijada. Este proceso es similar al tipo de envío que se realiza en un broadcasting, con la peculiaridad de generar un buffer o pila sobre el mensaje. Además, para realizar la trasmisión solo es necesario conocer el identificador del destino y se va generando un buffer con los identificadores recorridos con un modo de reintentos, si fuera necesario.

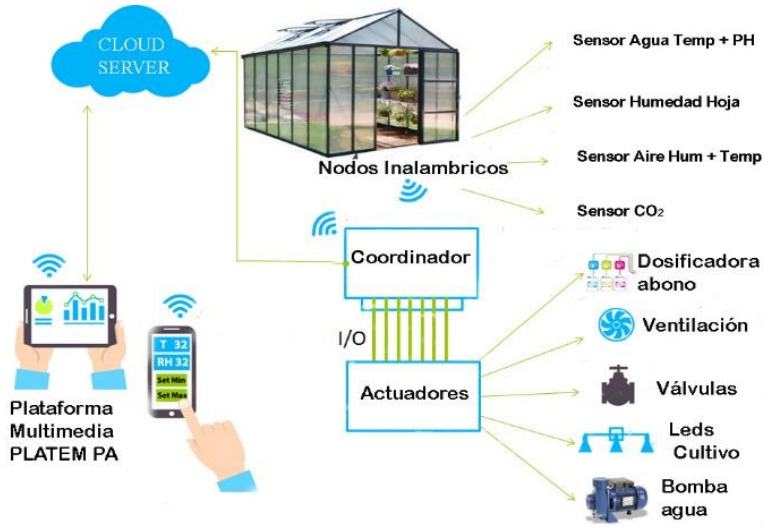


Fig. 1. Esquema de dispositivos del Sistema.

Como podemos ver, tanto la comunicación con las principales ventajas de esta configuración son las siguientes:

- Los nodos no necesitan tablas de enrutamiento para definir el camino más corto.
- Cada nodo usa una configuración de retardo (algoritmo de retardo según identificador) al envío de los mensajes, para evitar colisiones.
- Se pueden insertar nodos nuevos en la red sin la necesidad de modificar ni añadir parámetros identificativos a una tabla de encaminamiento.
- Los nodos se pueden mover, sin perder comunicación, a nivel de una tabla de enrutamiento, supone generar una nueva ruta más óptima.
- No requiere de un nodo central para la gestión de las tablas de enrutamiento.

Creemos que esta propuesta supone un gran avance para el sector, ya que las redes de sensores y en nuestro caso la descrita en este trabajo, va a poder ser usada por agricultores y profesionales del campo que quiere tecnificar sus sistemas de cultivo, pero no desean costear soporte técnico cualificado. Además la inclusión de un nuevo modo no implicará que tengan que modificar parámetros en el sistema hidropónico, es decir, nos encontramos frente a un sistema sencillo y de fácil adaptación al entorno de trabajo.

IV. Arquitectura del Sistema

En esta sección se describe los diferentes dispositivos usados como nodos con sensores, coordinadores de red y actuadores. A nivel de comunicaciones se ha implementado un framework basado en comunicaciones sin infraestructura con encaminamiento multisalto lo que permite una movilidad de nodos dentro de la red mallada WSN, existen diferentes trabajos relacionados con las arquitecturas de protocolo [13] que nos proporcionan la base de las funciones de encaminamiento.

A. Dispositivos

Los dispositivos usados en la red son nodos inalámbricos, con un consumo de energía ultra bajo, capaces de transmitir la información de los sensores conectados. La Fig. 2 muestra nos nodos inalámbricos empleados. Una de los motivos por los que se ha elegido este dispositivo, además de por su bajo consumo, es su pequeño tamaño (Dimensiones 30,5 x 67,0 mm). Los sensores están conectados a los nodos, de forma muy sencilla a través de los sockets, dedicados a la toma de diferentes datos necesarios para la toma de decisiones de funcionamiento del sistema de cultivo y análisis del crecimiento del cultivo. Por un lado, como muestra la Fig. 3 incorporamos sensores para el control de los sistemas de producción para comprobar su correcto funcionamiento, como puedan ser nivel de flujo de agua, nivel de caudal o detección de agua en un punto. Otro tipo de sensores integrados en el sistema son los dedicados a la medición de parámetros de calidad y niveles de elementos nutricionales de las plantas, los encargados de determinar las actuaciones de aporte nutricional, luz, aire, temperatura, etc. En la Fig. 4 se muestra algunos ejemplos de sensores usados como humedad de la hoja, nivel de pH y temperatura del agua.

En cuanto al nodo base (Ver Fig. 5), este se considera el punto central dedicado a recoger la información de la red de sensores y comandar operaciones a los actuadores añadidos al sistema de cultivo, este a su vez está conectado a internet a través de la red móvil con tecnología 3G y mantiene comunicación bidireccional con la Plataforma multimedia PLATEM PA, que es la encargada de almacenar todos los datos, crear patrones de notificaciones o eventos y tomar imágenes, gracias a una pequeña cámara.

El protocolo de comunicaciones seleccionado está orientado al uso de la red de radio en la frecuencia de 2,4 GHz en modo mallado o Mesh donde los identificadores de los nodos son identificadores únicos y es una red dinámica ya que es posible que los

dispositivos entren y salgan de la red. Debido a esta estructura, es sencillo para cualquier nodo comunicarse con el nodo maestro.

Una cosa a tener en cuenta es la naturaleza dinámica del protocolo RF24Mesh, y la necesidad de verificar la conectividad a la red. Para los nodos que están transmitiendo constantemente es conveniente comprobar la conexión y/o renovar la dirección cuando la conectividad falla cada pocos segundos. Los nodos que no están transmitiendo activamente, deben configurarse para probar su conexión a intervalos predefinidos, para permitir que se vuelvan a conectar según sea necesario. En el caso de nodos en modo *Sleep* sólo estarán en línea temporalmente, por tanto, es adecuado liberar la dirección antes de desconectarse y solicitar una dirección al despertar.

Este protocolo es capaz de transmitir cargas útiles sin que la red devuelva una respuesta. Si utiliza únicamente este método de transmisión, el nodo también debe configurarse para verificar periódicamente su conexión a través de la instrucción mesh.checkConnection (). La comunicación de nodo a nodo requiere que las consultas de dirección se envíen al nodo maestro, ya que los nodos individuales pueden cambiar de dirección en cualquier momento.

El motivo por el que nos hemos declinado por el uso de este dispositivo se centra en varias ventajas que a priori hemos visto que se adaptan bien a sistemas de sensores. Por una parte, el NRF24L01 es muy barato (hasta 10 veces) aunque por el contrario, el tamaño de mensaje es más pequeño, siendo de 32 Bytes, es decir, si la cabecera del protocolo es grande, el espacio para la trama de datos es pequeño [14]. El protocolo utiliza sólo 8 Bytes cuando se rellenan todas las condiciones, redundancia, ACKs, checksum y repeticiones.

El campo de dirección es muy similar al protocolo IPV4, pero en nuestro caso, usamos valores hexadecimales con un tamaño de 5 bytes. Otra peculiaridad es el uso de tuberías de comunicación, tiene como máximo 5 flujos por lo que cada direccionamiento en uso se asigna a una de ellas. El Pid (Packet ID) que se ha empaquetado se particiona en el tipo de mensaje (por ejemplo, la temperatura, la humedad,...), el valor y si es una difusión o una solicitud punto a punto al nodo final de solicitud o respuesta. La parte del origen y el destino del mensaje viene asignado por las direcciones y el uso de las tuberías de comunicaciones. El CRC se calcula sólo cuando se alcanza el nodo final. Si no es así, se activa el reintento. Las marcas de las repeticiones son más eficientes y más rápidas para reducir el tiempo de trabajo de la red.

V. RESULTADOS

Esta sección muestra los resultados de rendimiento, por un lado a nivel de datos erróneos en los envios de los sensores y por otro en cuanto a tráfico generado para 2 escenarios de transmisión diferentes.

Como se observa en la Fig. 6, es posible que por problemas del ambiente o defectos en componentes físicos del sensor, se puede dar el caso que los datos sean erronos y estén fuera de un umbral de trabajo. Por ello mostramos una tasa de error en un sensor usado, mostrando una tasa de error cada 60/70 minutos de capturas realizadas cada minuto (1,6%).

Fig. 2. Imagen de nodos del sistema

Fig. 3 Sensores del funcionamiento sistema.

Fig. 4. Sensores de estado del agua y cultivo

Fig. 5. Nodo central con actuadores, conectividad radio y 3G.

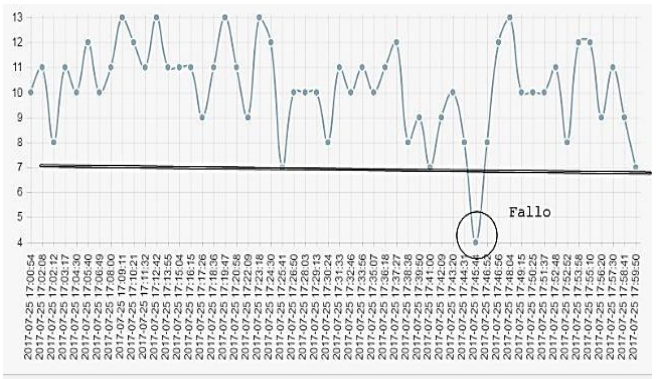


Fig. 6. Detección de datos erroneos, sensor flujo

A. Escenarios de pruebas de la red de comunicaciones

El sistema de radio soporta una función de confirmación automática (autoACK), activada de forma predeterminada, en la que la radio receptora cambia a modo de transmisión después de la recepción y reconoce la recepción de datos. Esto supone mantener un canal bidireccional.

En el primer test se crea un escenario sencillo con 3 nodos, A, B y C, con el siguiente intercambio de mensajes:

1. A envía a B
2. A envía a C
3. B responde A
4. A ya sabe que B tiene el mensaje correcto.
5. C responde A
6. A ya sabe que C tiene el mensaje correcto.

La Fig. 7 muestra el consumo en bytes de la red para el intercambio de mensajes explicado. Acorde con las recomendaciones de los fabricantes, si mantenemos la confirmación automática activa, estamos limitados a tener solamente 5 nodos en la red. Por lo que para este trabajo será necesario desactivar esta función.

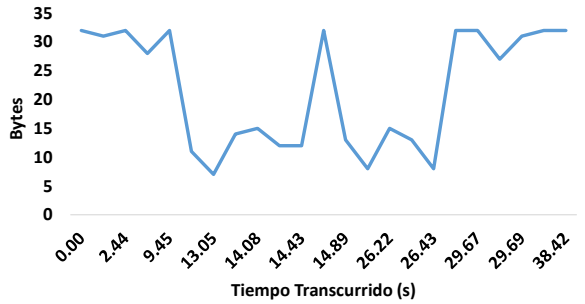


Fig. 7. Tráfico generado en la conexión punto a punto.

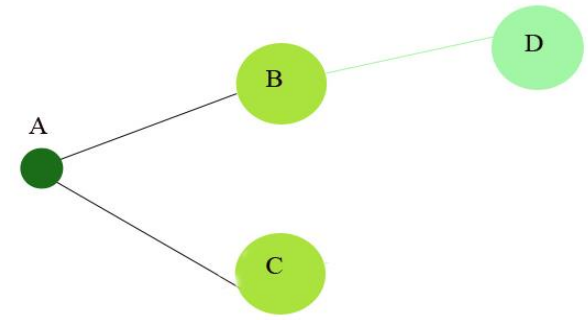


Fig. 9. Árbol de nodos en pruebas.

En la Fig. 9 se muestra el árbol de nodos de las pruebas con red mallada donde B depende de A y da comunicación a D. Como podemos observar en la Fig. 10, existen ocasiones en el que los paquetes no alcanzan su destino, de ahí que tengamos algún CRC error (longitud 10/12 bytes) y se reintente el envío.

El modo de difusión elegido puede tener problemas de colisión, aun con el algoritmo de desfase introducido. Esta tarea será un futuro análisis [17] a estudiar en siguientes trabajos.

La siguiente prueba de envío de datos se realiza con la función Auto-ACK desactivado. Se emplean en este caso, 4 nodos, A, B, C y D en forma de red mallada. El intercambio de mensajes que se realiza es el siguiente:

1. A envía
2. B, C, D escuchan
3. B tiene el mensaje con el ID correcto.

Como nos muestras la Fig. 8, se ha realizado un envío de mensajes sin auto respuesta ACK y se puede ver alguna pérdida de trasmisiones, posiblemente por colisión de los paquetes. A la hora de implementar una red mallada comprobaremos la recepción de mensajes con el ACK de respuesta del destinatario.

Por último, realizaremos una prueba con 4 nodos con Auto-ACK activo donde cada radio está vinculado a una dirección [15] (Bit Masks) y tubería que utiliza para comunicar con su padre o hijo. El nodo, una vez que escucha el paquete puede aceptarlo [16], encaminarlo a su padre o encaminarlo a su hijo: A flujo 1, B flujo 2, C flujo 3 y D flujo 4.

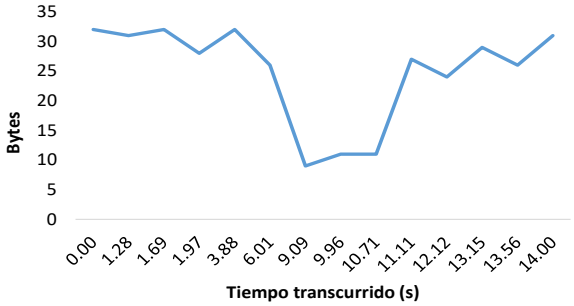


Fig. 8. Tráfico generado en las pruebas de red mallada.

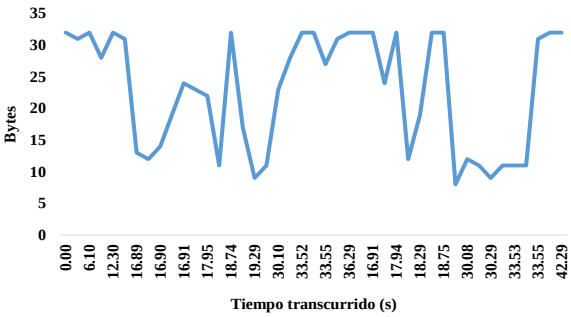


Fig. 10. Tráfico generado sin auto respuesta ACK.

B. Control de bajo consumo energético

Para maximizar el tiempo de funcionamiento de una batería cuando se emplea el módulo NRF24L01, debemos configurar una serie de mecanismos de ahorro de energía. Las líneas de código implementadas que nos permiten obtener este ahorro de energía se muestran en la Fig. 11. Para asegurar un ciclo de vida de batería más largo, se puede reducir la frecuencia del microcontrolador a 1MHz y bajar el voltaje de funcionamiento. Podríamos elegir trabajar a 1.8 V en vez de 3.3V.

```
radio.powerUp();  //turn the power on NRF24

// sending data

...

radio.powerDown();  //turn off the power on NRF24
```

Fig. 11. Encendido y apagado del módulo de radio.

```
s328o1.name=Sensor328p (int1MHz, 1.8V)

s328o1.upload.protocol=arduino
s328o1.upload.maximum_size=30720
s328o1.upload.speed=19200

s328o1.bootloader.low_fuses=0x62
s328o1.bootloader.high_fuses=0xda
s328o1.bootloader.extended_fuses=0x06
s328o1.bootloader.path=atmega

s328o1.bootloader.file=ATmegaBOOT_168_atmega328_pro_8MHz.hex

#s328o8.bootloader.file=ATmegaBOOT_168_atmega328.hex

s328o1.bootloader.unlock_bits=0x3F
s328o1.bootloader.lock_bits=0x0F

s328o1.build.mcu=atmega328p
s328o1.build.f_cpu=1000000L
s328o1.build.core=arduino
s328o1.build.variant=standard
```

Fig. 12. Código para reducción de ciclos de reloj.

```
//sleep mode for the microcontroller
void system_sleep() {
delay(2); // Wait for serial traffic
_SFR_BYTE(ADCSRA) &= ~_BV(ADEN); // Switch ADC off
set_sleep_mode(SLEEP_MODE_PWR_DOWN);
sleep_enable();
sleep_mode(); // System sleeps here
sleep_disable();
_SFR_BYTE(ADCSRA) |= _BV(ADEN); // Switch ADC on
}
void wdt_interrupt_mode() {
wdt_reset();
WDTCSR |= _BV(WDIE); // Restore WDT interrupt mode
```

Fig. 13. Código de control de estados del nodo e interrupciones

La Fig. 12 muestra el código perteneciente a la disminución de frecuencia de trabajo del chip, para descender el consumo energético. Para pasar el nodo a modo dormido será necesario implementar una función en la cual comparamos varias casuísticas que dependerán de los tipos de sensores conectados, estas pueden ser:

- Si el módulo es activado por una interrupción de una entrada (nivel de agua en un deposito)
- Si se despierta cada un tiempo definido de captura de datos (temp, humedad,..)
- Si se despierta por un tiempo para hacer comprobaciones de red (pulling) por si tiene alguna recepción pendiente.

En nuestro caso, se ha definido un tiempo de recepción cada dos segundos y de captura de datos cada 15 segundos. En el caso de pulling es necesario despertar el nodo y el radio cada dos segundos y mantenerlo despierto 500 ms ya que el trasmisor necesita dos segundos de preámbulo en cada paquete. La función implementada se muestra en la Fig. 13.

Existes diferentes enfoques en el modo de trabajo de este tipo de dispositivos y el modo de dormirlos y despertarlos [18], pero en definitiva, estos será elegidos en función de las condiciones de uso de los sensores. En cualquier caso, siempre nos permitirá alargar la vida útil de las baterías o capacitadores de almacenamiento.

Finalmente, la Fig. 14 muestra los valores de consumo de corriente de estos nodos, cuando están funcionando a 3.3 V. Cuanddo el nodo está transmitiendo datos los valores de corriente se sitúan en torno a los 30 mA. Cuando el nodo está transmitiendo datos, su consumo se incrementa hasta los 49 mA. Finalmente, en relación al consumo de energía cuando el nodo se encuentra en el modo sleep, el consumo de corriente se reduce radicalmente hasta los 90 µA.

En este tipo de sistemas, existe la posibilidad de usar placas solares para una alimentación extra, con el inconveniente de la movilidad y orientación forzada hacia la luz solar. En estos casos, se debe hacer un balance entre la energía consumida para el funcionamiento del nodo y el movimiento de la placa solar y la energía solar que es capaz de captar la placa usada.

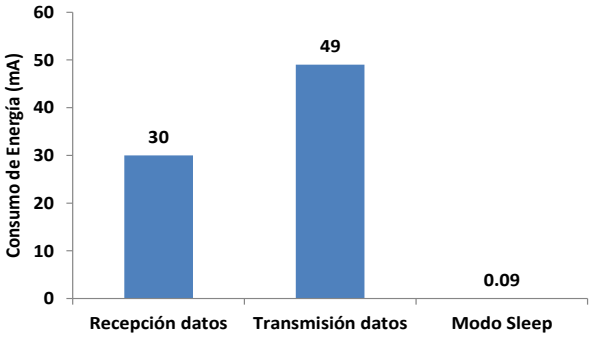


Fig. 14. Consumos de energía del nodo

VI. CONCLUSIONES

La evolución de la sociedad tiende a proponer nuevos retos alimentarios relacionados con la evolución de nuevos sistemas de cultivo, como es el caso de los hidropónicos que está fuertemente ligada a las redes de sensores [19]. Las WSN son sin duda la mejor solución para el seguimiento cultivo y su correcta producción.

Se está tendiendo al uso de tecnologías ecológicas energéticamente eficientes a través de los algoritmos y patrones de uso dedicados dentro de la Plataforma multimedia específicas. Este es el caso de PLATEM PA.

En este trabajo se ha presentado una WSN de bajo consumo que implementa un sistema autónomo construido con el objetivo de automatizar el proceso

productivo en el cultivo intensivo [20]. El diseño de la red de sensores inalámbricos se centra en el uso de nuevos diseños de redes malladas caracterizadas por alcanzar un consumo realmente bajo y la independencia de los nodos con el mínimo soporte técnico posible.

Dentro de este sistema la captación de parámetros ambientales, datos del cultivo y calidad de agua y nutrientes son cuantificados en pocos segundos gracias a la red de sensores desplegada en el invernadero, la corrección de parámetros puede ser llevada a cabo en un corto espacio de tiempo y reaccionar con antelación a posibles problemas o casuísticas distintas. Hemos observado que los nodos de sensores en algún momento emiten datos erróneos asíncronamente Aprovechando el gran número de nodos de sensores desplegados, es necesario tener una regla de control bajo un umbral de variables asignadas, para filtrar datos totalmente alejados de parámetros reales del sistema. Si un nodo es problemático continuamente, se detecta y se desecha esa información.

La sociedad demanda cada vez más cultivos más saludables y que puedan ser transformados con el menor impacto ambiental posible [21], de ahí que en pocos años veamos un despliegue exponencial de huertos urbanos altamente tecnológicos. Como futuros desarrollos y avances en este tipo de redes de sensores de bajo coste, deseamos continuar con el análisis y depuración del protocolo de comunicaciones en redes malladas con esta tecnología, para solventar el problema leve de pérdida de paquetes, que suponen hacer mayores reintentos de envío del mensaje y conlleva a un mayor consumo de energía. Una vez estabilizado ese problema, las líneas de trabajo se orientarán a crear mayor inteligencia a la plataforma tecnológica, tanto para la difusión de mensajes entre los responsables de los sistemas, como en actuaciones automáticas o acciones para recortar los periodos de anomalías en los sistemas de producción. Compartir los datos de la plataforma PLATEM PA con robots encargados de tareas de supervisión y de procesos mecanizados será clave en la integración de sistemas 100% autónomos sin tener el factor humano físicamente en el lugar de operaciones y pueda hacerlo en remoto.

REFERENCIAS

[1] D.D. chaudhary, S.P Nayse, L.M. Waghmare, “Application of Wireless Sensor Network for Greenhouses Parameter Control in Presicion Agriculture”, International Journal of Wireless & Mobile Networks, 2011, vol 3.No 1, Pp. 140-149.

[2] S.R. Boselin, M. Pradeep, E. Gajendran, “Monitonig Climatic Conditions Using Wireless Sensor Networks”, A multidisciplinary Journalof Scientific Research & Education, 2017, Vol. 3, No. 1, Pp.179-184.

[3] S Sendra, J Lloret, C Turro, JM Aguiar, IEEE 802.11 a/b/g/n short–scale indoor wireless sensor placement, International Journal of Ad Hoc and Ubiquitous Computing, 2014, Vol.15, No. 1-3, pp. 68-82

[4] S. Janos, G, Martinovic, I. Matijevics, “WSN Implementation in the Greenhouses Environment Using Mobile Measuring Station”, International Journal of Electrical and Computer Engeenering Systems. 2010, Vol 1, No. 1, Pp. 37-44.

[5] C. Cambra, S. Sendra, J. Lloret, L. Garcia, “An IoT Service-Oriented System for Agriculture Monitoring, In proceedings of the Intenational Conference on Communications 2017, May 21-25, Paris (Francia).

[6] A. Tozounis, N. Katsoulas, K.P Ferentinos T. Bartzanas, C. Kattas, “Development of a WSN for Greenhouse Microclimate Distribution Monitoring”, University of Targoviste-Agriculture, 2016, Vol.10, No. 1, Pp. 7-13.

[7] J. Lloret , M. Garcia, D. Bri, S. Sendra, “A Wireless Sensor Network Deployment for Rural and Forest Fire Detection and Verification”, Sensors, 2009, Vol. 9, No. 11, Pp. 8722-8747.

[8] C Cambra, S Sendra, J Lloret, L Parra. “Ad hoc network for emergency rescue system based on unmanned aerial vehicles”, Network Protocols and Algorithms, 2016, Vol 7, No 4, Pp. 72-89.

[9] H. Ibayashi, Y. Kaneda, J. Imahara, N. Oishi, M. Kuroda and H. Mineno, A Reliable Wireless Control System for Tomato Hydroponics” Sensors 2016, 16, 644 .

[10] G. Barrenetxea, F. Ingelrest, , G. Schaefer, M. Vetterli, “Wireless sensor networks for environmental monitoring: The sensor scope experience”, In proceedings of the 2008 IEEE International Zurich Seminar on Communications, March 12-14, 2008, Zurich (Switzerland). (pp. 98-101).

[11] C. Yeng, S. Yuling, W. Zhongyi, “Connectivity of Wireless Sensor Networks for Plant Growth in Greenhouse”, International Journal of Agricultural and Biological Engineering of Beijing, 2016, Vol. 9, No 1, Pp. 89-98.

[12] Nordic Semiconductor, RF specialist in ultra low power wireless communications. Disponible en: http://www.nordicsemi.com/eng/Products/2.4GHzRF/nRF24L01 , [Último Acceso, 20 de Abril de 2017]

[13] Caracteristicas de Arduino Nrf24L01. Disponible en: http://playground.arduino.cc/InterfacingWithHardware/ Nrf24L01, [Último Acceso, 20 de Abril de 2017]

[14] C. Zhurong, H. Chao, L. Jingsheng, and L. Shoubin, “Protocol architecture for wireless body area network based on nrf24l01”, In proceedings of the 2008 IEEE International Conference on Automation and Logistics (ICAL 2008), Sept 1-3, 2008,Chindao, (China). (pp. 3050–3054).

[15] V. Gupta, P. Raspaile, “Low Cost Standard Internet of Things”, International Journal of Engineering Science & Advanced Technology, 2015.Vol.5, No.2,pp. 78-80.

[16] Arduino Bit Masks definitions. Disponible en: https://www.arduino.cc/en/Tutorial/BitMasks, Mayo 2017 [Último Acceso, 20 de julio de 2017]

[17] P. Mutukhumaran, R. Paz, R. Spinar, “ MeshMAC: Enabling Mesh Networking over IEEE 802.15.4 through Distributed Beacon Scheduling”, In proceedings of the International Conference on Ad hoc Netwroks, ADHOCNETS 2009, September 22-25, 2009. Niagara Falls, Ontario,(Canada). (pp. 561-575)

[18] T. Camilo, C. Carretero, J. da Silva, “An Energy-Efficient Ant-Based Routing Algorithm for Wireless Sensor Networks”, In proceedings of the 5th International Workshop on Swarm Intelligence (ANTS 2006), September 4-7, 2006, Brussels, (Belgium). (Pp.49-59).

[19] N. Pantacis, D. Vergados. “Energy efficiency in wireless sensor networks using sleep mode TDMA scheduling”, Adhoc Networks, 2009, Vol. 7, No. 2, Pp. 322-343.

[20] L. Ruiz-Garcia, P. Barreir, J.I. Robla, “Performance of ZigBee-Based wireless sensor nodes for real-time monitoring of fruit logistics”. Journal of Food Engineering. 2008, Vol. 87, No. 3, Pp. 405–415.

[21] A. Pawloswki, J. Guzman, F. Rodriguez, “Simulation of Greenhouse Climate Monitoring and Control with Wireless Sensor Network and Event-Based Control”, Sensors, 2009, Vol. 9, No. 1, Pp. 232-252.

[22] FAO, Organización de las Naciones Unidas. Disponible en: http://www.fao.org/climate-smart-agriculture/es/, [Último Acceso, 20 de julio de 2017]

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6491

Feasibility assessment of a fine-grained access control model on resource constrained sensors

Mikel Uriarte, Jasone Astorga, Eduardo Jacob, Maider Huarte
Departament of Communications Engineering,
University of the Basque Country UPV/EHU
Bilboko Ingenieritza Eskola, Urkixo Zumarkalea S/N, 48013 Bilbo, Bizkaia
muriarte@nextel.es, jasone.astorga@ehu.eus, eduardo.jacob@ehu.es, maider.huarte@ehu.es

***Resumen*—Upcoming smart scenarios enabled by the Internet of Things (IoT) envision smart objects that expose services that can adapt to user behaviour or be managed for higher productivity. In such environments, smart things are cheap and, therefore, constrained devices. However, they are also critical components because of the importance of the information they provide. Therefore, strong security is a must, but not all access control models are feasible. In this paper, we propose the feasibility assessment of an access control model that deals with a hybrid architecture and a policy language that provides dynamic fine-grained policy enforcement in the sensors, which requires an efficient message exchange protocol called Hidra. This experimental performance assessment conveys a prototype implementation, a performance evaluation model, the measurements and the related discussions, which demonstrate the feasibility and adequacy of the analysed access control model.**

***Palabras Clave*—access control, authorization policy language, constrained device, Internet of Things, security**

I. INTRODUCTION

The Internet of Things (IoT) concept conceives an interconnected network of things, the smarter the better, contributing to a higher awareness, enhanced decision making, and more adaptive behaviour of systems supporting any business process integrating pervasive and ubiquitous ICT technologies. IoT also implies a massive deployment of sensors and actuators, which, aiming at being cheap, are implemented in a range of constrained devices, constrained device sensors (CDSs) from now on, classified according to IETF [1], from severely constrained C0 to less constrained C2. Moreover, depending on the use case and location, they may require power autonomy, and therefore, require low power consumption mechanisms.

In such IoT applications, security (more specifically, access control) remains an insufficiently solved problem since existing approaches are challenged by divergent properties as tightness and feasibility. Consequently, in this paper, we propose the feasibility assessment of an access control model based on an expressive policy language enabling tight enforcement in CDSs. Such access control model includes a protocol that enables secure provisioning and enforcement of dynamic security policies as well as an audit trail, and this protocol is the subject of the performance evaluation driven through a prototype implementation.

Beyond the traditional producer behaviour of CDSs, which publish measurements and events to message brokers, in more advanced IoT scenarios, CDSs behave as tiny information servers. Specifically, requesting clients directly query the tiny CDS servers, establishing a secure end-to-end (E2E) communication. These exposed services enable usage, operation, maintenance and manageability of CDSs over their entire life-cycle and protect the value stream of the connected objects. For example, an end-user can access directly to tune personal parameters such as gender, age, weight, etc. in a health constant monitoring sensor. Moreover, the use of intermediary proxies is avoided because on one hand, they are specific for each protocol or application and are not flexible enough, and on the other hand, breaking the security association into two or more sub-transmissions might not be considered acceptable from a security point of view.

Fig. 1 shows an IoT schema that conveys different roles in various domains, operating, monitoring and controlling related business process through applications and fully aligned with the functional decomposition view of the IoT architecture reference model (ARM) [2].

In this context, the accuracy and correctness of the information exchanged with CDSs is crucial. Protecting this information requires the implementation of appropriate security mechanisms that include fine-grained access control mechanisms based on expressive policies and that can guarantee essential security properties such as confidentiality, integrity, availability, authenticity and non-repudiation [3], [4], [5]. However, implementing these appropriate security mechanisms in resource-constrained CDSs is not straightforward. Currently, one of the key challenges for enabling broader adoption of smart things is

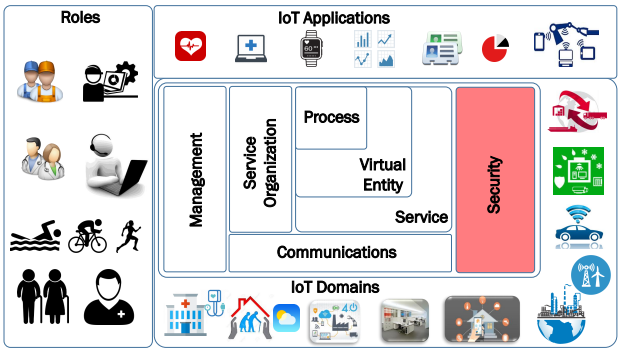


Fig. 1: Scenario schema where several stakeholders playing different roles access IoT applications on different IoT domains. The core shows the functional decomposition view of the IoT reference architecture, where security functional group is highlighted.

the availability of feasible access control solutions. Moreover, due to the extremely dynamic nature and purpose of applications based on services in sensors, policy-based security must be enforced locally in the CDSs, where resources are scarce.

The main contribution of this paper is a feasibility assessment of such a highly expressive E2E access control model in severely constrained devices (C0 and C1 CDSs), based on an experimental performance evaluation.

The rest of the paper is organized as follows. Related works are presented in Section II as the state of the art. The proposed access control model is specified in Section III. The performance evaluation conveying an experimental prototype is discussed in Section IV. Finally, the main conclusions of the paper are presented in Section V.

II. STATE OF THE ART

In the last years, the research area related to security in IoT has received a significant attention, dealing with the design of different architectures, security protocols and policy models. But security still remains as the main obstacle in the development of innovative and valuable services [3]. In fact, traditional security countermeasures cannot be applied directly to CDSs in IoT scenarios, because they are too resource consuming and not optimized for resource deprived devices. Additionally, existing feasible E2E access control approaches do not implement an expressive and therefore fine-grained and tight security policy enforcement [6].

For feasibility reasons, a centralized architecture based on traditional standards and protocols, where a central access control server (ACS) with no resource constraints makes authorization decisions for each access request, could be initially a possible option. But this approach does not consider local context conditions in CDSs, and it implies high energy consumption as well as network overhead due to continuous communications between the CDSs and the ACS.

A recent alternative approach is the distributed capability based access control (DcapBAC) [7], where an unforgeable token exchangeable as a capability, grants access to its holder in a more agile way. However, the token is designed in a XML schema and it has not been validated in constrained devices.

In any case, this approach has been adopted by some other designs involving technologies specifically defined for IoT, which enable CDSs to make local authorization decisions based also on local conditions [8], since the capabilities might include conditions represented as tuples (type, name, value). Per contra, this approach is based on public key cryptography (PKC) which is heavier than symmetric key cryptography (SKC) by means of resource consumption. Additionally, the conditions are limited to matching because the approach does not support expressions. Moreover, its syntax is not optimized by means of codification since it uses JSON, it does not support the enforcement of additional obligations and it has been validated in not so constrained C2 devices.

In this line, the delegated CoAP [9] authentication and authorization framework (DCAF) [10] defines a token to distribute pre-shared keys, and if authorized, a handshake is done to establish a DTLS channel. Local authorization policies are specified as conditions serialized in a concise binary object representation (CBOR), instead of JSON, aiming at compacted payloads in CoAP protocol. But CBOR is a general purpose serialization solution [11] and the resulting compression is not sufficiently optimized for security policies in very constrained C0 and C1 devices, where fine-grained access control is aimed through a higher but feasible policy language expressiveness, beyond the conditions consisting of simple constant matching of existing local attributes.

In other line, the usage control model and the attribute based policy schema [12] extend traditional access control systems to a continuous protection of the resource during access by the definition of obligations to enforce usage control. However, there is not an approach addressing the feasibility in CDSs.

Finally, attending to the protocols for the instant provisioning of the policy during the E2E security association in a secure session, Ladon [13], which is inspired in Kerberos, is susceptible to be evolved for that purpose. In fact, Ladon is specifically designed for very constrained devices, but it does not directly support the provisioning of an expressive policy.

Consequently, currently no suitable solution exists to provide authentication and fine-grained authorization processes in the envisioned scenarios of constrained but manageable sensor networks, and additionally, neither of the above considered approaches implements any accounting feature.

III. ACCESS CONTROL MODEL

The E2E access control model subject of feasibility assessment is based on an efficient policy language and codification, which are specifically defined to gain expressiveness in the authorization policies and to keep the viability in very constrained devices. Besides the

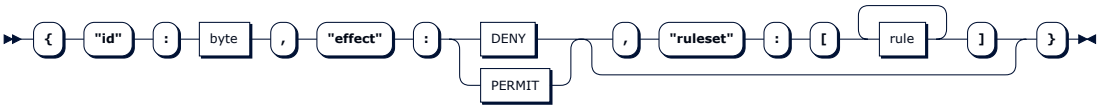


Fig. 2: Policy construct definition.

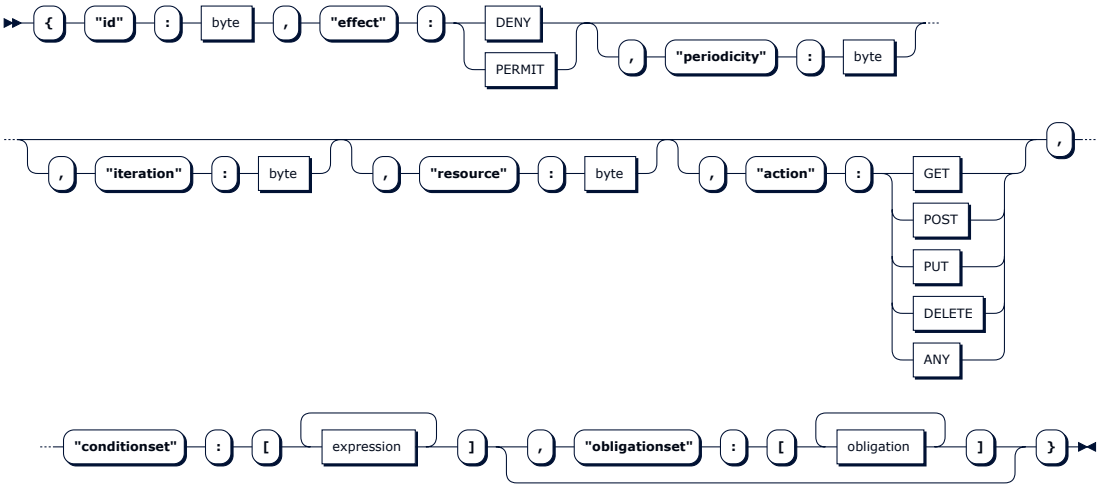


Fig. 3: Rule construct definition.

policy language, the access control model conveys the E2E feasible security association between two mutually authenticated peers, and consists of an architecture to enable multi-step authorization as well as a protocol for the provisioning and enforcement of a dynamic security policy in the CDSs.

A. Authorization policy language and codification

In this section, an expressive policy language is briefly presented. The goal of this policy language is to enable the enforcement of tighter access control policies in CDSs, overcoming the resource constraints. In fact, this policy language definition enables both to make granting decisions based on local context conditions, and to react accordingly to the requests by the execution of additional tasks defined as obligations.

A resulting policy instance is defined, like in the general event-condition-action approaches, as an optional set of rules, which specifies both the conditions to be checked and the related reactions, in enforcement time. Specifically, this policy language stands for a sequence of constructs with particular meaning in the decision making and enforcement time.

Some of the constructs are defined as mandatory, and some others as optional, enabling to shorten the length of the policy when a simple policy is enough. Additionally, some constructs are extended through other nested constructs, and some of them can be instantiated many times within a container construct. Related to this elasticity feature, the more constructs, the higher the expressiveness of the policy, so the more granular the policy is, and consequently, the tighter the enforcement is. Consequently, the challenge to overcome is to be feasible even in the most expressive use-case.

The policy language enables a policy instantiation through the *policy* construct, with three nested constructs as depicted in Fig. 2. First of all, a policy instance identification, *id*, is specified for logging, tracking and auditing purposes. Then, a default policy granting *effect* is specified. This effect will prevail in the case of absence of rules, or any rule evaluation conflict. Lastly, optionally, an array of rules may be instantiated as a *ruleset* to specify the conditions and related reactions. Each *rule* in the array is an extensible construct.

The *rule* construct depicted in Fig. 3 is defined as a sequence of eight nested constructs, where the order is crucial. Some of them, such as *id*, *effect*, and *conditionset* are mandatory, and the rest named *periodicity*, *iteration*, *resource*, *action* and *obligationset* are optional. The *conditionset* and the *obligationset* are arrays of expressions and obligations respectively. These repeatable and extensible *expression* and *obligation* constructs are defined in a similar way enabling the instantiation of rich expressions on attributes declared as inputs as well as reactive tasks declared as obligations.

The length of any policy instance, in a human readable format, grows proportionally with the aimed tightness, and it would impact negatively in the performance. So a specific policy instance codification is considered, distinguishing from existing ones that serialize policy instances through standardized generalist solutions such as CBOR.

The considered policy codification serializes each construct and concatenates them in a bit stream. In fact, it takes profit of beforehand knowledge of the defined sequence of the constructs, and their format. An additional crucial factor is the injection of some agreed bit masks,

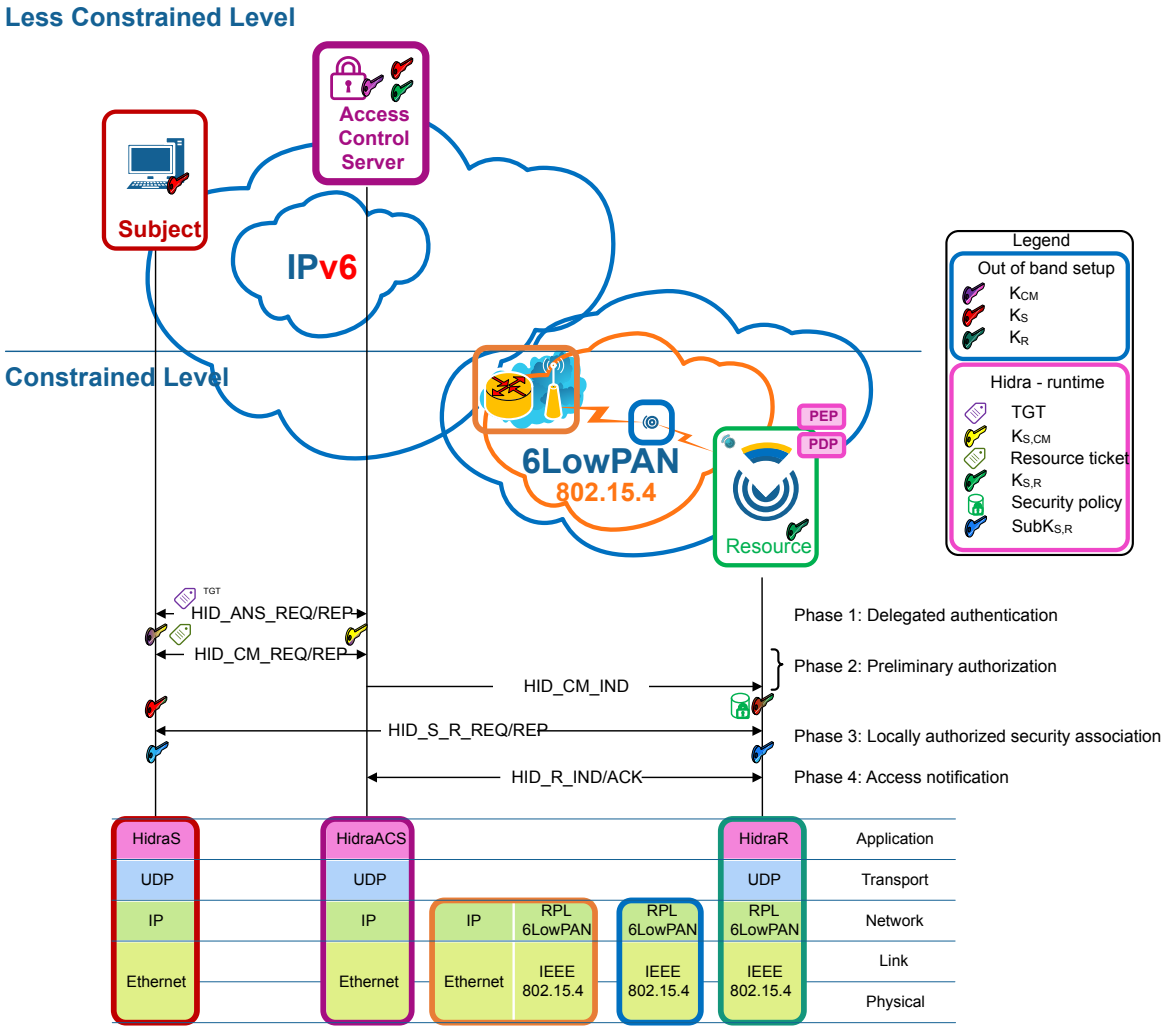


Fig. 4: Performance evaluation scenario. Hidra protocol messages and security association establishment related authentication, authorization, key exchange and notifications.

to specify the existence or not of optional constructs. It enables to deal optimally with the elasticity defined in the policy language, avoiding unused but expected fields of expressive policies, greatly reducing the length.

With respect to covered policy formats, this policy codification can easily be applied to any original policy instance format (XACML, JSON, *etc.*), from textual files to structured policy instance representations. For example, four different instances of sample policies in the testlab, represented in JSON with lengths of 23, 118, 174 and 554 bytes, which are codified in CBOR with lengths of 14, 81, 123 and 391 bytes respectively, are notably reduced to 2, 7, 9, and 32 bytes using the considered codification.

B. Hidra messaging protocol

In order to efficiently convey the presented access control policies to the CDSs, the Hidra protocol is considered. Hidra, depicted in Fig. 4, is based on a three parties architecture, and provides authentication, authorization in two steps, dynamic policy provisioning and accounting.

Hidra is based on Ladon [13], which is a validated solution for the establishment of E2E security associations, through pair-wise keys, guaranteeing mutual authentication and authorization in very CDSs.

Both Hidra and Ladon are based on symmetric key cryptography and they assume that each endpoint owns a secret key shared with the ACS. The operation is based on the use of tickets, a capability distributed by the ACS that contains a proof of the identity of the subject that requests it. Tickets are encrypted so that only the entities which they are intended for, are able to decrypt them.

After a successful authentication in the ACS (Phase 1) the subject that wants to access a service in the CDS obtains a ticket granting ticket (TGT). This TGT is used by the subject to obtain resource tickets (Phase 2) required to access any resource on the CDSs.

This approach enables the attribute based access control (ABAC) authorization enforcement in two steps. On the first one, as condition to release any resource ticket, fine-grained preliminary access control is performed in the ACS (Phase 2), focusing on the attributes of the subject, resource and expected actions. If this first authorization step is successful, the ACS sends a message to the subject including a resource ticket, and also sends a message

to the CDS conveying an expressive authorization policy instance. This instantaneous custom policy provisioning avoids permanent policies' storage in the CDS and reduces network overhead comparing with approaches enclosing the policy in the resource ticket.

On the second authoritative step, once the subject has obtained a resource ticket, the local context based access control is performed in the CDS (Phase 3). First, the proper rule is evaluated to make the granting decision, and then the corresponding reactive actions are enforced. In a positive authorization case, the result is a shared session key to be used on further E2E resource access exchanges.

Another novelty of Hidra with respect to Ladon is the addition of a pair of messages to enable precise accounting (Phase 4). By means of these messages, the CDS will notify details like who performed what, where and when in each and every access request received from the subject. These notifications are gathered, normalized, and treated properly by the ACS. Additionally, the ACS can react and send a related policy message, enabling the dynamic delegation, request, cancellation and revocation of permissions.

Then, while the security association is not finalized, the access control is enforced in the CDS autonomously in each and every further request attempt, since the received expressive policy (Phase 2) includes related rules.

Consequently, unified, coherent and adaptive management of the policies by the ACS is achieved. Additionally, the proposed Hidra protocol and the adopted architecture enable to rely the most expensive features on the ACS, which entails the usage of standard security and access control technologies in the non constrained interactions. It also achieves that most unauthorized access attempts are refused before reaching the CDS, avoiding unsuccessful message exchanges and thus, saving energy in the CDS, which is a crucial aspect.

IV. PERFORMANCE EVALUATION

In this section, the experimental performance analysis of an access control model for E2E security in CDSs is carried out conveying the establishment of a security association between a requesting subject and a CDS. For performance analysis there are three main options: analytical evaluation, specific network simulation and prototype implementation. Even though this third one is more complex and expensive, this proposal conveys an Hidra protocol implementation prototype in order to present more accurate and realistic performance results. Therefore, the performance analysis covers the measuring and evaluation of the crucial performance parameters of such resource constrained sensing environments.

First, the reference scenario, some assumptions and its implementation through software codification and configurations are discussed. Then, the crucial performance parameters are identified, and their measurement methods and computation are described. Finally, an analysis of the measurements is presented, describing and discussing the evaluation results.

The overall goal is to demonstrate the suitability of the designed access control model for CDSs in the envisioned scenarios.

A. Test-bed implementation

The testing scenario for the performance evaluation is graphically depicted in Fig. 4. In this scenario, a subject is connected to the Internet and establishes an E2E connection with a resource running on a CDS in an IEEE 802.15.4 network. A 6LoWPAN router (in orange) acts as the LowPAN coordinator and connects a beacon-enabled lineal structure to the Internet. The IEEE 802.15.4 network is 2-hops deep, which is considered significantly large for validation purposes. The PAN router coordinator has a child coordinator implemented in a TelosB [14] sensor, which controls one leaf node. In this node the CDS exposes resources as management services, and it is implemented in an C0 hardware platform IRIS [15] sensor with Contiki OS [16].

In the implementation two aspects are distinguished: on one hand, network protocols and connectivity, and on the other, Hidra messaging protocol as an application.

Fig. 4 shows implemented protocol stacks in all entities involved in the performance evaluation scenario. The implemented network enables E2E IPv6 connectivity through a multi-hop IEEE 802.15.4 sensor network, in 2.4 GHz band. Note that IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) [17] is configured to set a multi-hop hierarchical node network with neighbour discovery functionalities within the IEEE 802.15.4 network. The network configuration is done through sensor *ID* setting and registration as a first step, and then UDP messaging module is installed.

Then, Hidra messaging protocol application is made by different software modules covering different functions in each of the three parties involved in the protocol: subject, ACS, and sensor.

Hidra in the sensor side, which means the biggest impact, is codified and installed as a server application on top of Contiki OS. In fact, it receives an indication and a policy from the ACS, and a request from the subject. So it checks the validity of the messages, the identities and the service ticket, prior to evaluating the policy locally and making a granting decision related to the security association establishment requested by the subject. Finally, it notifies any activity to the ACS through log messages for further tracking and auditing.

This sensor side Hidra module implements also an unique UDP socket to wait for any of the possible messages in order to be more efficient in memory footprint. Once a received message is identified, it is parsed and proper protocol related checks, validations and reactions are performed following an specific flow.

During development some design decisions were made. Specifically, where message authentication codes (MACs) are 16 bytes long; message cyphering is done with AES-128 and it is combined with Ciphertext Stealing algorithm to avoid size increments with respect to cleartext

messages. Cryptographic libraries are TomCrypt and tiny-AES-128 [18].

B. Performance modelling

To conduct a performance evaluation of the proposed access control model for E2E security in CDSs, the experimental performance model focuses on three critical parameters: (1) the response time of the access control model to establish an authorized E2E secure session, (2) the energy cost of this model for the protected CDS running on finite batteries, (3) the model's impact on the local storage on the CDS and memory footprint.

The response time needs to be below an accepted value if the proposal is to be useful, and the energy consumption, local storage, and memory footprint, due to the nature of the CDSs and their constraints on resources, cannot exceed rational and proportional limits.

1) Response time: During the establishment of a security association, five messages are exchanged, as detailed in Section III.B. This response time includes the steps where the subject requests and obtains the service ticket, the notification that the ticket is granted, policy provisioning in the CDS by the ACS, and the security association request and response between the subject and the CDS.

In order to measure this time, some few code lines are inserted in the subject's side software code, setting two timestamps: one at the beginning of the security association establishment and the other at the end of this establishment.

2) Energy consumption: Regarding the energy cost measurement, the energy consumed by the transmission and reception of bits over the air and the message processing are considered.

For the measurement of the processing energy consumption, two timestamps are inserted in the sensor's side, at the beginning and the end of the message processing software code. Once measured the time to process each message, a constant instantaneous power consumption (P_C) provided by the manufacturer in the datasheet is considered in order to compute the energy consumption of each message.

For the computation of the power consumption due to the transmission and reception of each message, involved message lengths in bytes and packet fragmentation are computed (considering 50 bytes of longest IEEE 802.15.4 plus UDP/6LowPAN headers). According to Ladon protocol message lengths [13], in which Hidra protocol is based, the lengths of the messages exchanged during the authentication and authorization protocol range from 15 to 63 bytes. Enclosing the policy in the HID_CM_IND message (33 bytes), which is one of the smallest, implies the minimum fragmentation of 6LowPAN IPv6 packets over IEEE 802.15.4 links. This design decision makes a difference with respect to the approaches for enclosing the policy in the ticket, which is included with larger request messages. Therefore, it can be anticipated a proportional and optimized impact in the length of a message from injecting the compressed policy into the shortest one. In particular, the length of the instantiated policy in this proposal is 2 bytes and consequently, the total length of the HID_CM_IND message is 35 bytes.

Tabla I: Parameters used to characterise the energy consumption of sensor nodes

Name	Description	Value
B_N	Effective network wireless link data bit rate	70Kbps
P_{RX}	Power consumption in reception mode	48 mW
P_{TX}	Power consumption in transmission mode (3dBm)	51 mW
P_C	Power consumption in message processing mode	8 mW

Additionally, constant reception and transmission power consumption rates provided by the manufacturer in the data-sheet and a constant propagation bit rate are also considered. Table I shows testlab real-conditions (non-optimal) network data bit rate and the different instantaneous power consumption values used for the analysis. Note that these power consumption values correspond to a MEMSIC IRIS mote (XM2110CA) powered with a $3V$ power supply [15].

Finally, the power consumption is calculated as the sum of the individual power consumptions of each of the involved messages in the sensor.

3) Storage and memory footprint: In this subsection, the increase of permanent storage and the memory footprint generated by the proposed access control model are considered. Specifically, the storage and memory footprint for the instant provisioning of a received access control policy and the code needed to parse it, as well as the data blocks related to the messages of the Hidra protocol, are considered together.

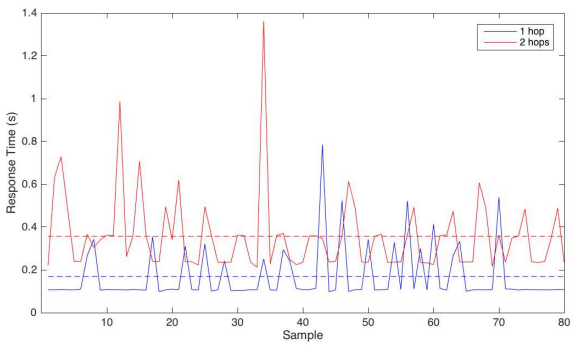
On one hand, regarding the permanent storage for the mentioned entities, the symmetric key shared with the ACS, the access control policy, the code to parse the received policy and the code to run the Hidra protocol are considered the additional minimum entities that should be permanently stored in a CDS along with some original resources and sensing applications.

On the other hand, regarding the memory footprint, the data required for the Hidra protocol message exchange are considered: namely, the session keys shared with the subject, the subject identity, the different data needed to identify the messages and to guarantee their freshness, the lifetime of the security association, and the log of each access provisionally wrapped until the reception of the acknowledgement from the ACS. However, some of these values are loaded and erased during the reception, processing and transmission of relatively consecutive messages.

Measurement of code plus data storage and memory footprint is done through a particular command (*size*), which provides with both static and dynamic occupied memory amounts.

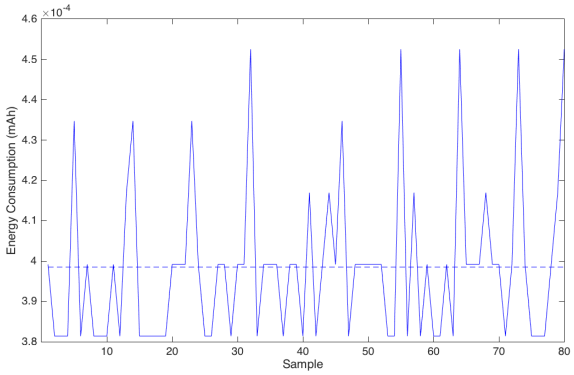
C. Performance analysis

In this section, the measurement results obtained on the experimental prototype described in the previous section are used to analyse the mean response time and energy consumption to establish a secure session, as well as the impact on the local storage and memory footprint. Response time and energy consumption have been measured



	Average (s)	Median (s)	St. Dev. (s)	95 % Conf. Int. (s)
1 hop	0.1689	0.1080	0.1293	(0.1405,0.1972)
2 hops	0.3576	0.3440	0.1834	(0.3174,0.3978)

Fig. 5: Response time of security association establishment considering 80 requests in two network configurations: one and two hops.



	Average (µAh)	Median (µAh)	St. Dev. (µAh)	95 % Conf. Int. (µAh)
1 hop	0.3985	0.3992	0.0199	(0.3941,0.4029)

Fig. 6: Energy consumption in security association establishment considering 80 requests in an one-hop network configuration.

by 80 samples, so potential measuring error is reduced. Additionally, response time has been measured with two different configurations: one hop and two hops respectively, so the impact of the intermediate node insertion can be measured.

1) Response time: First, impact of Hidra on performance is conveyed measuring response time of 80 subject's security association establishment requests, considering both network configurations: with one and two hops.

Fig. 5 shows that the maximum response time, even with a non-optimal network bit rate, is below $420ms$ and $637ms$ in both network configurations. This value is very good, considering that the maximum acceptable delay in interactive E2E data transactions specified by Stallings [19] is $1000ms$.

Fig. 5 showing a composition of the measurements with two configurations, points out that a second hop increases proportionally $200ms$ on average. This value could also be considered as referential increment per hop for further estimations aiming at large scale deployments.

Additionally, one related comparable response time value has been found in the literature, although there is no mention of additional performance indicators such as energy consumption. At the C2 level, [20] reveals that the comparable measured response time for the authorization response starting when the subject sends the request is $480.96ms$ (with one hop). The response time of Hidra is lower even with a worse network bit rate and, therefore, better.

2) Energy consumption: Attending to the impact of Hidra protocol on the energy consumption Fig. 6 shows measured values of power consumption related to each of 80 requests as well as the average.

This figure shows that the measured average value of power consumption is $0,3985\mu Ah$, which is a very low value. This very good result means that impact of Hidra in energy consumption is very low. In the case of two AAA batteries of $900mAh$, suming up to $1800mAh$, considering a 95% of battery performance, makes a real capacity of $1710mAh$. Therefore, more than 4 millions (4,291,006) of requests could be handled during the battery life.

In the envisioned scenario, the CDS is accessed by the subject to perform tasks such as personalization, parametrization, updating, upgrading, maintenance, and so on. These types of interactions do not occur often. As the most exigent scenario, we can consider one that requests access every hour to tune the user experience in an application where the users change hourly.

In this most exigent scenario, a subject making one request per hour, 24 per day, 8760 per year, could get response for approximately 490 years (489.840) of the battery life. Therefore, Hidra's energy consumption could be deprecated, and the battery life would depend basically on the main purpose application of the CDS.

3) Storage and memory footprint: Finally, from the storage point of view, at the CDS, the amount of RAM memory is the most limiting aspect, compared with permanent storage, which is usually an order of magnitude larger.

Assuming that measures are dependent on programming style, measurements of storage occupancy are 20836 bytes, and memory footprint is 1440 bytes. Therefore, the impact is considered acceptable considering available 128KB and 8KB of flash and RAM memory in so constrained devices such as the ones used in the implementation [15].

V. CONCLUSION

Incoming smart scenarios enabled by IoT envision smart objects exposing services to be adapted to user experience or to be managed aiming at higher productivity, often in multi-stakeholder applications. In such environments, smart things are cheap, therefore constrained devices, but they are also critical components, so security is a must. Existent approaches coping with the principle of least privilege, based on the expressiveness and updating of the policy to be enforced in the sensors, are challenged by feasibility constraints.

The proposed performance evaluation is focused on an innovative access control model dealing with a hybrid architecture and a policy language for dynamic fine-grained policy enforcement in the sensor. Such policy enforcement is based on local context conditions and correspondent obligations, not only during secure session establishment but also afterwards while the security association is in use, in order to control the behaviour of the access. Such a dynamic policy cycle avoiding local storage, is enabled by an efficient message exchange protocol, named Hidra. Actually the Hidra protocol assures the mutual authentication, the expressive policy injection, the tight policy enforcement in the secure association establishment and the derived resource access, as well as the accounting for further tracking and auditing purposes.

The proposed feasibility assessment is based on a prototype implementation of such an innovative access control model in very constrained devices. The experimental performance analysis focusing on three key performance indicators such as the response time, the power consumption and the memory footprint outcomes remarkable results. Based on these measurements, the performance evaluation of this proposal concludes the feasibility of this analysed access control model on resource constrained sensors.

REFERENCIAS

[1] C. Bormann, M. Ersue, and A. Keranen, "Terminology for constrained-node networks," Internet Requests for Comments, RFC Editor, RFC 7228, May 2014, http://www.rfc-editor.org/rfc/rfc7228.txt.

[2] F. Carrez, M. Bauer, M. Boussard, and N. Bui, "Final architectural reference model for the iot v3.0," http://www.iot-a.eu/public/public-documents/d1.5/at_download/file, July 2013.

[3] S. Sicari, A. Rizzardi, L. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," *Computer Networks*, vol. 76, pp. 146 – 164, 2015.

[4] K. T. Nguyen, M. Laurent, and N. Oualha, "Survey on secure communication protocols for the internet of things," *Ad Hoc Networks*, vol. 32, pp. 17 – 31, 2015, internet of Things security and privacy: design methods and optimization. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870515000141

[5] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things," *Computer Networks*, vol. 57, no. 10, pp. 2266 – 2279, 2013, towards a Science of Cyber SecuritySecurity and Identity Architecture for the Future Internet.

[6] Z. Yan, P. Zhang, and A. V. Vasilakos, "A survey on trust management for Internet of Things," *Journal of Network and Computer Applications*, vol. 42, pp. 120 – 134, 2014.

[7] S. Gusmeroli, S. Piccione, and D. Rotondi, "A capability-based security approach to manage access control in the Internet of Things," *Mathematical and Computer Modelling*, vol. 58, no. 5ï¿½6, pp. 1189 – 1205, 2013, the Measurement of Undesirable Outputs: Models Development and Empirical Analyses and Advances in mobile, ubiquitous and cognitive computing.

[8] J. L. Hernández-Ramos, A. J. Jara, L. Marın, and A. F. Skarmeta, "Distributed capability-based access control for the internet of things," *Journal of Internet Services and Information Security (JISIS)*, vol. 3, no. 3/4, pp. 1–16, 2013.

[9] Z. Shelby, K. Hartke, and C. Bormann, "The constrained application protocol (coap)," Internet Requests for Comments, RFC Editor, RFC 7252, June 2014, http://www.rfc-editor.org/rfc/rfc7252.txt.

[10] S. Gerdes, O. Bergmann, and D. C. Bormann, "Delegated CoAP Authentication and Authorization Framework (DCAF)," Internet Engineering Task Force, Internet-Draft draft-gerdes-ace-dcaf-authorize-04, oct 2015, work in Progress.

[11] C. Bormann and P. Hoffman, "Concise binary object representation (cbor)," Internet Requests for Comments, RFC Editor, RFC 7049, October 2013.

[12] Z. Su and F. Biennier, "On attribute-based usage control policy ratification for cooperative computing context," *CoRR*, vol. abs/1305.1727, 2013. [Online]. Available: http://arxiv.org/abs/1305.1727

[13] J. Astorga, E. Jacob, M. Huarte, and M. Higuero, "Ladon: end-to-end authorisation support for resource-deprived environments," *IET Information Security*, vol. 6, no. 2, pp. 93–101, June 2012.

[14] MEMSIC's TelosB mote (TPR2420CA) datasheet. [Online]. Available: http://www.memsic.com/userfiles/files/Datasheets/WSN/telosb_datasheet.pdf

[15] "MEMSIC's IRIS mote (XM2110CA) datasheet," http://www.memsic.com/userfiles/files/Datasheets/WSN/IRIS_Datasheet.pdf.

[16] Contiki: The Open Source OS for the Internet of Things. [Online]. Available: http://www.contiki-os.org/

[17] T. Winter, P. Thubert, A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, and R. Alexander, "Rpl: Ipv6 routing protocol for low-power and lossy networks," Internet Requests for Comments, RFC Editor, RFC 6550, March 2012, http://www.rfc-editor.org/rfc/rfc6550.txt. [Online]. Available: http://www.rfc-editor.org/rfc/rfc6550.txt

[18] Tiny AES128 in C. [Online]. Available: https://github.com/kokke/tiny-AES128-C

[19] W. Stallings and T. Case, *Business Data Communications: Infrastructure, Networking, and Security*, 7th ed. Pearson Education Limited, 2013, no. draft-ersue-constrained-mgmt-03, internet-draft 2, pp. 57–84.

[20] A. F. Skarmeta, J. L. Hernandez-Ramos, and M. V. Moreno, "A decentralized approach for security and privacy challenges in the Internet of Things," in *Internet of Things (WF-IoT), 2014 IEEE World Forum on*, March 2014, pp. 67–72.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6520

UGR'16: Un nuevo conjunto de datos para la evaluación de IDS de red

Gabriel Maciá Fernández*[1], José Camacho[1], Roberto Magán-Carrión[1], Marta Fuentes-García[1], Pedro García-Teodoro[1]
[1]Departamento de Teoría de la Señal, Telemática y Comunicaciones - CITIC, Universidad de Granada - España
C/ Periodista Daniel Saucedo Aranda, s/n 18071 Granada
{gmacia,josecamacho,rmagan,nmfuentes,pgteodor}@ugr.es

Roberto Theron[2]
[2]Universidad de Salamanca - España
theron@usal.es

***Resumen*—La evaluación de algoritmos y técnicas para implementar sistemas de detección de intrusiones depende en gran medida de la existencia de conjuntos de datos (*dataset*) bien diseñados. En los últimos años, se ha realizado un gran esfuerzo para construir estos *datasets*. En este trabajo se presenta un nuevo *dataset* que se construye a partir de tráfico real y donde se realizan ataques actualizados. La principal ventaja de este conjunto de datos sobre otros previos es su utilidad para la evaluación de IDSs donde se considera la evolución a largo plazo y la periodicidad del tráfico. También permite entrenar y evaluar modelos que contemplen las diferencias entre día/noche o entre días laborables/fines de semana.**

***Palabras Clave*—seguridad en redes, dataset, IDS, tráfico de red, netflow**

I. INTRODUCCIÓN

Los Sistemas de Detección de Intrusiones (IDS) aparecieron en la esfera de la seguridad como una solución al problema de identificar actividades maliciosas en redes y sistemas. En pocas palabras, un IDS consta de un módulo encargado de la obtención de datos, un módulo de pre-procesamiento que adapta esos datos para los siguientes pasos en el sistema, y un módulo de decisión capaz de determinar si un evento debe ser considerado malicioso o no.

Existen varios tipos de IDS [1]: los *IDSs basados en red* (NIDS) monitorizan eventos de red como flujos o logs de cortafuegos, entre otros, mientras que los *IDS basados en host* (HIDS) consideran eventos relacionados con el sistema, por ejemplo *syslog*, monitorización de sistemas de archivos, carga de la CPU, etc. Los IDS también se clasifican de acuerdo al proceso de detección. Así, los *IDS basados en firmas* (S-IDS) hacen uso de reglas para decidir si un comportamiento observado es malicioso o no, mientras que los *IDS basados en anomalías* (A-IDS) [2] construyen un modelo a partir de datos de entrenamiento y consideran que cualquier comportamiento que se desvía de este modelo es anómalo. Hay que destacar que, aunque existe una diferencia semántica entre un comportamiento anómalo y uno malicioso, un A-IDS los considera equivalentes.

Un problema esencial cuando se evalúan las capacidades de los IDS es la necesidad de un conjunto de datos representativo que permita la comparación entre distintas propuestas. En los años 90 DARPA llevó a cabo un proyecto para construir un conjunto de datos con este fin, generándose los *datasets* ***DARPA'98*** y ***DARPA'99*** del MIT. [3]. Después de ser utilizados y estudiados ampliamente por varios autores, se identificaron algunas limitaciones, como la existencia de registros duplicados, muestras no balanceadas entre ataques y conexiones normales, y otras limitaciones inherentes por considerar tráfico sintético. Desde entonces, muchos otros investigadores y proyectos han intentado proporcionar versiones mejoradas de estos conjuntos de datos, como NSL-KDD, o construir nuevos *datasets*.

Más recientemente se han propuesto otros conjuntos de datos. Por ejemplo, ***UNB ISCX 2012***, creado en 2012 por Shiravi *et al.* [4]. La contribución más relevante de este trabajo es el uso de perfiles para la generación de tráfico. Los autores definen ciertos perfiles α para el tráfico de ataque y perfiles β para el tráfico de *background*. Implementan su propuesta en una red con 17 estaciones de Windows XP y un único ordenador Windows 7, capturando datos durante 7 días. El principal inconveniente de este *dataset* en la actualidad son su duración reducida, el uso de algunos sistemas operativos anticuados (Windows XP), y el uso de

tráfico sintético. ***UNSW-NB15*** fue propuesto por Moustafa *et al.* [5] en 2015. Los autores utilizaron una herramienta de generación automático de ataques llamada *IXIA Perfect-Storm* para implementar nueve familias de ataques reales y actualizados contra varios servidores. Capturaron las trazas *tcpdump* del tráfico de red en una duración total de 31 horas a comienzos de 2015, obteniendo 2 millones de flujos. A partir de estas trazas, se construyó un *dataset* de 49 características para cada flujo. El principal problema de este conjunto de datos es la generación sintética de tráfico, que está asociada a comportamientos teóricos en lugar de realistas en Internet.

Además, existen distintos conjuntos de datos que son específicos de ciertas áreas. Por ejemplo, se están construyendo nuevos *datasets* para sistemas de control industrial (SCADA) [6] [7].

A pesar de este gran número de esfuerzos para contribuir con un conjunto de datos definido para la evaluación de IDS, y después de haber aprendido varias lecciones, es posible constatar que, hasta el momento, todas ellas son soluciones parciales. En una primera revisión, se puede comprobar que muchos de los conjuntos recientes carecen de tráfico real o estrategias de ataque actualizadas. Otra limitación importante está relacionada con la duración de las capturas de datos. Esto es, para hacer posible la evaluación de algoritmos de detección que consideran la evolución ciclo-estacionaria del tráfico, es decir, las diferencias en el tráfico entre día/noche o laborables/festivos, se necesita una traza de larga duración.

Un problema adicional es el del *nivel de dificultad* del *dataset* [8]: si algoritmos de detección simples proporcionan buenos resultados de detección, el nivel de dificultad del *dataset* es bajo. Se puede verificar que algunos de los algoritmos propuestos en los últimos años proporcionan tasas de detección próximas al 100% con índices de falsos positivos realmente bajos. Sin embargo, los sistemas de detección reales todavía están muy lejos de funcionar tan bien. Esto provoca la sospecha de que el problema podría no residir solamente en el diseño de los algoritmos, sino en los conjuntos de datos utilizados para la evaluación.

En este trabajo, se describe un nuevo *dataset* (el conjunto de datos UGR'16[1]) que contiene trazas reales de *netflow* anonimizadas capturadas en un ISP Tier-3 durante 4 meses. En este conjunto, se han incluido escenarios de ataque realistas y se ha llevado a cabo el etiquetado el tráfico. Además, se demuestra que el nivel de dificultad del *dataset* es suficientemente elevado para probar nuevos algoritmos de detección.

El documento está estructurado como sigue. En la Sección II se describe cómo se ha construido el *dataset* y la metodología seguida para insertar tráfico de ataque. En la Sección III se analiza el conjunto de datos, discutiendo y proporcionando una descripción global de la información contenida en él. A continuación, se discute el proceso de etiquetado y se realiza una evaluación con algoritmos de detección del estado del arte para comprobar el nivel de dificultad del *dataset*. Finalmente, se presenta la Sección V de conclusiones.

[1]Tomamos este nombre para el *dataset* del acrónimo de Universidad de Granada

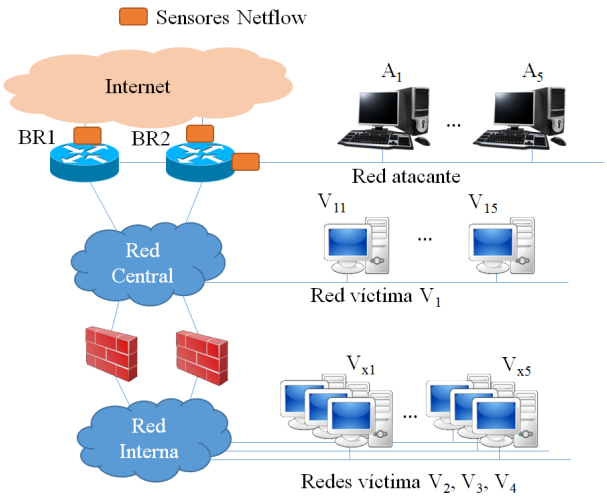


Fig. 1. Topología de la red.

II. METODOLOGÍA PARA LA GENERACIÓN DEL CONJUNTO DE DATOS

En lo que sigue, se describe la metodología seguida para producir el conjunto de datos. Además, se proporciona una extensa descripción de todos los detalles relevantes de cara a su utilización en la evaluación de un IDS.

A. *Infraestructura de red*

Los datos se obtienen de una red real de un ISP Tier-3. El ISP es un proveedor de servicios en la nube, por lo que algunos de los servicios típicos implementados en la red están virtualizados. Algunos de los servicios típicos de alojamiento que se encuentran son webs con configuraciones propietarias o estándares, por ejemplo Joomla o Wordpress, correo electrónico, servidores FTP y DNS, etc. Esta red se utiliza por muchas compañías que tienen tamaños dispares y que se centran en una gran variedad de mercados. Se espera así que el tráfico que atraviesa la red sea muy heterogéneo, pues incluye tanto accesos de clientes a Internet como recepción de tráfico por servidores típicos. Por tanto, una potencial ventaja de esta traza sobre otros conjuntos de datos es su representatividad de un amplio subconjunto de usuarios de Internet. Muchas otras bases de datos solamente recolectan tráfico de universidades o centros de investigación, donde sólo se presentan patrones de tráfico específico.

La topología esquemática de la red del ISP y la infraestructura usada para la recolección de datos se muestran en la Fig. 1. Los principales elementos son:

- Dos *routers* frontera redundantes, *BR1* y *BR2*, proporcionan acceso a Internet. En cada una de sus interfaces de red salientes, se configura un sensor de *netflow* que permite la recolección de todas las conexiones de entrada y salida. Nótese que, por razones de privacidad y volumen, no se proporciona información de carga útil, de modo que la información

se almacena con el formato de archivos *netflow* en lugar de *pcap* (*tcpdump*).
- El ISP tiene dos sub-redes distintas. Una denominada *red central*, donde se localizan los servicios que no están protegidos por cortafuegos. La segunda es la *red interna*, donde se proporcionan servicios de cortafuegos a los clientes.
- Una *red de atacantes* con 5 máquinas se despliega en el nivel superior. Nos referimos a estas máquinas como A_1-A_5.
- En la red central, se configuran 5 máquinas víctima que se utilizan solamente para la recogida de datos. Se colocan junto con otros clientes reales en una red existente que llamamos *red víctima* V_1. Estas máquinas se denominan V_{11}-V_{15}.
- En lo referente a la red interna, se emplazan un total de 15 máquinas víctima adicionales en tres redes existentes y distintas, cada una con 5 máquinas. Llamamos a esas redes: *red víctima* V_2 (V_{21}-V_{25}), *red víctima* V_3 (máquinas V_{31}-V_{35}) y *red víctima* V_4 (V_{41}-V_{45}).

B. Generación de tráfico de ataque

Algunos *datasets* existentes, como MAWILab [9] o CAIDA [10], solamente consideran el tráfico real capturado de ciertos sensores de red. Aunque es una ventaja para modelar el tráfico de *background*, esto conlleva ciertas limitaciones en la identificación de ataques. De hecho, el etiquetado de tráfico real implica la necesidad de confirmar que las conexiones señaladas como ataques son realmente tráfico malicioso. Por esta razón, se decide combinar tráfico de *background* real (que probablemente contenga instancias de ataque) con ataques que se generan intencionadamente para el experimento.

Con este objetivo, las 25 máquinas virtuales mencionadas se instalan con una configuración similar a las proporcionadas para los clientes ISP, es decir, se implementan servidores web, DNS, FTP y de correo electrónico. Las máquinas virtuales A_1 a A_5 se utilizan para lanzar un número de ataques específicos a lo largo del tiempo contra el resto ($V_{x1} - V_{x5}$, con $x = \{1 - 4\}$), que juegan el rol de víctimas de los ataques. Como se puede observar en la Fig. 1, tanto atacantes como víctimas se encuentran dentro de la infraestructura ISP para evitar la potencial detección y bloqueo de los ataques por otros ISP intermedios. Además, la red de atacantes se ubica en el *router* frontera para simular que el tráfico de ataque procede de Internet.

Implementación de ataques. Debido a que solamente se recoge tráfico de *netflow*, y por tanto, no se considera el *payload* de la información en la traza, no se incluyen tipos de ataque susceptibles de ser detectados mediante análisis del *payload*. Solamente se consideran ataques relacionados con la red. Los tipos de ataque implementados son:

- *DoS de baja tasa*: Se envían paquetes TCP SYN a las víctimas utilizando la herramienta `hping3`. El puerto destino es el 80, por lo que el tráfico se mezcla con el tráfico web de *background* real. El tamaño de cada paquete es de 1280 bits y la tasa es de 100 paquetes/s. Como puede comprobarse, la tasa de ataque es baja, por lo que no se afecta la operación normal de la red. Se consideran tres escenarios distintos de ataque:
  - *DoS11*: Ataque DoS uno-a-uno, donde el atacante A_1 ataca a la víctima V_{21}. La duración total de *DoS11* es de 3 minutos.
  - *DoS53s*: Los cinco atacantes A_1-A_5 atacan a tres de las víctimas, cada una en una red diferente, durante 3 minutos. En particular, estos ataques siguen esta estructura: (A_1, A_2) $\rightarrow$ V_{21}, (A_3, A_4) $\rightarrow$ V_{31} and A_5 $\rightarrow$ V_{41}. La letra '*s*' al final del nombre del ataque representa 'síncrono', lo que significa que los ataques se inician por todos los atacantes al mismo tiempo. Debido a esta sincronización, la duración de *DoS53s* es de 3 minutos también.
  - *DoS53a*: Los ataques se ejecutan como en *DoS53s*, pero ahora cada víctima es seleccionada secuencialmente, siendo atacada durante 3 minutos con un periodo de inactividad de 30 segundos entre los tres ataques. De esta forma, la duración total de *DoS53a* es de 10 minutos. En este caso, la letra '*a*' al final del nombre del ataque representa 'asíncrono'.
- *Escaneo de puertos*: Se ejecuta un escaneo SYN continuo a los puertos comunes de las víctimas durante 3 minutos, utilizando la herramienta `nmap`. Se implementan dos variantes para este ataque:
  - *Scan11*: Ataque de escaneo uno-a-uno, donde el atacante A_1 escanea a la víctima V_{41}.
  - *Scan44*: Ataque de escaneo cuatro-a-cuatro, donde los atacantes A_1, A_2, A_3 y A_4 inician un escaneo al mismo tiempo a las víctimas V_{21}, V_{11}, V_{31} y V_{41}, respectivamente. Como los ataques se llevan a cabo en paralelo (comienzan en el mismo instante), la duración total es de 3 minutos.
- *Actividad relacionada con una botnet*. Se simula tráfico de *botnet* mediante la exfiltración de datos desde algunas máquinas infectadas al puerto 80 de un *botmaster* localizado en A_1. Se consideran veinte *bots*, correspondientes a todas las máquinas víctima. Cada uno de los *bots* lleva a cabo estas variantes de exfiltración:
  - *Exf1KB*: Se envía un fragmento de información de 1KB al *botmaster*.
  - *Exf1MB*: Se envía un total de 1MB de información al *botmaster* en una única conexión.
  - *Exf1MBp*: El fragmento de información de 1MB a enviar al *botmaster* se divide en trozos de 1KB cada uno, y se envía al *botmaster* en conexiones distintas.

  De nuevo, la transmisión desde cada *bot* puede ser *síncrona* (sufijo *'s'*), lo que significa que todos ellos inician la transmisión de información al mismo tiempo, o *asíncrona* (sufijo *'a'*), donde cada uno de

Tabla I
PROGRAMACIÓN PLANIFICADA PARA CADA UNO DE LOS ATAQUES EN EL INTERVALO DE TIEMPO DE 2H COMENZANDO EN t_0.

Instante de inicio	Ataque	Duración
$t_0 + 0h00m$	DoS11	3m
$t_0 + 0h10m$	DoS53s	3m
$t_0 + 0h20m$	DoS53a	10m
$t_0 + 0h40m$	Scan11	3m
$t_0 + 0h50m$	Scan44	3m
$t_0 + 1h00m$	Exf1KBs	$\leq$10m
$t_0 + 1h10m$	Exf1MBs	$\leq$10m
$t_0 + 1h20m$	Exf1MBps	$\leq$10m
$t_0 + 1h30m$	Exf1KBa	$\leq$10m
$t_0 + 1h40m$	Exf1MBa	$\leq$10m
$t_0 + 1h50m$	Exf1MBpa	$\leq$10m

ellos selecciona un instante aleatorio en una ventana de 3 minutos para empezar el correspondiente procedimiento de exfiltración. En cualquier caso, es importante mencionar que los ataques relacionados con *botnet* en todas las variantes se restringieron a una duración menor de 10 minutos.

Programación de la ejecución de ataques. El tráfico de ataque se genera en lotes de 2 horas. En cada lote de ataque, se ejecutan todas las variantes de ataque siguiendo dos patrones posibles de programación:

1) *Programación planificada*: Se ejecuta cada ataque en el lote en un instante fijo y conocido dado por un desplazamiento desde el instante inicial de tiempo t_0. Los desplazamientos para los distintos ataques se muestran en la Tabla I. Nótese que no hay solapamiento en el tiempo para los distintos tipos de ataque.
2) *Programación aleatoria*: El instante inicial para la ejecución de cada uno de los ataques se selecciona aleatoriamente entre $t_0 + 00h00m$ y $t_0 + 01h50m$, por lo que se restringe la duración total del lote a un máximo de dos horas. En este caso, podría existir solapamiento temporal entre los ataques, lo que permitirá comprobar la adecuación de los detectores de anomalías cuando esta situación aparece.

Merece la pena destacar que los ataques se lanzaron mientras que el tráfico real de *background* atravesaba la red. De esta forma, el tráfico capturado por los sensores para la correspondiente ventana de monitorización incluirá instancias de tráfico relacionado tanto con ataques como con tráfico normal. Así, para permitir el estudio del tráfico de *background* para distintas horas del día junto con tráfico de ataque, se lanzan lotes de ataque durante 12 días consecutivos desde el comienzo del experimento de ataque, por lo que se cubren todas las horas posibles del día. Cada día en el experimento de ataque, se lanza un lote de programación planificada en el instante t_0, seguido de un lote de programación aleatoria que se inicia en t_0+12h. En cada día siguiente durante 12 días, t_0 se incrementa con un desplazamiento de $2h$. En la Tabla II se pueden ver los diferentes instantes de tiempo y fechas seleccionados para la ejecución de los lotes planificados y aleatorios.

Tabla II
FECHA Y HORA PARA LA EJECUCIÓN DE DISTINTOS LOTES DE ATAQUE EN EL CONJUNTO DE DATOS.

Fecha	Planificada	Aleatoria
Jue, 28/07/2016	00:00	12:00
Vie, 29/07/2016	02:00	14:00
Sab, 30/07/2016	04:00	16:00
Dom, 31/07/2016	06:00	18:00
Lun, 01/08/2016	08:00	20:00
Mar, 02/08/2016	10:00	22:00
Mie, 03/08/2016	12:00	N/A
Jue, 04/08/2016	14:00	00:00
Vie, 05/08/2016	16:00	02:00
Sab, 06/08/2016	18:00	04:00
Dom, 07/08/2016	20:00	06:00
Lun, 08/08/2016	22:00	08:00
Mar, 09/08/2016	N/A	10:00

C. Capturas del dataset

Los flujos se capturan y transfieren desde los sensores indicados en la red utilizando el formato Netflow v9 (ver Fig. 1). Los parámetros por defecto se mantuvieron durante la configuración *netflow* de los *routers* Cisco, esto es, el temporizador inactivo es de 15 segundos, mientras que el cronómetro activo para flujos es de 30 minutos.

El *dataset* completo contiene dos capturas distintas: un *conjunto de calibración* y un *conjunto de prueba*. El *conjunto de calibración* dura 100 días, y se extiende desde 10:52-18/03/2016 hasta 18:27-26/06/2016. Su principal propósito es ayudar a la construcción y calibración de modelos de normalidad, principalmente porque no se generaron ataques de forma artificial. Nótese que esto no implica que no haya presencia de ataques, como se discutirá a continuación en la Sección III.B.

Aunque la captura del *conjunto de calibración* se programó de forma automática para ser continua a lo largo del tiempo, se interrumpió dos veces por el ISP para llevar a cabo procedimientos de mantenimiento específicos de la red. Estos dos intervalos son:

1) [02:00 27/03/2016 — 03:00 27/03/2016]
2) [00:00 01/04/2016 — 17:20 06/04/2016]

El *conjunto de prueba* dura aproximadamente un mes, desde 13:43-27/07/2016 hasta 09:27-29/08/2016. Durante esta captura, los lotes de ataque se ejecutaron comenzando en 00:00-28/07/2016 y finalizando en 12:00-09/08/2016, como se muestra en la Tabla II. Esta captura pretende ser utilizada para la validación de los algoritmos de detección.

La Tabla III resume las diferentes características para los conjuntos tanto de *calibración* como de *prueba*. Se pueden ver las fechas para las dos capturas y el número y tamaño de los archivos incluidos en cada una. Adicionalmente, se muestra el número de conexiones incluidas en los dos conjuntos. Este número es mucho mayor que en los demás conjuntos de datos.

D. Pre-procesamiento y disponibilidad del dataset

Los archivos de formato binario `nfcapd` recolectados para los conjuntos de calibración y de prueba se agrupan

Tabla III
CARACTERÍSTICAS DE LOS CONJUNTOS DE CALIBRACIÓN Y PRUEBAS.

Característica	Calibración	Prueba
Inicio de la captura	10:47h 18/03/2016	13:38h 27/07/2016
Fin de la captura	18:27h 26/06/2016	09:27h 29/08/2016
Inicio de los ataques	N/A	00:00h 28/07/2016
Fin de los ataques	N/A	12:00h 09/08/2016
Número de archivos	17	6
Tamaño (comprimido)	181GB	55GB
# Conexiones	$\approx$ 13.000M	$\approx$ 3.900M

Tabla IV
CORRESPONDENCIA DE DIRECCIÓN IP ANONIMIZADA CON LAS MÁQUINAS EN LA CONFIGURACIÓN EXPERIMENTAL.

Máquinas/s	Direccion/es IP
A_1 - A_5	42.219.150.{*246,247,243,242,241*}
V_{11} - V_{15}	42.219.156.{*30,31,29,28,27*}
V_{21} - V_{25}	42.219.158.{*16,17,18,19,21*}
V_{31} - V_{35}	42.219.152.{*20,21,22,23,18*}
V_{41} - V_{45}	42.219.154.{*69,68,70,71,66*}

en un único archivo por semana para los dos periodos de captura. El tamaño medio de los diferentes archivos está en torno a $14GB$ en formato de compresión `tar`. En el conjunto de calibración hay 17 archivos, mientras que para el conjunto de prueba hay 6 archivos disponibles. Todos estos archivos están disponibles para descargar en nuestra página web: https://nesg.ugr.es/nesg-ugr16/.

La información disponible se corresponde con trazas *netflow* tanto en formato `nfcapd` como `csv`, éste último obtenido del posprocesado de los ficheros `nfcapd` mediante la herramienta `nfdump`. Las características que han sido seleccionadas para el formato `csv` son[2]: *timestamp* del final de un flujo (*te*), duración del flujo (*td*), dirección IP de origen (*sa*), dirección IP de destino (*da*), puerto origen (*sp*), puerto destino (*dp*), protocolo (*pr*), banderas (*flg*), estado de reenvío (*fwd*), tipo de servicio (*stos*), paquetes intercambiados en el flujo (*pkt*), y su correspondiente número de bytes (*byt*).

Las direcciones IP de las distintas máquinas en el *dataset* se han anonimizado utilizando CryptoPan, que proporciona anonimización preservando los prefijos de las direcciones IP [11], implementada en la herramienta `nfanon` [12]. Esta herramienta se ha utilizado tradicionalmente para la anonimización de las trazas en los conjuntos de datos CAIDA. La correspondencia entre direcciones de IP anonimizadas en las distintas redes víctimas y atacantes se muestra en la Tabla IV.

III. ETIQUETADO Y ANÁLISIS DEL CONJUNTO DE DATOS

A continuación se analizan los datos recogidos en las dos capturas de datos, Calibración y Prueba, cuyo fin es aportar ideas y resultados de cara a su evaluación por IDS. Después, se describe el proceso de etiquetado, indicando los procedimientos seguidos y sus limitaciones. Finalmente, se evalúa el nivel de dificultad asociado al *dataset*, mediante la utilización de tres algoritmos del estado del arte en detección de intrusiones basados en anomalías.

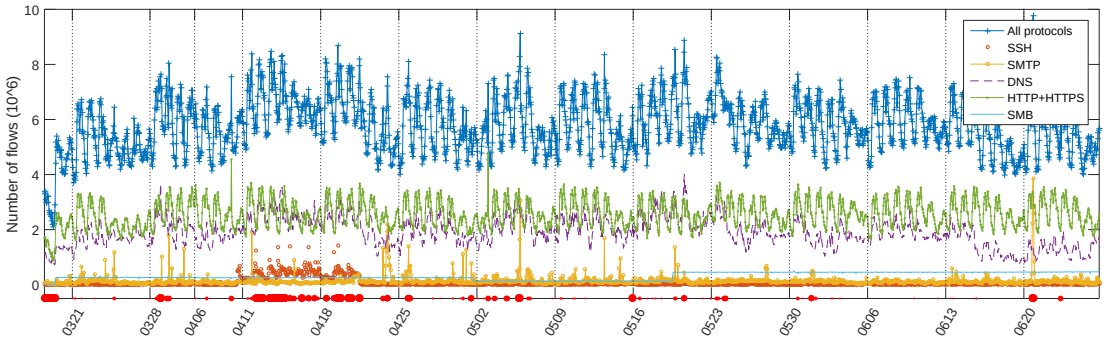
A. Cifras del dataset

Como se muestra en la Tabla III, el número de flujos total en el conjunto de datos es superior a 16.900 millones, y la duración de la traza es de más de 4 meses. Esto permite la evaluación de algoritmos de detección que hacen uso de la evolución ciclo-estacionaria del tráfico en diferentes patrones día/noche, así como fases días laborables/fin de semana. El número de IP externas observadas en la traza es mayor que 600 millones, correspondiente aproximadamente a 10 millones de sub-redes distintas.

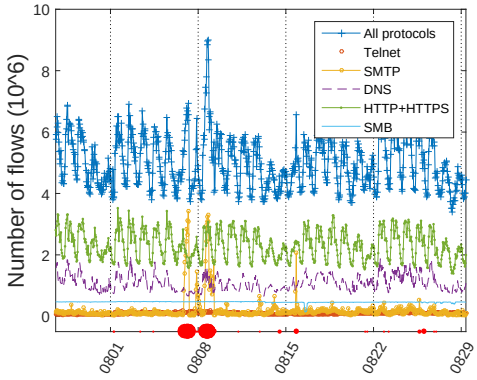
En la Fig. 2 se puede ver la evolución del número de flujos (en millones de flujos) para los conjuntos de calibración y de prueba. Las líneas punteadas verticales separan las distintas semanas (y por tanto archivos) en el conjunto de datos. Se señalan los protocolos con mayor cantidad de flujos. Los puntos rojos en el eje X muestran las anomalías encontradas por los tres detectores de anomalías utilizados. El tamaño del círculo está relacionado con el número de anomalías obtenidas en el intervalo (ver la discusión de este punto en la Sección III.A). A partir de estos resultados se pueden derivar algunas conclusiones interesantes:

- Se confirma que, tal y como se esperaba, predomina el protocolo HTTP/S y el tráfico DNS sigue un patrón coherente.
- Debido al hecho de que la mayoría de los clientes son compañías que alojan sus servidores en la red del ISP, el tráfico de BitTorrent o de otros servicios de P2P es casi residual.
- Nótese que hay un incremento de tráfico SSH en el intervalo 11/04/16 - 18/04/16. Se ha comprobado manualmente que este incremento de tráfico se debe a un ataque de escaneo SSH procedente de una única máquina alojada en el ISP. Las víctimas del ataque tienen un amplio rango de IP localizadas en un país de Sudamérica.
- Sorprendentemente, a pesar de su naturaleza insegura, se observa que muchas compañías todavía utilizan el servicio Telnet para gestionar sus equipos (ver Fig. 2(b)).
- Hay picos en el tráfico de SMTP dispersos a lo largo del tiempo en las dos capturas. Estos picos se refieren a veces a campañas de correo electrónico procedentes de compañías legítimas (bancos, servicios online, etc.), pero también se han encontrado campañas de *spam*. Por ejemplo, se identifica el pico en 20:39-06/08/16 - 05:59-07/08/16. Este se refiere a 12,5 millones de conexiones SMTP desde 5 IP públicas utilizando servidores de Yahoo. Un análisis minucioso de este tráfico lleva a concluir que sigue el patrón de una campaña de *spam*. Cada

[2]Se indica con paréntesis el nombre de las variables como viene dado por la herramienta `nfdump` para facilitar la identificación en el *dataset*.



(a) Conjunto de Calibración



(b) Conjunto de Prueba

Fig. 2. Evolución del número de flujos para (a) conjunto de Calibración y (b) conjunto de Prueba.

máquina generó alrededor de 2,5 millones de correos electrónicos. El equipo de IT del ISP confirmó que éste fue un cliente que alquiló máquinas virtuales durante 2 meses. Las IPs contratadas acabaron en las listas negras de Yahoo, y esta es probablemente la razón por la que el cliente dejó de alquilar las máquinas virtuales.

- Nótese el patrón constante para tráfico SMB. Este tráfico se corresponde con una sola empresa distribuidora con muchas tiendas. Cada establecimiento se comunica con un servidor central de forma periódica para comprobar el estado de algunos servicios o datos. Los picos en el tráfico de SMB se deben a la cancelación de la suscripción de ciertas tiendas a ese servicio.

B. Etiquetado

El etiquetado es un tema crítico en la producción de conjuntos de datos para la evaluación de IDS. Cuando se maneja tráfico real, es un reto decidir si un flujo dado corresponde a un ataque o no. Incluso cuando la naturaleza de un flujo es evaluada por un experto, hay ciertos casos en los que esto no está claro. Por esta razón, algunos autores deciden generar únicamente tráfico sintético, pues sólo en estos entornos se puede determinar de forma certera qué flujos se deben a un ataque y cuáles son legítimos. El problema con este enfoque es que el tráfico de *background* no es representativo del comportamiento de redes reales, por lo que se abre la posibilidad de que los algoritmos de IDS estén sesgados hacia la detección de falsos escenarios.

En nuestro caso, como se discutió anteriormente, se ha optado por utilizar tráfico de ataque artificial entrelazado con tráfico real de *background*. Este hecho supone un problema cuando se intenta etiquetar un conjunto de datos. El tráfico de *background* no está libre de ataques que podrían ser ejecutados por terceras partes durante la captura del *dataset* y, como ya se ha dicho, no está claro cómo identificarlos en algunos casos. Esto no ocurre cuando generamos ataques artificialmente. Incluso cuando se utilizan herramientas de generación de ataques reales y los escenarios de ataque son también reales, es posible etiquetar estos flujos, ya que se conocen las reglas para identificarlos (direcciones de IP y puertos, marcas de tiempo, etc.).

Algunos autores [13] que utilizan este mismo enfoque decidieron etiquetar flujos o paquetes utilizando tres etiquetas distintas: a) una etiqueta *attack* para los flujos que positivamente conocían que se correspondían con un ataque, b) una etiqueta *normal* para aquellos que se generaron de forma sintética con patrones normales, y c) una etiqueta *background* para aquellos que no sabían exactamente si eran ataques o no. Como en este caso no se está generando tráfico normal de forma sintética, finalmente se ha decidido etiquetar únicamente como *attack* aquellos flujos generados de forma artificial y dejar el resto etiquetados como *background*.

Tabla V
CORRESPONDENCIA ENTRE LAS ETIQUETAS Y ATAQUES EJECUTADOS.

Tipo de ataque	Etiqueta
DoS11	dos
DoS53s	dos
DoS53a	dos
Scan11	scan11
Scan44	scan44
Exf1KB	exf1KB
Exf1MB	exf1MB
Exf1MBp	exf1KB

Las etiquetas se aplican por flujo. Las conexiones son bidireccionales, por lo que una única conexión aparece como dos conexiones distintas unidireccionales. Las etiquetas de ataque se dividen en distintas sub-etiquetas, dependiendo del tipo de ataque que está siendo ejecutado. La Tabla V muestra la correspondencia entre las etiquetas utilizadas en el *dataset* y el tipo de ataque.

Nótese que no todos los ataques implementados (ver la Sección II.B) han sido etiquetados de forma individual. En su lugar, algunos de ellos han sido agrupados como uno solo, debido al hecho de que su patrón de tráfico es realmente el mismo. Tal es el caso de *DoS11*, *Dos53s* y *DoS53a*, todos ellos etiquetados como DoS. De manera similar, *Exf1KB* y *Exf1MBp* han sido etiquetados como *Exf1KB*, ya que realmente siguen el mismo patrón, esto es, los flujos *Exf1MBp* aparecen como muchas conexiones *Exf1KB* consecutivas.

IV. ANOMALÍAS EN EL CONJUNTO DE DATOS

A continuación se trata de ilustrar el aspecto de las 'anomalías' en el tráfico de *background* del *dataset* UGR'16. Nuestra única intención es motivar a otros investigadores a la identificación de estas anomalías utilizando sus propios métodos y algoritmos de detección. Para encontrar estas anomalías, se ha utilizado el detector de anomalías MSNM [14] y herramientas de análisis visual ad hoc con el propósito de describir, explorar y analizar los datos para descubrir el conocimiento subyacente de estos datos [15]. Adicionalmente, se ha contado con la ayuda del personal del ISP. Con fines ilustrativos nos centramos en la anomalía que tiene lugar en el intervalo de tiempo entre 04:10-01/08/16 hasta 04:14-01/08/16. Se observa un incremento de paquetes ACK y conexiones muy cortas que utilizan UDP. Inspeccionando los registros de Netflow para este periodo de tiempo se encuentra una única IP que crea 867.405 conexiones únicamente desde cuatro puertos origen (5061, 5062, 5066 y 5068) desde Alemania. Los destinos son 4.097 equipos distintos distribuidos en 16 sub-redes diferentes (máscara /24). Dependiendo del puerto origen de la conexión, cada *host* víctima se escanea en un rango específico de 60 puertos (por ejemplo, desde el puerto origen 5068 se escanearon los puertos 5000-5059). Debido a estos patrones de conexión, se concluye que parece tratarse de un ataque provocado por un *malware* orientado al escaneo de una vulnerabilidad específica.

A. Nivel de dificultad del dataset

Finalmente, estamos interesados en evaluar el nivel de dificultad del *dataset*. Recordemos que esta es una característica cualitativa que evalúa la probabilidad de que algoritmos de detección simple proporcionen muy buenos resultados de detección [8]. Un nivel de dificultad bajo no supondría un reto para el desarrollo de nuevos algoritmos que mejoren los existentes.

El *dataset* se ha evaluado utilizando tres detectores de anomalías del estado del arte. El primer par de detectores, MSNM_C y MSNM_S, se proponen en la metodología de monitorización de redes estadística multivariante basada en PCA dada en [14]. MSNM_C preprocesa los datos para centrarlos, mientras que MSNM_S los autoescala. Ambos esquemas muestran un rendimiento ligeramente distinto dependiendo de los tipos de ataque considerados [14]. Estos detectores han demostrado exhibir un comportamiento mejorado en varios escenarios distintos [16].

El tercer detector implementado es una máquina de vectores de soporte de una clase, OCSVM [17] [18]. OCSVM es un método de detección de anomalías de red basado en clasificación que se dice proporciona excelentes resultados según distintos estudios [19].

Para estos tres detectores se han obtenido las curvas ROC utilizando los datos de calibración obtenidos durante el mes de junio para la creación de los modelos de normalidad. Las ROC obtenidas se muestran en la Fig. 3. En la figura, se muestran dos tipos de conjuntos de curvas ROC distintas. Por una parte, las series MSNM_C, MSNM_S y OCSVM se corresponden con los resultados obtenidos cuando sólo se consideran ataques artificiales en la *ground truth*.

Los malos resultados relativos a las citadas ROC se deben principalmente al hecho de que no se han considerado todas las anomalías en el tráfico de *background*. Para resolver este problema, se ha identificado como anómalo todo el tráfico de *background* que activa una alarma en los tres detectores de anomalías al mismo tiempo. Para ello, primero se ha seleccionado el mejor punto de operación para cada detector de acuerdo con el criterio de Youven [20] (círculos rojos en la Fig. 3). Este criterio selecciona el umbral del detector de anomalías con la mayor distancia a la diagonal en la curva ROC.

Frente a esta experimentación, también se han añadido las anomalías detectadas por los tres detectores a la *ground truth* (observar estas anomalías como círculos rojos en el eje X de Fig. 2). Así se han vuelto a calcular las curvas ROC con esta nueva *ground truth*. Los resultados se muestran en la Fig. 3 como la serie $MSNM^*_C$, $MSNM^*_S$ and $OCSVM^*_C$. Se puede ver que incluso en el mejor de los casos, $MSNM^*_C$, los resultados son inferiores al 90% para la tasa de verdaderos positivos cuando el índice de falsos positivos está en torno al 10%, lo cual no es en absoluto excepcional. La principal razón es la dificultad para detectar ataques de exfiltración.

Aunque sería necesaria una evaluación más extensa de los algoritmos y sus resultados, éstos son suficientes para concluir que el nivel de dificultad del *dataset* resulta

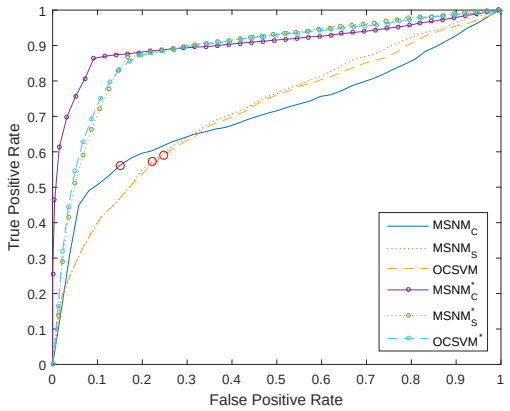


Fig. 3. Curvas ROC para los distintos detectores.

adecuado para la evaluación de nuevos algoritmos de detección.

V. CONCLUSIONES

La principal contribución de este trabajo es la generación de un nuevo *dataset* para la evaluación de algoritmos y sistemas IDS llamado UGR'16. El conjunto de datos se ha construido teniendo en cuenta lo aprendido sobre conjuntos de datos anteriores. UGR'16 es una colección de trazas *netflow* capturadas durante más de 4 meses de tráfico en una red real de un ISP Tier-3, junto con un conjunto de ataques de red de tipo real que se ha diseñado específicamente para entrenar y probar algoritmos IDS.

Las principales ventajas del *dataset* presentado frente a otros ya existentes se enumeran a continuación. Primero, el tráfico de *background* es muy representativo del tráfico de Internet, pues se captura de sensores de una red ISP donde se ubican perfiles muy diferentes de clientes. Esta es una diferencia principal con el resto de conjuntos de datos, en los cuales se recoge tráfico muy específico (como tráfico generado en una universidad o un centro de investigación). Segundo, el *dataset* tiene un nivel de dificultad que permite que nuevos algoritmos puedan ser comparados con otros ya existentes. Tercero, la duración del *dataset* lo hace adecuado para probar algoritmos que consideran la evolución ciclo-estacionaria del tráfico en día/noche y días laborables/fines de semana.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Gobierno Español-MINECO (Ministerio de Economía y Competitividad) y fondos FEDER, a través del proyecto TIN2014-60346-R.

REFERENCIAS

[1] R. Di Pietro and L. V. Mancini, *Intrusion detection systems*. Springer Science & Business Media, 2008, vol. 38.

[2] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 1, pp. 18–28, 2009.

[3] DARPA'98 and DARPA'99 datasets. [Online]. Available: https://www.ll.mit.edu/ideval/docs/index.html

[4] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers and Security*, vol. 31, no. 3, pp. 357–374, 2012.

[5] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, 2015.

[6] A. Lemay and J. M. Fernandez, "Providing scada network data sets for intrusion detection research," in *9th Workshop on Cyber Security Experimentation and Test (CSET 16)*. USENIX Association, 2016.

[7] T. Morris and W. Gao, "Industrial control system traffic data sets for intrusion detection research," in *International Conference on Critical Infrastructure Protection*. Springer, 2014, pp. 65–78.

[8] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*, no. Cisda, pp. 1–6, 2009.

[9] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking," in *ACM CoNEXT '10*, Philadelphia, PA, December 2010.

[10] CAIDA. The cooperative association for internet data analysis. [Online]. Available: http://www.caida.org/

[11] J. Fan, J. Xu, M. H. Ammar, and S. B. Moon, "Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme," *Computer Networks*, vol. 46, no. 2, pp. 253–272, 2004.

[12] P. Haag, "NFDUMP-NetFlow processing tools," *URL: http://nfdump.sourceforge.net*, 2011.

[13] S. Garcia, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Computers & Security*, vol. 45, pp. 100 – 123, 2014.

[14] J. Camacho, A. Pérez-Villegas, P. García-Teodoro, and G. Maciá-Fernández, "PCA-based multivariate statistical network monitoring for anomaly detection," *Computers & Security*, vol. 59, pp. 118–137, 2016.

[15] Juan Alvarado-Pérez and Diego H. Peluffo-Ordóñez and Roberto Therón, "Bridging the gap between human knowledge and machine learning," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 4, no. 1, 2015.

[16] G. Maciá-Fernández, J. Camacho, P. García-Teodoro, and R. A. Rodríguez-Gómez, "Hierarchical PCA-based multivariate statistical network monitoring for anomaly detection," in *Information Forensics and Security (WIFS), 2016 IEEE International Workshop on*. IEEE, 2016, pp. 1–6.

[17] B. Scholkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New Support Vector Algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[18] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[19] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[20] A. K. Akobeng, "Understanding diagnostic tests 3: Receiver operating characteristic curves." *Acta paediatrica*, vol. 96, no. 5, pp. 644–647, 2007.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6562

Establecimiento de claves seguro mediante códigos sonoros en dispositivos móviles

Ricardo Ruiz Tueros, Isaac Agudo
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga
{rrt,isaac}@lcc.uma.es

***Resumen*—El objetivo de este trabajo ha sido investigar el uso de canales sonoros para el intercambio de claves en teléfonos inteligentes. Para ello, se ha realizado un diseño y un análisis del problema a resolver, teniendo en cuenta factores como la necesidad de sincronización, los parámetros de escucha, la longitud de la clave que podemos obtener, los ataques a los que sería vulnerable el protocolo, la necesidad de cierta persistencia a nivel local, el uso de la nube, etc.**

Se ha realizado un análisis de tecnologías relacionadas, eligiendo aquellas que mejor se adaptan a los requisitos del problema y se ha implementado un prototipo capaz de realizar un intercambio de claves y autenticar mutuamente a un par de dispositivos, creando así un canal seguro a través del cual puedan comunicarse en el futuro, sin necesidad de que una tercera parte confiable que certifique la identidad de las partes.

***Palabras Clave*—Autenticación, Seguridad, Sonido , Intercambio de claves, iOS, Nube**

I. INTRODUCCIÓN

La motivación inicial detrás de este trabajo es la implementación de una aplicación para teléfonos inteligentes que permita establecer una clave compartida entre dos dispositivos de forma que los respectivos usuarios tengan un control total sobre el proceso. Para ello se han analizado diferentes aproximaciones al problema y se ha implementado un prototipo que resuelve el problema usando canales sonoros.

Existe una gran variedad de protocolos de intercambio de claves [1] en la literatura, si bien uno de lo más conocidos y usados en la actualidad, por ejemplo en los protocolos TLS e IPSEC, es Diffie Hellman (DH) [2]. Uno de los requisitos para que DH funcione correctamente es que los usuarios implicados en el protocolo tengan la certeza de que han calculado la misma clave compartida, para de esa forma evitar un ataque de Man-in-the-middle (MitM). Una posible solución al problema se basa en la autenticación de las claves públicas DH, tal y como ocurre en TLS cuando se utiliza el intercambio de claves basado en DH. Esto requiere de una tercera parte confiable (Autoridad Certificadora) que certifique la identidad de las partes, mediante un certificado de clave pública, y del establecimiento de relaciones de confianza que pueden ser complejas de administrar. Nuestra propuesta se basa en la utilización de un canal *"fuera de banda"* , que nos permita realizar la autenticación de los dispositivos sin necesidad usar claves autenticadas. Esto se engloba dentro de las técnicas de "Key Fingerprint Verification" usando canales fuera de banda [3].

Nuestro objetivo final no es saber quien es la otra persona, sino tener la certeza de que cuando queramos enviar algo a un dispositivo que ya hemos enlazado previamente, solo ese dispositivo será capaz de leerlo. Por tanto, se cuenta con una primera fase donde la autenticación en el intercambio de claves está fundamentada solo en la proximidad y la capacidad que tienen los usuarios de observar el intercambio de información, y una segunda fase de comunicación donde la autenticación de basa en la clave negociada en la primera fase.

Si bien inicialmente se valoraron diferentes protocolos inalámbricos para comunicaciones de corto alcance (Bluetooh, NFC, etc.) como canal fuera de banda, se descartó esta aproximación debido a dos factores principales:

- El uso de antenas de gran tamaño y/o potencia puede ampliar el rango de comunicaciones, de forma que el sentido de proximidad se pierda.
- Los usuarios no tienen una constancia directa de cuando se están comunicando los dispositivos ni pueden identificar visualmente que dispositivos son los que se están emparejando. Esto último es un problema reconocido en los sistema de pago sin contacto[4].

Existen diversas técnicas para el emparejamiento de dispositivos móviles que no recurren al uso de tecnologías de radio frecuencia. Un ejemplo sería el uso del canal háptico (vibración) [5], es decir, las vibraciones del teléfono. Una de las ventajas o limitaciones, según los requisitos que se tengan en cuenta, de este canal es que solo funciona entre un par de dispositivos y además, estos tienen que estar

en contacto. Esto podría ser bueno porque evitaría ataques de intermediarios pero puede resultar intrusivo al requerir que se coloquen los dos dispositivos en contacto.

Como alternativa se buscó un canal de comunicación aún más básico, las personas intercambian secretos mediante susurros, esto es, en definitiva sonido... ¿Y si un dispositivo pudiera "susurrar" un secreto a otro?. El canal sonoro cumple las condiciones que necesitamos: Es un canal de dispersión entre esos dos dispositivos que puede ser controlado sencillamente, mediante el volumen, y que es percibido por los usuarios de forma directa. Es independiente de la plataforma y fácilmente accesible desde cualquier sistema operativo y hardware, ya que el sonido sigue siendo el mismo (o al menos tan sólo ligeramente distinto), sumado a que hoy en día casi cualquier dispositivo digital incorpora un micrófono y un altavoz lo hacen el canal fuera de banda ideal para nuestro esquema.

Los canales sonoros también se han utilizado con otros propósitos. Por ejemplo, [6] utiliza el sonido ambiente recibido por dos dispositivos móviles como segundo factor en la autenticación, si el sonido ambiente grabado por los dos dispositivos es el mismo, podemos asumir que se encuentran en la misma localización. Si la calidad del sonido ambiente no es buena, se podrían utilizar dispositivos externos que emitan un patrón predecible para los dispositivos cuya escucha garantice que ambos se encuentran en la misma localización. El problema de esta aproximación es que requiere del despliegue de un sistema de "balizas" que emitan este tipo de señales.

II. TECNOLOGÍAS RELACIONADAS

A continuación se analizan algunas tecnologías relacionadas con el trabajo realizado.

Signal360 [1] (Anteriormente SonicNotify) es una tecnología propietaria en la que participa Oracle, permite que el dispositivo reciba notificaciones con información por ultrasonidos, se orienta a la obtención de información adicional de un evento concreto, por ejemplo, los grupos que está actuando en un festival de música o para fines de marketing y publicidad, sorteos en eventos, compartición de imágenes mediante redes sociales, etc... Desgraciadamente al ser una solución propietaria no cuenta con un SDK (Software Development Kit) o librerías de carácter público.

Audio Modem [2] ha sido desarrollada por la empresa francesa Appdilium, en su página se analizan las distintas frecuencias que podrían ser utilizadas y razona la utilización de un rango de frecuencias de 18.4kHz-20.8kHz (ultrasonidos). Además, utilizan su propia implementación para la modulación de onda, siguiendo el algoritmo DBPSK [7] que permite hacer más robusto el canal gracias a la redundancia de datos, así como facilitar la demodulación con un oyente de tipo no coherente.

En la actualidad las tecnologías basadas en ultrasonidos son el objetivo de grandes empresas, prueba de ello es la publicación de la API Nearby de Google para Android e iOS [3] que combina Bluetooth, Bluetooth Low Energy, WiFi y ultrasonidos para intercambiar códigos de emparejamiento entre dispositivos.

Al utilizar ultrasonidos, los usuarios no pueden identificar a los dispositivos que participan en el emparejamiento. De hecho, recientemente, investigadores de la Universidad de Brunswick han descubierto aplicaciones Android en Google Play Store que hacen uso de técnicas de ultrasonidos en anuncios, para poder relacionar cuales son los dispositivos del usuario con objeto de realizar campañas de marketing dirigido sin el conocimiento del usuario, es lo que se conoce como *"ultrasound Cross-Device Tracking (uXDT)"* [8]. Todo esto nos lleva a buscar tecnologías audibles de forma que los usuarios sean conscientes de los intercambios realizados.

Para el desarrollo del prototipo se recurrió a "Chirp" (ver sección IV) que tiene como finalidad el intercambio de contenido multimedia entre dos o más dispositivos cercanos mediante sonidos similares a los módem analógicos. En realidad lo que se envía a través del sonido es un enlace al archivo, que ha sido alojado en su servidor desde el dispositivo, al recibirlo el otro usuario lo descarga en su aplicación y lo visualiza casi de forma instantánea. Chirp incluye un SDK tanto en Android como en iOS así como en plataformas web. Chirp permite el uso de la SDK durante un periodo de tiempo limitado de forma gratuita a los desarrolladores que lo soliciten. Este fue uno de los puntos que decantó la elección de esta tecnología.

III. SOLUCIONES PROPUESTAS

En el desarrollo de nuestro prototipo hemos asumido que el atacante no tiene acceso de forma simultanea a los dos canales de comunicación que utilizamos: el canal sonoro implementado usando la SDK de Chirp y un canal a través de Internet implementado usando Firebase (ver sección IV). Con respecto al canal de comunicaciones a través de Internet asumimos que todos los usuarios tienen un ID temporal diferente único, por lo que el atacante no puede apoderarse de la sesión de un usuario existe, aunque si puede leer los mensajes de todos los usuarios y enviar mensajes con su propio ID a cualquier usuario del sistema. También puede conseguir nuevos IDs simplemente estableciendo una nueva conexión. Con respecto al canal sonoro o canal fuera de banda, asumimos que el atacante no puede enviar ningún código sonoro, ya que en ese caso, el usuario receptor podría identificar que la fuente del sonido no se corresponde con el dispositivo del otro usuario y abortar el emparejamiento. Estamos por tanto ante un atacante pasivo en el canal fuera de banda, que si podría enviar tráfico en el canal de comunicaciones pero no puede secuestrar sesiones existentes.

Durante el desarrollo del prototipo se han tenido en cuenta diferentes diseños para el protocolo, en función de las capacidades del atacante. También se ha tenido en cuenta una limitación de Chirp en cuanto al envío de información, que no puede superar los 50 bits[4].

[1] http://newatlas.com/sonicnotify-audio-signals/21385/
[2] https://applidium.com/en/news/data_transfer_through_sound/
[3] https://developers.google.com/nearby/
[4] http://developers.chirp.io/docs/online-and-offline-operation

La primera aproximación (Figura 1) consiste en el envío directo de la clave mediante un código sonoro. Esta opción es la más simple de implementar pero presenta el problema de que un usuario que se encuentre en la misma zona de influencia puede tener acceso a la clave e interceptar todas las comunicaciones, por tanto solo sería segura antes atacantes sin acceso al canal fuera de banda.

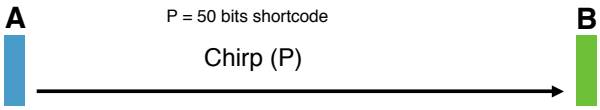


Fig. 1. Primera iteración del protocolo

También se podría impersonar al segundo usuario frente al primero en el canal de comunicaciones al disponer de la clave de comunicaciones. Por tanto solo es recomendable en escenarios donde solo los dos dispositivos tengan acceso al canal sonoro. Por ejemplo, en entornos con mucho ruido de fondo, ajustando el nivel de volumen, la distorsión del ruido haría que un dispositivo que estuviera fuera del rango visual de los usuarios no fuera capaz de recibir correctamente la clave P. Sin embargo, en entornos sin ruido, y con un volumen suficientemente alto, el atacante podría ser capaz de capturar la clave sin ser visto. Otro problema es que la clave al tener solo 50 bits como máximo tiene una entropía limitada.

La primera idea que podríamos plantear para mejorar la situación sería usar un protocolo DH donde cada usuario enviara su clave pública usando el canal fuera de banda, desgraciadamente el limite de dicho canal hace que esta aproximación sea insegura. Por tanto, se definió una segunda aproximación (Figura 2) que intenta resolver esos problemas, implementando el protocolo Diffie Hellman dentro del canal de comunicaciones con una verificación de la clave intercambiada usando el canal fuera de banda.

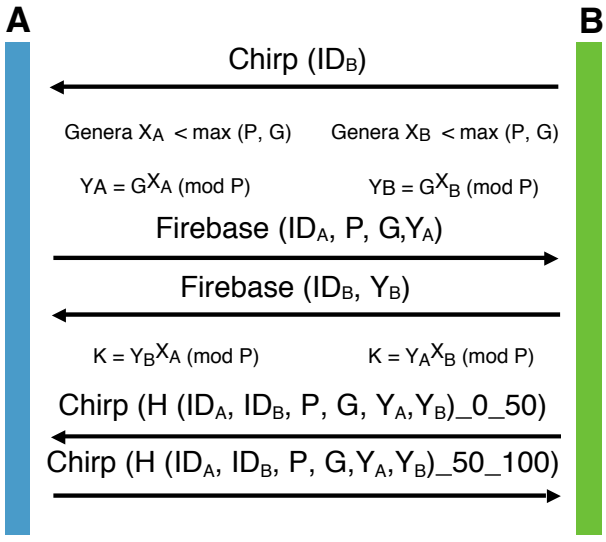


Fig. 2. Segunda iteración del protocolo

En concreto, se realiza la autenticación mediante códigos sonoros de las identidades de los usuarios en canal de comunicaciones, así como de las claves públicas y parámetros del protocolo Diffie Hellman. Para ello se comparte el hash de todos los elementos mencionados anteriormente por el canal fuera de banda, aunque debido a la limitación de 50 bits, se divide el código de verificación en dos partes, teniendo que enviar cada usuario una de las dos partes para que el otro la verifique.

En este protocolo un atacante no podría impersonar a ninguno de los usuarios legítimos, ya que aunque intercepte el ID del primer usuario no podría utilizarlo dentro del canal de comunicaciones con lo que sus mensajes sería descartados por el segundo usuario. Tampoco podría impersonar al segundo usuario porque no sería capaz de enviar el código de autenticación de vuelta usando el canal fuera de banda.

Aunque este esquema es más seguro que el anterior, el principal problema que presenta es que se necesita enviar al menos tres mensajes con códigos sonoros, uno para iniciar la comunicación con el ID y dos más para la autenticación con el HMAC, lo cual introduce problemas de sincronía entre ambos canales y además resulta poco práctico debido al tiempo que tardan en enviarse los mensajes por el canal fuera de banda. Se podría reducir el envío de mensajes a dos, solo para la verificación de la clave, pero en ese caso habría que acompañar a las claves públicas de ambos usuarios de alguna información que le permitiera relacionar ambos mensajes. Una forma de calcular un identificador único compartido entre ambos sería usar una función HASH sobre los dos parámetros siguientes:

1) **Hora actual:** En horas y minutos, de esta forma sabemos que la sincronización se está llevando a cabo en un instante de tiempo determinado. Esto también permitiría evitar *"Replay Attacks"* [9].
2) **Redes WiFi disponibles:** Podemos comprobar que ambos dispositivos se encuentran próximos si detectan las mismas redes WiFi. Para ello utilizamos el SSID de las redes con mayor señal. De esta forma podemos también mitigar los conocidos como *"Wormhole Attacks"* [10] en redes ad-hoc.

Enviado las claves públicas en difusión, acompañadas de este identificador, por el canal de comunicaciones, cada usuario podría identificar la clave pública de la otra parte y completar el intercambio de claves.

Si queremos implementar un prototipo con un solo mensaje por el canal fuera de banda, siempre será posible que el receptor de ese mensaje sea impersonado por un atacante que se encuentre próximo a ambos usuarios. Es decir, el iniciador nunca podrá tener la certeza que el dispositivo que está viendo es el que responde a través del canal de comunicaciones. Lo más que se puede hacer en ese escenario es una validación mutua a posteriori de la identidad de ambos usuarios, usando otro canal fuera de banda como puede ser el visual. Una opción es enviar información personal de cada usuario, como el nombre o alguna foto, que permita a cada usuario reconocerla en el dispositivo de la otra parte. Aún así, un atacante que tenga acceso, aunque solo sea de escucha, a ambos canales de

forma simultanea sería capaz de leer todos los mensajes.

Bajo al asunción de que el atacante no puede observar el canal de fuera de banda, se puede mantener la misma usabilidad de la primera versión, pero usando una clave de mayor calidad. Para ello se definió una tercera aproximación (Figura 3) que hace uso de la función PBKDF2 [11] (Password Based Key Derivation Function) para a partir de la clave de 50 bits enviada por el canal fuera de banda generar una clave de mayor extensión, usando como *"salt"* la hora y las redes wifi disponibles.

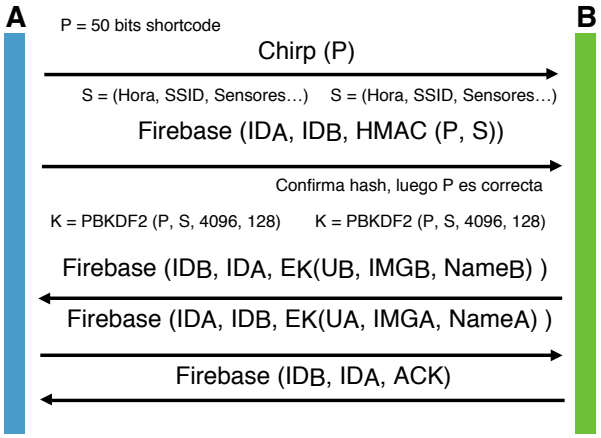


Fig. 3. Tercera iteración del protocolo

Una vez calculada la clave por ambos dispositivos se procede a enviar la información privada del usuario (nombre e imagen) cifradas, así como un código HMAC que servirá para autenticar al receptor y confirmar que la clave se ha calculado correctamente. Si el código HMAC es correcto se procede a añadir la información transmitida como un nuevo contacto y establecer una ID para el canal de comunicación con el mismo. En caso contrario la información se descarta y no se responde con la información propia al mensaje recibido.

La información de los contactos añadidos se almacena de forma local en la aplicación y no es accesible por el servidor, por otra parte las IDs de los usuarios se generan aleatoriamente en cada comunicación, por lo que se almacenan las ID utilizadas por el emisor y el receptor en la sincronización y se emplean (en un orden determinado) para establecer la ID del canal de comunicación entre ellos.

El objetivo de esta sincronización es que únicamente se envíen al servidor dos tipos de información: IDs aleatorias de comunicación y mensajes cifrados. De esta forma, aunque alguien pudiera atacar y observar la información del servidor no podría conocer el contenido de los mensajes enviados, ni siquiera podría identificar usuarios concretos, ya que en cada canal de comunicación utilizan ID distintos, garantizamos así no solo la confidencialidad de los mensajes sino también la privacidad de los usuarios.

En la Tabla I se puede ver un resumen comparativo de los tres esquemas que se han mencionado.

Tabla I
COMPARATIVA DE LOS DISTINTOS PROTOCOLOS PLANTEADOS

Protocolo	Bits de clave	Mensajes	Resiste ambos canales
$Claro$	50	1	No
DH	> 50	3	Si
$PBKDF2$	$50 + salt$	1	No

IV. TECNOLOGÍAS UTILIZADAS

La plataforma sobre la que se desarrolla el prototipo es iOS, esto se debe principalmente a que la primera tecnología que tratamos de utilizar fue Audio Modem que trabaja sobre iOS, aunque posteriormente se optó por Chirp pensando en la portabilidad de la aplicación a otros sistemas operativos.

A. Chirp

Tal y como se ha adelantado, Chirp[5] es una tecnología que permite la comunicación de información entre dispositivos utilizando el canal sonoro. Para ello emplean una asociación unívoca entre un byte y una determinada frecuencia, de forma que el carácter 'a' podría, por ejemplo, tener asociada la frecuencia de 500Hz, el carácter 'b' la de 510Hz sucesivamente.

Distinguen dos tipos de mensajes en su SDK:

1) **Shortcodes:** Son mensajes de un máximo de 50 bits, enviados en forma de 10 caracteres que no pasan por el servidor de Chirp y que se envían a través del canal sonoro de un dispositivo a otro siguiendo la codificación en frecuencias anteriormente explicadas.
2) **Mensajes de diccionario:** Son estructuras de datos más complejas que se envían con una clave, esta clave es, precisamente, un *"shortcode"* de los anteriormente mencionados. Estos mensajes con diccionario son, en última instancia, una clave de 10 caracteres que identifica un mensaje en formato JSON con la información estructurada.

Los mensajes con diccionario, a diferencia de los *"shortcode"*, sí necesitan subirse al servidor de Chirp. En su funcionamiento, el usuario final recibe mediante el canal sonoro un *"shortcode"* que utiliza como *"token"* frente al servidor de Chirp para obtener la estructura de datos en formato JSON.

Además, la SDK de Chirp también permite entre otras cosas la visualización en tiempo real de la onda sonora captada por el micrófono.

B. Firebase

Firebase[6] es una tecnología relativamente nueva, que comenzó como un proyecto independiente y ha sido comprada por Google. Aunque ahora ha integrado muchos servicios con Google podemos definir Firebase en su origen como una "Base de datos NoSQL online basada en JSON". En la actualidad incorpora gran cantidad de servicios adicionales en colaboración con Google como

[5]Chirp: http://chirp.io
[6]Firebase: https://firebase.google.com/

Analytics, Autenticación, Almacenamiento de ficheros, Hosting, Monetización de aplicaciones mediante Google Admob.

Dentro de Firebase la información se estructura en subramas en forma de arbol, con un nodo inicial: nuestro identificador de Firebase. Esto permite definir fácilmente ciertas políticas de control de acceso, por ejemplo, dando acceso a un elemento se concede acceso a todos los elementos de la rama.

C. CyptoSwift

CryptoSwift[7], tal y como su nombre sugiere, es una librería criptográfica en el lenguaje de programación Swift (utilizado conjuntamente con Objective-C a lo largo del desarrollo del proyecto iOS).

Se trata de un proyecto de código abierto en Github que incluye las operaciones básicas de criptografía que se utilizan actualmente, entre ellas: Funciones hash (MD5, SHA1.. 512), CRC, Algoritmos de cifrado (AES128..256, ChaCha20, Rabbit), Códigos de autenticación HMAC (MD5, SHA1..256, Poly1305), Modos de operación en bloque (ECB, CBC, CTR...), Funciones de derivación de claves (PBKDF 1 y 2) y Esquemas de relleno (PKCS5/PKCS7).

Para las aislar las operaciones criptográficas del control de la aplicación hemos creado un interfaz en una clase llamada *"CustomCrypto"* que contiene los métodos necesarios para realizar todas las operaciones criptográficas con CryptoSwift ofreciendo el resultado esperado al control de la aplicación.

D. JSQMessages

Por último, JSQMessages[8] es otro proyecto *"open source"* que conforma una librería de gráficos y control para crear de forma sencilla los elementos básicos y con un diseño estandarizado de una ventana de chat en iOS. Entre los elementos se incluyen diferentes burbujas con mensajes: texto, contenedoras de vídeos o imágenes, localización geográfica. Todo a nivel de interfaz de usuario.

Dado que se trata de una librería compleja y la elaboración de un chat es sólo una excusa para poner de manifiesto la solución teórica propuesta hemos hecho un uso reducido de esta librería limitándonos a la creación de ventanas de chat, burbujas de mensajes e indicadores de "escribiendo...".

V. NUESTRO PROTOTIPO: CHATCHAT

Como aplicación prototipo para la utilización del canal sonoro con Chirp e implementación del protocolo propuesto hemos elaborado una aplicación de chat para dispositivos iOS, que hemos denominado ChatChat.

En la pantalla principal (Figura 4) podemos apreciar tres botones:

1) **General:** Chat general compartido por todos los usuarios de ChatChat. No tiene cifrado.
2) **Profile (Perfil):** Permite editar nuestro perfil, nombre de usuario e imagen.
3) **Contacts (Contactos):** Permite establecer un chat con un usuario sincronizado o sincronizar la información de dos usuarios, poniendo en práctica nuestro protocolo

La ventana de chat general (Figura 5) muestra en burbujas los mensajes enviados al "Chat General" de Firebase. En este chat general pueden enviar mensajes todos los usuarios, usando un nuevo ID cada vez que entremos. Estos mensajes "anónimos" se almacenarán en Firebase en claro y serán visibles para el resto de usuarios de la sala. Nótese que si salimos y volvemos a entrar en el chat general Firebase nos asigna un nuevo ID aleatorio, por lo que no hay forma de identificar los mensajes pertenecientes a un mismo dispositivo o usuario.

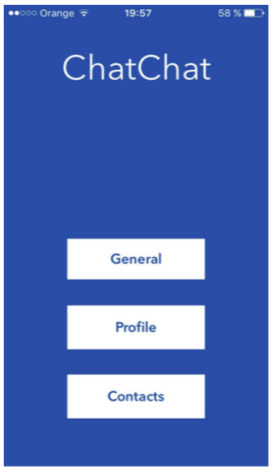


Fig. 4. Menú principal

Fig. 5. Pantalla de chat

En la ventana de perfil (Figura 6) podemos editar la información de usuario: nuestro nombre y nuestra imagen. Esta información se almacena en un fichero local y es persistente aunque cerremos la aplicación. Ésta es también la información que se envía al otro dispositivo después de un intercambio correcto y es la que deben usar los usuarios para verificar que no hay un impostor.

Si en el menú principal pulsamos sobre el botón *"Contacts"* iremos a la ventana de contactos existentes y sincronización de los mismos (Figura 7). En la parte central encontramos un botón *"Add New"* que al pulsarlo pondrá en marcha nuestro protocolo de sincronización, generando el *"shortkey"* de chirp para reproducirlo a través del altavoz del dispositivo, calculando la clave con PBKDF (fecha + WiFi) y enviando los mensajes correspondientes a través de Firebase. En esta ventana y sin haber pulsado el botón nos encontraríamos en la situación de receptor de la sincronización (inicialmente ambos), estando suscrito a una rama de mensajes de Firebase que se utiliza en exclusiva para gestionar las sincronizaciones. En la parte inferior de la pantalla, de un color azul más claro, hemos incluido la visualización en tiempo real de la captación de sonido del micrófono.

Después de un intercambio exitoso en ambos dispos-

[7]CryptoSwift: https://github.com/krzyzanowskim/CryptoSwift
[8]https://github.com/jessesquires/JSQMessagesViewController/tree/master

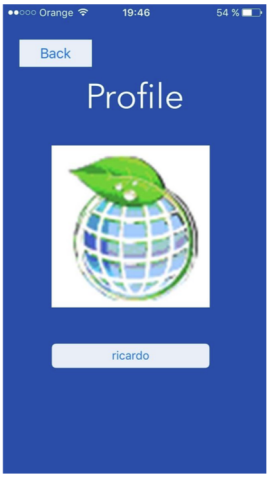


Fig. 6. Pantalla de perfil.

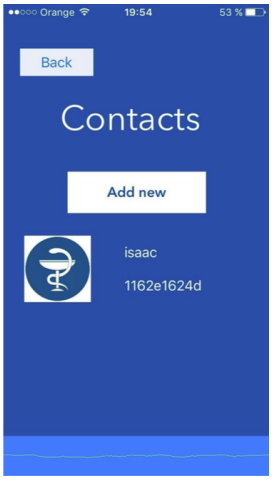


Fig. 7. Pantalla de contactos.

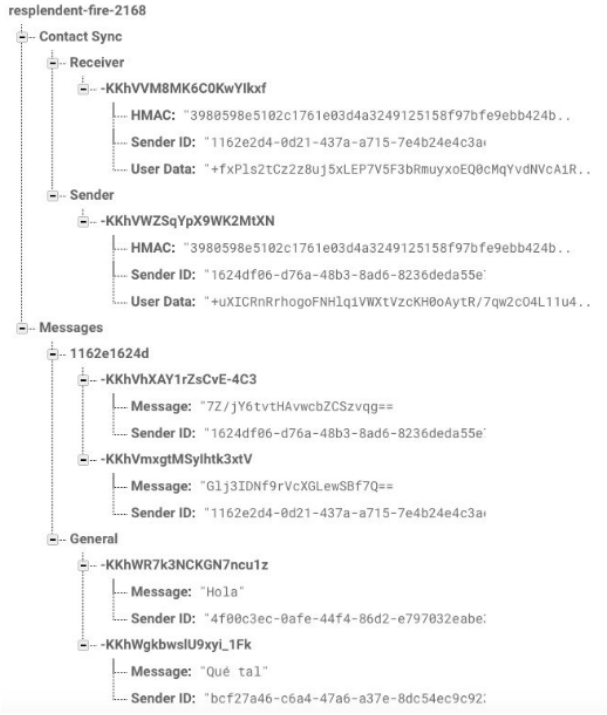


Fig. 8. Esquema de chat de Firebase

itivos se agregará el dispositivo opuesto como contacto. El número inferior al nombre de usuario es la ID del canal (construida a partir de la ID de ambos contactos en la sincronización) que será la rama en Firebase en la que ambos realizarán la comunicación. Si tocamos la imagen del contacto agregado podemos acceder a tener una conversación con él en una ventana de chat similar a la del canal general.

Distinguimos en nuestro esquema de Firebase (Figura 8) dos grandes ramas:

1) **Contact Sync:** Es la rama en la que se envían los mensajes de sincronización. Tiene dos subramas: Receiver y Sender, en función de quién haya iniciado la sincronización con Chirp y quien haya recibido el mensaje sonoro. Se envía la información de usuario cifrada, el HMAC y el ID aleatorio
2) **Messages:** Es la rama dedicada al envío de mensajes, distinguimos una rama General en la que podemos observar que se envían los mensajes en claro (texto ”Hola” y ”Que tal”) y una rama con un ID de canal coincidente con el ID que aparecía debajo del nombre del contacto agregado constituido por el ID aleatorio que ambos tenían. En esta subrama (canal de contactos sincronizados) podemos observar que los mensajes se envían efectivamente cifrados

De esta forma, por el servidor de Firebase únicamente pasan dos tipos de información: IDs aleatorios cada vez que establecemos un canal de comunicación (que garantizan la privacidad de los usuarios al no poder identificarse con los mensajes enviados) y mensajes cifrados. También se almacenan mensajes en claro, pero únicamente en el canal general.

VI. CONCLUSIONES Y TRABAJO FUTURO

Podemos concluir sobre este proyecto que los resultados que pretendíamos alcanzar son viables con la tecnología actual y se ha conseguido implementar un prototipo funcional de código abierto que está disponible en la plataforma GitHub [9].

Por desgracia, la usabilidad está reñida con el nivel de seguridad que se deseé obtener por lo que uno de los puntos a mejorar sería la tecnología para el envío de información utilizando el canal sonoro, que aún está en vías de desarrollo y, como hemos visto, en Chirp tiene una extensión máxima de 50 bits.

En cuanto a las posibles mejoras sobre el prototipo elaborado, una de las deficiencias de nuestra propuesta es que no proporcionan ”Perfect Forward Secrecy (PFS)” de forma directa. Es decir, si un atacante consiguiera hacerse con la clave privada de un canal, mediante el robo de uno de los dispositivos, podría descifrar todos los mensajes pasados de ese canal. Una forma fácil de solucionar este problema, sería usar la clave negociada en la sincronización de los dispositivos solamente como mecanismo de autenticación y no para establecer un canal confidencial. De esta forma, se requeriría una fase extra que podría consistir en un protocolo Diffie-Hellman donde al final, ambos usuarios utilizan la clave secreta negociada usando el canal sonoro para confirmar que no ha habido intermediarios en el protocolo Diffie-Hellman.

Otra línea de trabajo futuro sería integrar este desarrollo con alguna aplicación de mensajería instantánea, como por ejemplo Telegram.

REFERENCIAS

[1] Boyd C., Mathuria A: ”Key establishment protocols for secure mobile communications: A selective survey”. In: Boyd C., Dawson E. (eds) Information Security and Privacy. ACISP 1998. Lecture Notes in Computer Science, vol 1438. Springer, Berlin, Heidelberg

[9] ChatChat: https://github.com/RicardoRuizTueros/ChatChat

[2] Whitfielf Diffie, Martin E. Hellman, , "New Directions in Cryptography", 644 IEEE Transactions on Information Theory, Vol. IT-22, No. 6, 1976
[3] N. Unger et al., "SoK: Secure Messaging," 2015 IEEE Symposium on Security and Privacy, San Jose, CA, 2015, pp. 232-249. doi: 10.1109/SP.2015.22
[4] Roland, Michael, and Josef Langer. "Cloning Credit Cards: A Combined Pre-play and Downgrade Attack on EMV Contactless." WOOT. 2013.
[5] Sebastian U. et al. "Tactile One-Time Pad: Leakage-Resilient Authentication for Smartphones", Ruhr-University Bochum, Germany, 2015
[6] Nikolaos K et al. "Sound-Proof: Usable Two-Factor Authentication Based on Ambient Sound", 24th USENIX Security Symposium, Washintong DC, 2015
[7] F. Gardner "A BPSK/QPSK Timing-Error Detector for Sampled Receivers", IEEE Transactions on Communications. Volume: 34, Issue: 5, 1986
[8] Daniel A. et al. "Privacy Threats through Ultrasonic Side Channels on Mobile Devices", Technische Universität Braunschweig, Brunswick, Germany. 2017
[9] Han G. et al. "A Formal Analysis for Capturing Replay Attacks in Cryptographic Protocols", Informatics and Mathematical Modelling, Technical University of Denmark and Dipartimento di Informatica, Universita di Pisa
[10] Mariano G, Adrián P. "Detection of wormhole attacks in wireless sensor networks using range-free localization", IEEE 17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD). 2012
[11] K. Moriarty et al. "PKCS 5: Password-Based Cryptography Specification Version 2.1", RFC 8018, Internet Engineering Task Force (IETF), 2017

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6558

Sistemas de gestión de contenido web: Uso y estudio comparativo inicial de su seguridad

Antonio-José Aledo-Hernández[1], Antonio Guillen-Pérez[1], Jose-Manuel Martinez-Caro[1], Ramón Sánchez-Iborra[2], María-Dolores Cano[1]

[1]Dpto. Tecnologías de la Información y las Comunicaciones

[1]Universidad Politécnica de Cartagena

[2]Dpto. Ingeniería de la *Información* y las *Comunicaciones*

[2]Universidad de Murcia

antonio.aledo@alu.upct.es, agp4@alu.upct.es, jmmc0@alu.upct.es, ramonsanchez@um.es, mdolores.cano@upct.es

***Resumen*- Los Sistemas de Gestión de Contenido Web (*Web Content Management Systems,* WCMS) han ganado mucha popularidad debido a la facilidad que aportan a la hora de crear páginas o portales web, *sites* de comercio electrónico, etc. En este trabajo se explica de forma resumida cómo es el manejo los WCMS y qué se puede lograr con su uso. Para ello, trabajaremos con tres de los más populares WCMS de tipo *open-source* empleados hoy en día, Joomla, Wordpress y Drupal, y veremos las ventajas e inconvenientes de trabajar con cada uno de ellos. Con este fin, crearemos tres web iguales en requisitos y funcionalidades, una con cada WCMS, y se analizará cualitativamente la complejidad de cada uno de ellos. Finalmente, realizaremos un análisis básico de seguridad de las webs creadas, informando de sus posibles vulnerabilidades, explicando cómo mejorar su seguridad, qué fallos no debemos cometer y qué WCMS es inicialmente más seguro/vulnerable.**

***Palabras Clave*- sistemas de gestión de contenidos web, seguridad, Joomla, Wordpress, Drupal**

I. Introducción

Los Sistemas de Gestión de Contenido Web (*Web Content Management Systems*, WCMS) se utilizan generalmente en entornos donde es necesario crear un portal online de contenidos web sin necesidad de disponer de cocimientos sobre programación web y donde además se permite la creación de gran variedad de roles de usuario [1-3]. Un ejemplo de lo citado sería la redacción de un periódico, donde están interesados en lanzar su edición electrónica online. En este caso los editores sólo tienen conocimientos informáticos de ofimática, no de programación web para crear su portal, y es en este punto donde harían uso de los gestores de contenido web para poder llevar a cabo tal fin.

Los Open Source WCMS, también denominados WCMS de segunda generación, son plataformas de código abierto (*open-source*) desarrolladas a menudo en PHP y alimentadas con sus propias comunidades de usuarios, que aportan tanto soluciones como funcionalidades a estos [4]. La estructura básica de un WCMS la componen las siguientes partes: i) los archivos propios del gestor de contenidos, ii) un proveedor de hospedaje para almacenar en él los archivos propios del WCMS, y iii) una base de datos (p.e.: MySQL) para vincularla y almacenar en ella la información de nuestro futuro *site.* Los WCMS ofrecen una zona de administración o desarrollo, donde podremos añadir artículos, funcionalidades o dotar de un aspecto concreto conocida como *back-end.* Por otro lado la parte visible para el visitante del *site* es conocida como *front-end.*

En este artículo, ampliación del trabajo realizado en [5], se explica cómo es el manejo los WCMS y qué se puede lograr con su uso. Para ello, trabajaremos con tres WCMS populares y con buenas prestaciones [6]: Joomla [7], Wordpress [8] y Drupal [9]. Aunque Joomla y Drupal han ido perdiendo protagonismo en los últimos años frente a otras propuestas su todavía amplio uso nos ha hecho optar por incluirlos en esta comparativa [10]. Veremos también las ventajas e

inconvenientes de trabajar con cada uno de ellos en la actualidad. Con este fin, crearemos tres web iguales, una con cada WCMS, y se analizará la complejidad de cada uno de ellos. Finalmente, realizaremos un análisis preliminar de seguridad de los *sites* creados, informando de sus posibles vulnerabilidades, explicando cómo dotarlos de seguridad, qué fallos no debemos cometer y qué WCMS es más seguro/vulnerable.

El resto del artículo se distribuye de la siguiente forma. La sección II describe brevemente los trabajos relacionados en la temática. En la sección II se detallan los WCMS con los que trabajaremos. La sección IV proporciona una guía y una comparación de cómo crear un *site* utilizando los WCMS bajo estudio. En la sección V se presenta un análisis de seguridad básico. El documento finaliza con las conclusiones más relevantes de este trabajo.

II. Trabajos Relacionados

En la literatura especializada podemos encontrar algunos trabajos que también abarcan este tema. En [11] los autores presentan un estudio comparativo de siete WCMS, incluyendo los tres empleados en este trabajo. No obstante, la comparativa se basó únicamente en la implementación de un *site* con cada WCMS y en determinar si alguna de las funcionalidades establecidas como requisitos previos estaban disponibles o no. Por otro lado, [12-13] son trabajos muy interesantes que introducen y analizan la seguridad en en los WCMS. En [12], se identifican las vulnerabilidades y ataques a los que están expuestos los WCMS, así como posibles medidas de respuesta. Como caso práctico, experimentaron con Joomla y Drupal con versiones previas a las empleadas en este trabajo y realizaron algunos test de penetración sencillos usando las herramientas WebScarab [14] y TamperData [15]. Como principal resultado los autores indican que a pesar de disponer de mecanismos de seguridad, ambos WCMS pueden ser víctimas fáciles de ataques. En [13], los autores presentan un estudio en profundidad de los sistemas de gestión de contenidos, aunque desde una perspectiva puramente cualitativa, incluyendo otros aspectos como el efecto de los servidores web (i.e, Apache, Nginx, etc.) en base a estadísticas disponibles. Como veremos en las siguientes secciones, este trabajo actualiza el llevado a cabo en [12] y contribuye de forma práctica a los resultados obtenidos en [13].

III. Sistemas de Gestión de Contenido Web bajo estudio

Joomla! [7] es uno de los WCMS más populares para crear portales web dinámicos [10]. En la fecha de preparación de este trabajo, tenía su compatibilidad limitada únicamente a bases de datos de tipo MySQL. Su característica fundamental es que ofrece la gama más alta de funcionalidades, como pueden ser galerías de imágenes, foros, chats, blogs, deslizadores de imágenes, noticias y un largo etc. Por su parte, Drupal [9] también está destinado a la creación de portales dinámicos. Nos permite una compatibilidad con gran diversidad de tipos de bases de datos. Su característica principal es su seguridad, rapidez de carga y la diversidad en los roles de usuario, ya que nos permite por ejemplo limitar el acceso de cierto usuario hasta el punto de sólo poder éste modificar las propiedades de una determinada funcionalidad e incluso sólo de ciertos parámetros de esta funcionalidad. Wordpress [8] es un WCMS destinado sobre todo a la creación de *blogs*, y ha evolucionado hasta proporcionar soluciones de aplicaciones web y comercio electrónico. Tal y como ocurre con Joomla! su compatibilidad está limitada (en la fecha de realización de este trabajo) a la base de datos MySQL. Su característica principal es su gran posicionamiento SEO (*Search Engine Optimization*), ya que una web Wordpress dispone de numerosos *plugins* para facilitar una rápida aparición en los motores de búsqueda en comparación con otros WCMS [16]. Además nos permite crear un *blog* sencillo de forma gratuita sin necesidad de tener contratado un hospedaje, bajo la extensión de un subdominio propio (*.wordpress.com*). De forma detallada las características fundamentales de los tres WCMS a estudio para creación de portales web se muestra en la Tabla I obtenida a partir de [4-11, 16].

Cabe destacar algunos aspectos de la Tabla I como pueden ser que Wordpress ofrece muchas extensiones para realizar una funcionalidad concreta pero no tiene un amplio abanico de funcionalidades como puede ser el caso de Joomla!. También decir que Wordpress ofrece la posibilidad de añadir extensiones integradas en su entorno de administración, y que el más complejo de los tres WCMS para llevar a cabo su extensibilidad es Drupal, ya que una funcionalidad concreta puede tener dependencias de librerías u otras funcionalidades relacionadas, lo que hace que este proceso de extensibilidad pueda ser en ocasiones largo y tedioso.

Tabla I
Algunas características fundamentales de Joomla!, Wordpress y Drupal

Característica	Joomla!	Drupal	Wordpress
Tipo principal de contenido	Web *sites*, online apps	Blog	Blog, e-commerce, online apps
Disponibilidad de extensiones	Alta	Media	Alta
Variedad de funcionalidades	Alta	Media	Alta
Repositorio de extensiones	Distribuido	Centralizado	Distribuido
Documentación	Muy buena	Buena	Muy buena
Comunidad de usuarios	Muy activa	Limitada	Muy activa
Facilidad de uso	Sencilla	Compleja	Sencilla
Personalización de roles de usuario	Media	Muy alta	Media
Posicionamiento SEO manual	Sí	Sí	Sí
Posicionamiento SEO automático	Extensiones	Modules	Plugins and tools

IV. Creación de un portal web con Joomla!, Drupal y Wordpress

A la hora de comparar los tres WCMS, se va a crear un portal web con cada uno de ellos. Este portal deberá tener las siguientes funcionalidades (algunas de ellas se muestran en la Fig. 1) y además se desea que las funcionalidades estén dispuestas siguiendo la distribución de la Fig. 2.:

- Deslizador de imágenes: *Slider* o *banner* de imágenes en movimiento para la página principal del *site*, basado en Javascript
- Módulo de acceso o login: Permitirá crear zonas privadas en el *site* y el registro de usuarios al mismo
- Integración redes sociales: Twitter y Facebook
- Módulo multilenguaje: Traducción basada en el traductor de Google para ofrecer la posibilidad de traducir los contenidos
- Módulo buscar: Para búsqueda de contenido indexado por parte del visitante del *site*
- Formulario de contacto
- Vídeos
- Mapa
- Descargas: Descargas multiusuario personalizadas
- Boletín de noticias y eventos: Dotar de la posibilidad al usuario de estar informado de novedades acerca del *site* por e-email y por otro lado situar noticias importantes en la página principal de nuestra página

Fig. 1. Algunos elementos que deseamos tener en nuestras 3 versiones

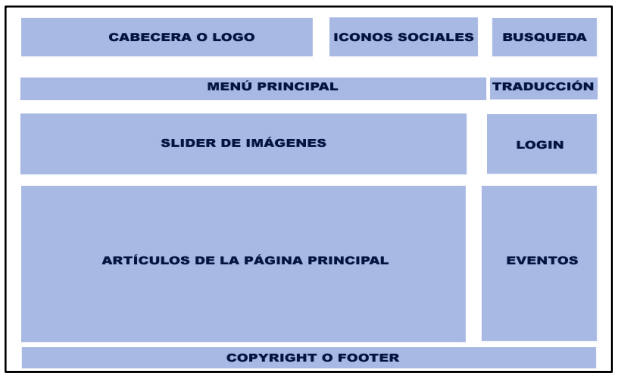


Fig. 2. Disposición de las funcionalidades

Para comenzar, veremos el proceso de instalación de un WCMS. Independientemente de con cuál estemos trabajando, se suelen seguir siempre las siguientes pautas:

1. Contratar un proveedor de hospedaje (por ejemplo 1and1.es) que incluya un sistema de base de datos (por ejemplo MySQL)
2. Crear una base de datos para el CMS
3. Descargar el paquete de instalación del CMS desde la web oficial y descomprimir en el directorio virtual proporcionado por el proveedor de hospedaje
4. Instalar el CMS usando el asistente que proporciona y que enlaza con la base de datos

Por otra parte, el proceso de configuración y personalización de un WCMS suele seguir el siguiente patrón:

1. Descargar (o diseñar con el software creador de plantillas conocido como Artisteer [17]) la plantilla base del sitio
2. Asignar la plantilla al sitio web a partir del gestor de plantillas
3. Buscar las extensiones (funcionalidades) necesarias
4. Activar extensiones dentro del sitio web y situarlas en las posiciones deseadas (ver bloques o posiciones disponibles de plantilla previamente en la configuración de esta)
5. Generar contenidos por parte de los editores, para ello los WCMS incorporan su propio editor para incluir imágenes, videos, personalizar textos, etc.

De forma resumida, la Tabla II incluye las diferencias y similitudes detectadas a la hora de implementar el *site* con Joomla, Drupal y Wordpress. Obsérvese que las plantillas seleccionadas disponen de un *framework* propio que les permite cambiar la disposición de los bloques y su tamaño, el ancho de la hoja, las fuentes, colores, etc. Algo que nos permite dar el aspecto que deseemos al *site* además de asemejarse en las tres versiones a realizar. Referente a Artisteer [17] decir que es un software muy intuitivo en su utilización que nos permite plasmar él aspecto, disposición y estilos que deseemos para la plantilla de un *site* y exportarlo en un archivo .zip listo para ser instalado en nuestro gestor de contenido web. En cuanto a la cabecera y pie de página, el módulo Artical para Joomla! convierte un artículo creado con el editor en un módulo para situarlo en la posición que deseemos del *site*. Respecto al *slider* de Drupal, comentar que éste necesita una librería adiccional llamada *jquery.cycle.all.js* para funcionar correctamente y que en el caso de Wordpress para mostrar el *slider* tenemos que incluir dentro de una página o entrada el código que nos devuelve este en su configuración. En la configuración de sendos *sliders* cabe destacar la elección de las dimensiones para estos y la elección de

Tabla II
SIMILITUDES Y DIFERENCIAS DETECTADAS A LA HORA DE IMPLEMENTAR LAS FUNCIONALIDADES DEL *SITE*

Funcionalidad	Joomla!	Drupal	Wordpress
Diseñador de plantillas propio	Artisteer		
Plantilla usada	Yoo Downtown	AT commerce	Yoo Downtown
Cabecera y pie de página	Módulo Artical	Incluido en plantilla	Incluido en plantilla
Slider de imágenes	Nivo	Incluido en plantilla	
Redes sociales	ITPsocialbuttons	Web de Linksalpha nos proporciona el código social	
Traducción	GTranslate		
Eventos	JNews	Bloque de contenido reciente	Widget enlaces permanentes
Gestor de descargas	Jdownloads	CMS Mollify	
Boletín de noticias	Web de Mailchimp		
Formulario de contacto	CKForms	Ya incluido en núcleo	CformsII
Creador de módulos	Jumi	-	-
Editor de artículos	JCE	Propio	Propio
Youtube, Twitter y Mapa	Mediante inserción del código que nos devuelve la webs oficiales de cada una en páginas nuevas o artículo		
Busqueda y login	En el núcleo		

la ruta donde se encuentran las imágenes que queremos que estos contengan. En cuanto a los botones sociales, hay páginas web que nos devuelven el código necesario para insertarlo en un nuevo bloque en nuestra web. Cuando un visitante del *site* haga clic sobre uno de estos botones enlazará con su propia cuenta de la red social asociada al botón para que este visitante pueda publicar como recomendado el enlace de nuestro *site*. Respecto a la disponibilidad de un visitante de cambiar el idioma de todo el contenido del *site* se puede decir que Google se encarga de ello disponiendo de un módulo donde prácticamente sólo tenemos que indicar en su configuración qué idiomas queremos ofrecer (con banderas de elección para el visitante) y el idioma por defecto del sitio.

Para llevar a cabo la gestión de eventos en la versión de Drupal se ha hecho uso del bloque de contenido reciente para indicar qué páginas creadas situaremos como eventos. Para la versión Joomla!, se ha configurado el módulo usado para este fin indicándole que identificadores de artículos queremos resaltar como los más importantes, y por último en la de Wordpress se ha empleado el *widget* de enlaces permanentes para pegar en él los enlaces de las páginas que nos interesen para eventos. Importante reseñar que cuando creamos una página Wordpress ésta nos devuelve un enlace permanente.

La funcionalidad de gestión de descargas multiusuario no estaba disponible en Drupal y Wordpress (en la fecha de desarrollo de este trabajo), por lo que se hizo uso del gestor de archivos Mollify. Para el boletín de noticias se utilizó la web oficial de Mailchimp [18] en las tres versiones. Esta web devuelve un código para insertar en nuestra web para que los usuarios puedan subscribirse y estar por tanto informados vía email de las novedades que deseamos ofrecer a nuestros visitantes.

Una de las funcionalidades que más configuración y personalización requiere es la del formulario de contacto ya que se deben crear las etiquetas, cajas de texto simples, áreas de texto, botones de envío y reseteo que deseemos que éste contenga y además indicar en su configuración parámetros como pueden ser el email destino de las consultas. Se cambiará el editor por defecto de Joomla! por otro más completo que sí permite inserción de fotos, videos, etc., y por último indicar que la página principal de cada una de las versiones son entradas o artículos seleccionados para formar la presentación de la página principal del *blog*.

V. ANÁLISIS PRELIMINAR DE SEGURIDAD

Se ha llevado a cabo un estudio básico de las vulnerabilidades de las implementacions hechas con Joomla! y Drupal, ya que están destinadas a la creación del mismo tipo de portales web. Por este motivo hemos desestimado realizar el de la versión de Wordpress (destinada principalmente para realización de *sites* de tipo *blog*). Antes de comenzar, citamos algunas recomendaciones de seguridad cuando trabajamos con WCMS:

- Hacer copias de seguridad periódicas tanto de los archivos de nuestro WCMS como de la base de datos (exportarla)
- Contratar proveedores de hospedaje profesionales, así además estaremos más seguros contra ataques del tipo SQL (como explicaremos más adelante)
- Usar la versión más reciente del WCMS y de los *plugins* instalados en él
- Usar *plugins* específicos para seguridad como puede ser JHackGuard para Joomla!, para dotar de seguridad extra
- Restringir el acceso a archivos y carpetas de administración
- Eliminar el *script* de instalación (por ejemplo *install.php* en Drupal o la carpeta llamada *installation* en el caso de Joomla!)
- Modificar las contraseñas por defecto y definir roles de usuario seguros
- Habilitar CAPTCHA para usuarios no registrados, de este modo evitaremos el spam
- Activar URLs amigables
- Cambiar la configuración por defecto en los parámetros globales del *site*
- Cambiar el prefijo por defecto para las tablas de la base de datos durante la instalación si el WCMS en concreto nos lo permite
- Evitar mostrar por error información sensible sobre el WCMS en la parte visible al usuario (*front-end*)

Los ataques más comunes sufridos por los WCMS son del tipo SQL *injection*, que consiste básicamente en que un usuario accede al *site* pudiendo alterar la base de datos, y de tipo *Cross-site scripting* (conocido como XSS), que se produce cuando un usuario mal intencionado encuentra la forma de insertar un fragmento de código malicioso en nuestra web [12]. Más específicamente, *SQL injection* es una técnica en la que el atacante inserta caracteres o palabras clave (*keywords*) en una sentencia SQL mediante parámetros de entrada de usuario sin restricciones para cambiar la lógica de la consulta deseada [19]. Por su parte, un ataque XSS consiste en inyectar una secuencia de comandos maliciosa en un sitio web de confianza que se ejecuta en el navegador de un visitante sin el conocimiento del visitante y, por lo tanto, permite al atacante acceder a datos de usuario sensibles, como *tokens* de sesión y *cookies* almacenados en el navegador [20]. Mientras que la inyección de SQL se dirige a la función de consulta que interactúa con la base de datos, XSS explota la función de salida HTML que envía datos al navegador.

Para llevar a cabo el análisis de vulnerabilidades de nuestras dos versiones Joomla! y Drupal vamos a hacer uso de una herramienta diseñada para tal fin denominada Acunetix [21]. Tras realizar las pruebas de *SQL injection* y XSS con Acunetix sobre nuestras dos versiones del *site* obtenemos los resultados que se muestran en la Tabla III. Lo primero que debemos decir acerca de los resultados es que respecto a inyección SQL y XSS ambos son seguros, en contraposición a lo ocurrido en [12]. Todas las alertas son de nivel de riesgo bajo y de índole similar en todos los casos. No hay que alarmarse tras ver las alertas que aparecen ya que la gran mayoría de éstas son referentes a contenido o extensiones que han sido borradas o desinstaladas de nuestros WCMS durante el desarrollo del proyecto, y por este motivo aparece contenido sin indexar, para el cual nos recomienda Acunetix borrar manualmente.

Por otro lado nos muestra una alerta relacionada con el módulo de *login* en las cuatro pruebas, ya que tenemos activada la opción de autocompletar, y nos recomienda desactivarla. Por último en este apartado citar que en el caso de las pruebas sobre Joomla! nos indica que a través de la extensión *Jdownloads* se permite la subida de archivos por parte de usuarios, algo peligroso aunque en este caso somos conscientes de ello pues se trata de un requisito de diseño.

Tabla III
ALERTAS TRAS MEDICIONES CON ACUNETIX

	Joomla!	Drupal
SQL injewction	Total de alertas 133 Riesgo de nivel 1 o bajo	Total de alertas 78 Riesgo de nivel 1 o bajo
XSS	Total de alertas 122 Riesgo de nivel 1 o bajo	Total de alertas 9 Riesgo de nivel 0

Tanto Joomla! como Drupal tienen una comunidad extensa de usuarios y desarrolladores muy activa y siempre pendiente de incluir mejoras de las versiones disponibles en concepto de seguridad y por supuesto de reportar los fallos y las posibles soluciones a estos. Por supuesto las versiones posteriores corrigen los fallos o vulnerabilidades que presentaban las anteriores. Ambos deberían de tener más recomendaciones durante la instalación con respecto a carpetas y archivos importantes que hay que proteger a posteriori, y sobre todo deberían indicarnos que no es seguro usar el nombre de usuario de administración por defecto y los prefijos por defecto para las bases de datos. Por el contrario, sólo nos avisan de que borremos los archivos de instalación tras realizar esta. En ambos WCMS tenemos módulos adicionales desarrollados por terceros para dotar de seguridad extra al *site* como puede ser *Taxonomy Access Control* que se basa en el uso de roles de usuario para dotar de seguridad extra a un *site* Drupal o el módulo Marco's SQL *Injection* que permite proteger contra inyección SQL e inclusión de ficheros en un *site* Joomla!.

Por otro lado ambos WCMS están programados e incorporan directivas de protección contra ataques de tipo *SQL injection* y XSS, además de estar dotados de un sistema de reconocimiento interno para comprobar las terminaciones y las extensiones de los archivos que sean subidos mediante sendos gestores. Además los componentes de terceros que puedan ser configurados para permitir subidas de archivos por parte de un visitante, como puede ser una galería de imágenes, un formulario de contacto o un gestor de descargas, disponen de medidas propias de seguridad en su configuración, que van desde el uso de *captcha* hasta filtrar por IP a los usuarios que accedan al *site*.

Por último comentar que en el caso de Drupal debemos desactivar las alertas desde su propia configuración para no facilitar el trabajo a un usuario mal intencionado, ya que de no hacerlo estaríamos mostrando nuestras debilidades al atacante. La Tabla IV resume el análisis básico de seguridad realizado.

VI. CONCLUSIONES

Los WCMS permiten distribución de contenidos online para usuarios con escasos conocimientos informáticos y son relativamente sencillos de utilizar y gestionar. En cuestión de funcionalidades y apariencia podemos conseguir prácticamente cualquier resultado que deseemos para una web. En nuestro caso, hemos conseguido realizar tres *sites* prácticamente iguales en apariencia y aspecto mediante el uso de diferentes WCMS, aunque hemos requerido del uso de otro WCMS para dotar de la funcionalidad de gestión de archivos multiusuario en las versiones de Wordpress y Drupal debido a que las funcionalidades que estos ofrecían en el momento de realización de este trabajo no era tan extensa como en el caso de Joomla!. Todos los WCMS son bajo licencia libre y se pueden obtener en sus páginas web oficiales, necesitan de una base de

Tabla IV
RESUMEN DEL ANÁLISIS BÁSICO DE SEGURIDAD

	Joomla!	Drupal
¿Tiene comunidad?	Sí	Sí
¿Tiene informes de vulnerabilidad?	Sí	Sí
¿Hay sugerencias de seguridad en la instalación?	Sí. Se nos indicará que borremos la carpeta de instalación.	Sí
¿Existen módulos o componentes a terceros para dotar de seguridad extra?	Sí	Sí
Existen contramedidas para *SQL injection*	Sí	Sí
Warnings cuando vas a usar módulos de terceros	Siempre	Siempre
Comprueba las terminaciones de los archivos en subidas	Sí	Sí
Apis contra XSS	Sí	Sí
Protegido contra XSS	Sí	Sí
Usuario por defecto de administración	Inseguro	Inseguro
Protección contra *spam*	Básica. Suelen tener ambos puntos débiles en este campo en el registro de nuevos usuarios.	Básica. Es recomendable usar *Captcha*
Versiones recientes que solucionan problemas de versiones anteriores	Sí	Sí
Avisos de nuevas actualizaciones de componentes y módulos	Sí	Sí
Incluyen archivos que es necesario proteger mediante permisos	Sí. En ambos CMS debemos proteger de forma extra algunos archivos claves	Sí
Warnings peligrosos en *front-end*	No	Sí. Desactivarlos

datos para almacenar en ellas la información y se instalan a través del navegador web a través de un asistente.

En base al trabajo realizado, concluimos que en cuanto a la administración para gestión de contenidos y funcionalidades Joomla! ofrece el entorno más intuitivo, Drupal el más complicado de gestionar, y Wordpress tiene la ventaja de poder buscar estas funcionalidades desde su propio *back-end*. Respecto al importante apartado de dotar de aspecto a un *site*, los tres gestores disponen de un creador de plantillas Artisteer que además es muy sencillo de utilizar, esto nos otorga la posibilidad de conseguir prácticamente el resultado estético deseado.

Por otra parte las características más destacables de los WCMS analizados son las siguientes. Joomla! ofrece un enorme número de funcionalidades comparado con los otros WCMS estudiados y posee una comunidad de usuarios altamente activa en comparación con Drupal y Wordpress, lo que permite tener siempre disponible una ayuda importante ante cualquier dificultad que pueda aparecernos en el camino. Drupal posee el concepto de seguridad y de roles de usuario más potente y Wordpress es el mejor en términos de posicionamiento SEO, dando además la posibilidad de obtener un alojamiento gratuito. Por último comentar que tras dotar de seguridad extra a los *sites* y llevar a cabo el estudio de seguridad podemos afirmar que tanto Joomla! como Drupal son robustos frente a los ataques más comunes en la web como son SQL *injection* y *Cross-site scripting* (XSS). No obstante, pensamos que la seguridad de este tipo de sistemas seguirá siendo objeto de estudio. Desde el uso de *phising* hasta *malware* específico para desarrollos en PHP, nuevas amenazas surgen constantemente. Lo que implicará una continua actualización y aplicación de medidas de prevención, detección y recuperación especificas para este tipo de sistemas.

AGRADECIMIENTOS

This research was supported by the AEI/FEDER, UE project grant TEC2016-76465-C2-1-R (AIM).

REFERENCIAS

[1] W. F. Cody, J.T. Kreulen, V. Krishna, W. S. Spangler, "The integration of business intelligence and knowledge management", IBM Systems Journal, Vol. 41 (4), pp. 697-713, 2002.

[2] S. Bergstedt, S. Wiegreffe, J. Wittmann, D. Moller, "Content management systems and e-learning systems -a symbiosis?", Proc. 3rd IEEE International Conference on Advanced Technologies, pp. 155-159, 2003.

[3] R. McDaniel, J. R. Fanfarelli, R. Lindgren, "Creative Content Management: Importance, Novelty, and Affect as Design Heuristics for Learning Management Systems", IEEE Transactions on Professional Communication, Vol. PP (99), pp. 1-18, 2017.

[4] D. Barker, "Web Content Management: Systems, Features, and Best Practices", O'Reilly Media, 1st Ed, 2016. ISBN 978-1491908129.

[5] A. J. Aledo Hernández, "Sistemas de gestión de contenidos web: uso y estudio comparativo de su seguridad". Proyecto Fin de Carrera, I.T.T. especialidad Telemática. Directora: María Dolores Cano Baños. Octubre 2015

[6] S. K. Patel, V. R. Rathod, J. B. Prajapati, "Performance Analysis of Content Management Systems - Joomla, Drupal and Wordpress", International Journal of compuer Applications, vol. 21 (4), pp. 39-43, 2011.

[7] Joomla. Disponible en: <http://www.joomla.org>. Último acceso mayo 2017.

[8] Wordpress. Disponible en: <http://www.wordpress.com>. Último acceso mayo 2017.

[9] Drupal. Disponible en: <http://www.drupal.org>. Último acceso mayo 2017.

[10] W3Techs. Web Technology Surveys. Disponible en: https://w3techs.com/technologies/overview/content_management/all. Último acceso julio 2017.

[11] A. Mirdha, A. Jain, K. Shah, "Comparative Analysis of Open Source Content Management Systems", Proc. IEEE International Conference on Computational Intelligence and Computing Research, pp. 1-4, 2014.

[12] M. Meike, J. Sametinger, A. Wiesauer, "Security in Open source Web Content management systems", Internet Security & Privacy, Vol. 7 (4), pp. 44-51, 2009.

[13] H. Jerković, P. Vranešić, S. Dadić, "Securing web content and services in open source content management systems", Proc. 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1402 – 1407, 2016.

[14] Proyecto Webscarab OWASP. Disponible en: < https://www.owasp.org/index.php/Proyecto_WebScarab_OWASP>. Último acceso julio 2017.
[15] A. O'Donell, "Tamper Data: The Firefox Add-on", Lifewire, abril 2017. Disponible en: < https://www.lifewire.com/firefox-addon-that-hackers-dont-want-you-to-know-about-2487289>. Último acceso julio 2017.
[16] S. K. Shivakumar, "Enterprise Content and Search Management for Building Digital Platforms", Wiley-IEEE Press, 2017. ISBN 9781119206842
[17] Artisteer. Disponible en: <http://www.artisteer.com>. Último acceso mayo 2017.
[18] Mailchimp. Disponible en: <http://www.mailchimp.com>. Último acceso mayo 2017.
[19] L. Khin Shar and H. Beng Kuan Tan, "Defeating SQL Injection", Computer, vol. 46 (3), pp. 69-77, 2013.
[20] I. Yusof and A. K. Pathan, "Mitigating Cross-Site Scripting Attacks with a Content Security Policy", vol. 49 (3), pp. 56-63, Computer, 2016.
[21] Acunetix. Disponible en: <http://www.acunetix.com>. Último acceso mayo 2017.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6576

Sistema de cifrado basado en contexto aplicado a prevención de fuga de datos

Alberto García[1], Pilar Holgado[1], Jose Javier Garcia[2], Jorge Roncero[2], Víctor A. Villagrá[1], Helena Jalain[1].

[1]Departamento de Ingeniería y Sistemas Telemáticos, Universidad Politécnica de Madrid, Avenida Complutense, 30, 28040, Madrid, España.

[2]Nokia, Departamento de Innovación, Calle de María Tubau, 9, 28050, Madrid, España

garciamoro@dit.upm.es, pilarholgado@dit.upm.es, jose_javier.garcia_aranda@nokia.com, jorge.roncero_mayoral@nokia.com, villagra@dit.upm.es, hjalain@dit.upm.es.

***Resumen*- Las herramientas DLP (Data Leak Prevention) están adquiriendo un valor elevado en los últimos años debido a la importancia de proteger los datos sensibles de una organización. Muchas de las herramientas DLP se basan principalmente en la analítica de datos, ya sea un análisis de archivos almacenados o estando en tránsito por la red. La solución DLP propuesta usa el cifrado basado en contexto para evitar fugas de información. La clave de cifrado y descifrado se obtiene a partir de la ejecución de un conjunto de retos basados en el contexto de entorno y en las políticas de la empresa. En este artículo se explica la arquitectura y el diseño de la solución DLP y de los retos propuestos.**

***Palabras clave*- Data Leakage Prevention, cifrado basado en contexto.**

I. Introducción

Hoy en día, muchas empresas se ocupan de datos sensibles, incluyendo la propiedad intelectual, información financiera o información personal de los usuarios. La distribución accidental o no intencionada de datos privados a una entidad no autorizada es un problema grave para las empresas. El daño potencial de la fuga de datos puede influir en la reputación de la empresa, en la exposición de la propiedad intelectual a los competidores, o en la pérdida de ventas futuras.

En el contexto de la fuga de datos, el atacante puede ser un empleado interno o un atacante externo que intenta obtener información sensible. Incluso, no siempre es causada con mala intención, sino también por un error inadvertido. Además, un usuario autorizado no es el mismo que un usuario de confianza. En muchos casos, las organizaciones son víctimas de sus propios empleados que comparten intencionadamente datos confidenciales con personas externas con fines personales [1]. En este caso, el usuario está autorizado a acceder a información sensible y no es detectado a partir de medidas externas clásicas como cortafuegos.

La Prevención de Fugas de Datos o Data Leak Prevention (DLP) [2] se ha propuesto como una solución a estos problemas. Distintas soluciones DLP han sido estudiadas tanto en áreas de investigación académica como en aplicaciones prácticas. Sin embargo, la fuga de datos y el uso indebido de información se sigue considerando una amenaza emergente para las organizaciones, especialmente cuando son llevadas a cabo por sus propios empleados. En muchos casos, es muy difícil detectar a los usuarios internos porque hacen mal uso de sus credenciales para realizar un ataque.

En este artículo, proponemos una solución DLP que aplica el concepto de cifrado basado en el contexto. Esta propuesta se basa en un proceso de cifrado/descifrado de documentos confidenciales, donde la clave de cifrado se obtiene a través de la ejecución de un conjunto de retos. Estos retos utilizan el contexto del entorno y las políticas de la empresa en el momento de cifrado/descifrado. De esta manera, los archivos sensibles están cifrados en todo momento y sólo se pueden leer dentro de nuestro sistema DLP.

El resto del trabajo se organiza como sigue. Los antecedentes sobre soluciones DLP se describen en la Sección 2. La Sección 3 describe el estado actual de distintas técnicas de cifrado basado en contexto. La Sección 4 explica el sistema DLP usando cifrado basado en contexto. Los retos propuestos se explican en la

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Sección 5. La Sección 6 describe cómo se genera la clave de cifrado en base a los resultados de los retos. Finalmente, las conclusiones finales obtenidas durante este estudio se incluyen en la Sección 7.

II. Herramientas DLP

Una herramienta DLP [3] es una utilidad que ayuda a mantener segura la información confidencial, evitando posibles filtraciones o difusiones no autorizadas de información.

A. *Qué se protege*

Dentro de esta categoría podemos encontrar: Datos permanentes almacenados en el disco duro protegidos mediante cifrado o control de acceso, datos en uso utilizando medidas de limitación de acciones como copiar-pegar y realizar capturas de pantalla, y datos en tránsito por la red, ya sea entre equipos de la red interna o con un host externo utilizando controles sobre la red, como son la detección y la inspección de los datos transmitidos.

B. *Dónde se protege*

El despliegue se puede llevar a cabo en dispositivos finales para monitorizar y controlar el acceso a los datos desde los dispositivos finales. En este caso es necesario un servidor de supervisión remoto que se haga cargo de las tareas administrativas, la distribución de políticas y la generación de eventos de registro. O despliegue en red mediante análisis del tráfico de la red y sujeto a una política predefinida, incluso activar eventos y bloquear transmisiones sospechosas.

C. *Cómo se protege*

La protección se puede llevar a cabo de distintas maneras. Una de ellas es la inspección basada en contexto que consiste en inspeccionar el contexto de un fichero, como el tamaño o el tipo de archivo. Otra opción es la inspección basada en contenido que detecta la fuga de información mediante el análisis de contenido, utilizando distintas técnicas como análisis de lenguaje natural o estadísticas.También existen métodos basados en el establecimiento de una política para el cifrado de los ficheros confidenciales y el acceso a dicha información. Además también son utilizados métodos como el etiquetado o el control de acceso.

Nuestra propuesta de solución DLP añade una política de seguridad adicional, haciendo que las claves de cifrado de los ficheros sean obtenidas de una función dentro de un contexto específico, el cual puede ser personalizado y parametrizado por un administrador.

III. Cifrado basado en contexto

El cifrado basado en contexto es una forma de autorizar a los usuarios para acceder a información si cumplen una serie de requisitos de entorno. Este tipo de cifrado puede ser usado en diferentes entornos, como por ejemplo en entornos de Internet of Things (IoT) [4], utilizando diferentes metodologías o con servidores exteriores que comprueben el contexto.

El cifrado basado en contexto para entornos de IoT se puede definir de dos formas [4]: pueden establecerse los datos de contexto en el usuario o pueden establecerse en la propia información a proteger. A continuación se detallan ambos modelos.

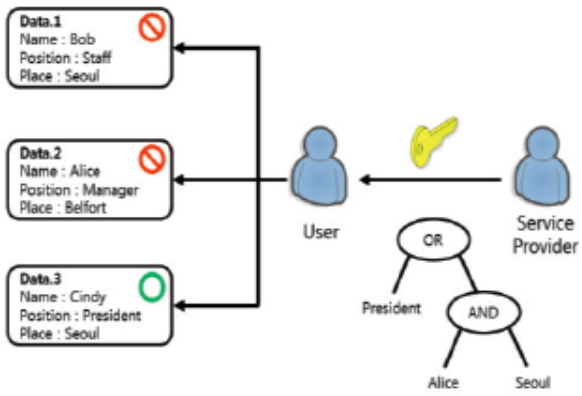


Fig. 1. Ejemplo de un árbol de autorización

En la Figura1 se puede ver un árbol de autorización basado en funciones lógicas. En este caso, lo que hace el proveedor de servicios es generar un árbol de operaciones lógicas que tiene que resultar en TRUE para poder acceder a los datos y que son suministrados al usuario junto con la clave. De esta forma, los datos tienen una serie de atributos y si cumplen el árbol de autorización del usuario, este dato estará accesible para ese usuario específico.

Por otro lado, se puede realizar lo opuesto: los usuarios tienen una serie de parámetros asociados y a los datos se les asigna un árbol de autorización, como se puede ver en la figura 2. Estos datos están cifrados, aunque la clave de cifrado se guarda en el mismo lugar que el dato en sí. En este caso, el usuario que tenga los parámetros necesarios que validan el árbol están autorizados a obtener el fichero, pero dichos parámetros no forman parte de la clave. Este sistema puede ser menos seguro debido a que la contraseña de descifrado está en el propio fichero y que su valor no depende de los parámetros.

Por otro lado, el cifrado basado en contexto también puede ser implantado mediante un servidor externo que haga las comprobaciones de contexto [5]. De esta manera, las comprobaciones del contexto se resuelven de forma externa, sin necesidad de que el árbol esté con el usuario o con el dato. Este servidor devuelve la respuesta en forma booleana para indicar si se puede acceder o no a los datos.

El cifrado basado en contexto se puede utilizar para distintas aplicaciones [6]. Un caso de ejemplo es cuando un proveedor desea compartir o habilitar el acceso a datos basándose en las credenciales del usuario receptor. El proveedor de datos proporciona una función f (*) donde describe cómo quiere compartir o habilitar el acceso a los datos y asigna al usuario una clave secreta

con credenciales X. Si f (X) = 1 (u otro resultado fijado), el usuario puede descifrar los datos.

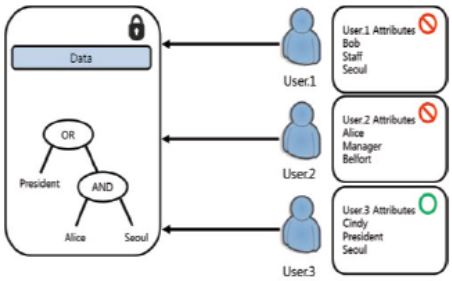


Fig. 2. Ejemplo en árbol de autorización

Las credenciales de este usuario están representadas por un conjunto de datos de tipo *string* y se conoce como cifrado basado en atributos o Attribute-Based Encryption (ABE). La función es representada por una fórmula de tipo matemático sobre estos atributos, que en este caso se corresponden con cadenas de texto pero pueden ser números u otros parámetros. Esta fórmula variará su estructura y tipo dependiendo de cómo se quieran manejar los parámetros y de su tipo.

Podemos definir dos formas de cifrado basado en atributos. En la primera de ellas, llamada Key-Policy ABE, los atributos se anotan en los datos cifrados y la fórmula f (*) se da al usuario, como en la figura 1. Por otra parte, existe la otra modalidad, llamada Ciphertext-policy ABE, donde los atributos se utilizan para las credenciales del usuario y la fórmula f (*) se encuentra junto con los datos, como en la Figura 2.

En otra metodología [7] se propone un control de acceso que se expresa con una matriz Linear Secret Sharing Scheme (LSSS) sobre los atributos en lugar de las clásicas estructuras de árbol equivalentes. Sin embargo, estas matrices LSSS son mucho menos intuitivas de usar que las fórmulas booleanas o árboles de acceso [8].

La diferencia principal entre las metodologías radica en la situación de los datos o atributos de autorización, que pueden encontrarse en la información en sí o pertenecer al propio usuario. Además, también hemos visto comprobaciones de contexto con servidor exterior, que da más seguridad y flexibilidad a la hora de configurar y comprobar contextos de usuarios.

En nuestra propuesta aplicamos este concepto de contexto de los usuarios, como su nombre, su identificador de usuario o incluso su localización para hacer que estos parámetros formen parte de la clave que dará acceso a un fichero cifrado. Así, si no se cumple ese contexto específico, la clave generada será incorrecta.

IV. Cifrado basado en contexto aplicado a DLP

Nuestra propuesta se basa en construir una herramienta de DLP utilizando el contexto de los usuarios para la autenticación y el cifrado/descifrado de los datos sensibles de una organización. Se propone utilizar, no sólo los atributos tradicionales de usuarios como son el nombre de usuario y su rol en la empresa, sino que vamos un paso más allá, incluyendo el contexto de entorno de los equipos, como la hora, la geolocalización, etc.

Aplicar el cifrado basado en contexto en una herramienta DLP proporciona mayor seguridad en el acceso a los ficheros por parte de los usuarios autorizados, ya que estos usuarios acceden a los datos confidenciales o sensibles en condiciones controladas y seguras. En nuestra propuesta, los usuarios solo podrán escribir y leer datos en el contexto que ha sido configurado por el administrador, y por tanto, únicamente cumpliendo este contexto se puede acceder a la información sensible. De esta forma se puede asegurar que el contexto de entorno del usuario es seguro tanto en el momento de escritura como en el de lectura. Cuando el contexto no sea correcto, ya sea por parte de un usuario no autorizado o un empleado autorizado que no se encuentra en el contexto apropiado, los datos no podrán ser descifrados correctamente con lo que se evita la fuga de información.

Para el correcto funcionamiento de esta propuesta, suponemos que el contexto siempre debe ser el mismo tanto en el momento de escritura de un nuevo documento como en el de lectura del mismo, es decir, los usuarios utilizan los equipos en el mismo intervalo de horas, en el mismo lugar, etc. De esta manera se pueden generar las claves de cifrado/descifrado de los distintos ficheros, debido a que siempre se calcularán las mismas si el contexto es el adecuado y serán diferentes si el contexto es inadecuado.

Por tanto, la clave para el cifrado/descifrado de un documento se obtiene a partir de la ejecución de una serie *retos*. Cada uno de estos *retos* procesan la información de contexto recibida, para calcular una subclave. Todas estas subclaves darán lugar a la clave final para el cifrado y descifrado de dicho fichero.

Esta propuesta de cifrado basado en contexto aplicado a DLP se integra en dos proyectos de investigación CiberNoid y DroneFS [9].

En el proyecto DroneFS se propone una arquitectura para el cifrado basado en contexto de la información que recopilan los drones (figura 3).

Esta arquitectura es la más adecuada para su uso en drones, por ejemplo para uso militar, donde los *retos* solo pueden ser ejecutados de forma local. Sin embargo, para integrar esta herramienta DLP en una organización, en la que el acceso a los datos se puede realizar a través de distintos tipos de dispositivos, como móviles o portátiles, añadir un servidor exterior puede facilitar la administración de nuevas políticas basadas en el contexto, aportar mayor seguridad dificultando la manipulación de los datos de contexto y permitir la ejecución de los retos fuera del dispositivo del cliente aliviando por ejemplo la carga de procesamiento y el gasto de batería de los dispositivos móviles. Este es el caso de la arquitectura del proyecto CiberNoid (Figura

4) que incluye un servidor externo para comprobar la autorización del usuario y ejecutar *retos remotos*.

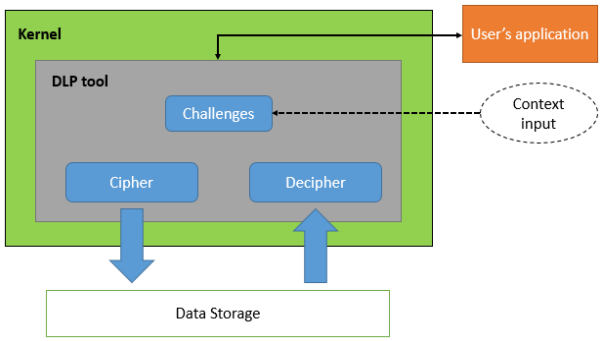


Fig. 3. Arquitectura DLP

En la figura 4 se puede ver que la arquitectura propuesta para la protección de ficheros sensibles está compuesta principalmente por la *Herramienta DLP* instalada en cada dispositivo cliente y el *Servidor externo* de la organización.

La *Herramienta DLP* propuesta se encarga de capturar a nivel de *Kernel* las llamadas que las aplicaciones realizan para el manejo de ficheros. En concreto, cuando el usuario crea un nuevo fichero y lo quiere guardar en el disco duro se cifran los datos. De igual forma, cuando un usuario abre un fichero, la herramienta DLP se encarga de descifrarlo antes de que sea visualizado en la aplicación final del usuario. Todas estas operaciones se realizan de manera totalmente transparente al usuario, sin que sea consciente de esta protección aplicada a la información, ni de los *retos* y parámetros necesarios. La herramienta propuesta no necesita estar instalada en todo el disco duro destinado a datos de la máquina, es decir, que se puede tener distintas particiones en disco y que sólo sobre una de ellas se aplique el cifrado de los ficheros. Por tanto, sólo las peticiones sobre estos ficheros, realizadas desde distintas aplicaciones, serán capturadas por nuestra herramienta a bajo nivel. Las peticiones de aplicaciones que no utilicen datos almacenados en dicha partición de disco no serán capturadas por nuestra herramienta, siguiendo su funcionamiento habitual.

El *Servidor externo* de la organización es un servidor HTTP con una API REST, el cual se utiliza para atender distintas peticiones de ejecución de los *retos remotos* recibiendo el contexto como parámetro y devolviendo un JSON con las subclaves calculadas. Además, tiene una base de datos donde el administrador almacena la información de los usuarios autorizados y cada uno de los parámetros necesarios para ejecutar los distintos *retos* configurados siguiendo la política de la empresa.

Cuando se quiere abrir o guardar un fichero, la llamada de la aplicación de usuario se captura a nivel de *Kernel* para que la herramienta DLP pueda realizar el cifrado/descifrado de la información. El primer paso es obtener los *retos locales* y *los retos remotos* asociados al equipo a partir de un fichero de configuración incluido por el administrador. La herramienta DLP ejecuta los *retos locales* directamente en el dispositivo y realiza una petición POST al servidor externo con los parámetros necesarios para la ejecución de los *retos remotos*.

En la Figura 5 se muestra el proceso que se lleva a cabo en la ejecución de *retos remotos*. Cuando un dispositivo necesita cifrar o descifrar un fichero, la herramienta DLP recoge los datos de contexto y los envía al servidor externo. Este servidor ejecuta todos los *retos remotos* utilizando tanto el contexto recibido para cada uno de ellos como las políticas de la empresa almacenadas por el administrador en la base de datos. Cada uno de los *retos remotos* calcula una subclave. Estas subclaves puede ser correctas o no dependiendo de los datos de contexto enviados al servidor.

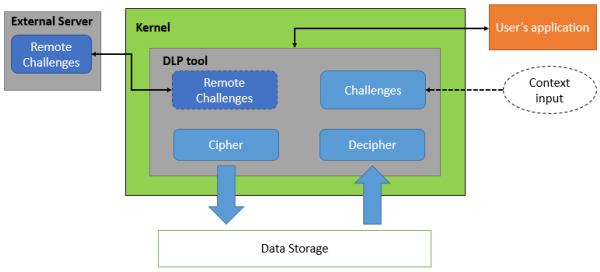


Fig. 4. Arquitectura DLP con servidor exterior

Una vez obtenidas todas las subclaves, tanto de los *retos locales* como de los *retos remotos*, la herramienta DLP calcula la clave de cifrado, como se detalla en la sección VI-C, y procede al cifrado/descifrado de la información.

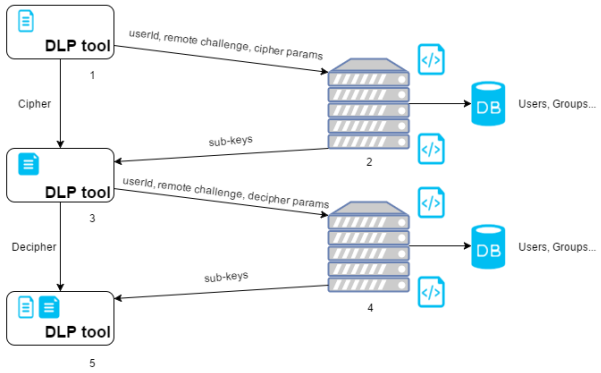


Fig. 5. Flujo del servidor con el FileSystem

V. Retos

Los *retos* son el mecanismo que permite calcular la clave de cifrado a partir de un conjunto de datos de contexto. En este sentido, se han diseñado una serie de *retos* capaces de generar una subclave a partir de unos datos de entrada. Cabe destacar que *los retos* que ejecuta el servidor tanto en el proceso de cifrado como en el descifrado de documentos son las mismas, por lo que no hace falta hacer ningún tipo de diferenciación entre las peticiones de cifrado y descifrado.

Los *retos remotos* propuestos se basan en geolocalización, la hora, la fecha, el operador de

telefonía y las redes wifi al alcance. En concreto, se proponen tres *retos* para determinar dónde se encuentra el usuario, basadas en la geolocalización, el operador de telefonía y las redes wifi al alcance; y el cuándo se determina mediante la ejecución de los *retos* de fecha y hora. En los siguientes subapartados se explica en mayor detalle cada uno de ellos.

A. Reto de localización GPS

Gracias al GPS se puede localizar a un usuario y comparar su situación geográfica con la que debería tener, por ejemplo la situación de la oficina. En este caso, el reto consiste en limitar el acceso a los ficheros confidenciales en todos los lugares que no estén acotados dentro de una zona delimitada por el administrador del sistema. Esta zona queda delimitada por dos parámetros incluidos por el administrador en la base de datos del servidor:

- Centro: coordenadas geográficas del centro del área permitida. Por ejemplo, en un edificio de oficinas, se almacena el centro del edificio.
- Área a cubrir: por defecto, el área a implementar será circular, por lo que el administrador solo deberá dar como dato la medida del radio del círculo a cubrir por sistema.

Con estos dos parámetros, la zona queda perfectamente delimitada. Siempre que se acceda desde dentro de este área se obtiene la misma subclave para un fichero concreto, mientras que fuera de ella se retorna un valor de subclave aleatorio y no válido.

Cuando un equipo necesita cifrar o descifrar un fichero realiza una petición POST al servidor con las coordenadas geográficas en las que se encuentra. En primer lugar, el *reto* utiliza el área y el centro almacenado en la base de datos por el administrador, para calcular los 4 puntos del círculo cuyas latitudes sean mayores y menores (es decir, los valores mínimos y máximos de latitud que se puedan dar en todos los puntos del interior del círculo) y comprueba cuál es la parte común entre los puntos. Por ejemplo, los puntos de latitud 3.4567º y 3.4589º tienen en común la parte 3.45º. O dicho de otra manera, tienen en común los 3 primeros dígitos de la coordenada. Este proceso se hace tanto para la latitud como para la longitud, obteniendo el número de dígitos invariantes en cada uno de ellos, generando 2 variables que contienen el número de dígitos invariantes en cada una de las dimensiones del área. Una vez hecho esto, el servidor sabe con cuántos dígitos de cada coordenada enviada como parámetro en la petición se tiene que quedar , obteniendo 2 valores para generar la clave.

Una posible complejidad podría venir dada por los lugares cuya coordenada cambie completamente. Por ejemplo, desde la latitud 3.9986 hasta la latitud 4.0056. En este caso, el sistema redondea las coordenadas al número x,y más cercano, siendo “x” la parte entera e “y” el primer decimal. Si la longitud presentara el mismo problema, se haría de la misma forma para obtiener los 2 valores necesarios para generar la clave.

B. Reto de fecha

Un reto básico es la comprobación de la fecha actual en el momento de intentar descifrar un fichero. Así, se puede limitar su acceso según la fecha, como por ejemplo poder descifrar los archivos pertenecientes a un proyecto en las fechas en las que se está trabajando en él. Este *reto* podría resolverse de manera local, pero, dado que el cambio de fecha en un dispositivo suele ser muy fácil de llevar a cabo y se podría engañar al sistema, es mejor ejecutarlo desde el servidor.

En este caso, se comprueba si un fichero se puede abrir en una franja de fechas predefinidas en el equipo. Nuestra propuesta se basa en utilizar una máscara que determine la duración del rango de fechas válido, como en direccionamiento IP. La idea es que variando la máscara y haciendo una operación del tipo *Fecha AND Mascara*, se pueda calcular siempre la misma subclave si se está en el rango correcto.

Para llevarlo a cabo, se van a codificar los meses en binario según su orden, de forma que los meses cercanos entre sí compartan el mayor número de bits posibles para poder hacer uso de la máscara y que se puedan determinar distintos rangos de fecha. Además, los meses se van a dividir en quincenas, de forma que la primera quincena de un mes tendrá una codificación distinta a la segunda. Por tanto, para codificar las 24 quincenas que hay en un año, se tienen que utilizar 5 bits (como mínimo). Una posible codificación sería: enero (00000, 00001), febrero (00010, 00011).... Con esta forma de codificar los meses, se consigue configurar distintos períodos de tiempo según la longitud de la máscara. Por ejemplo:

- Máscara 11111: Se corresponde con un periodo de una quincena, ya que al hacer *Time AND Mascara* (entendiendo *Time* como la fecha actual según nuestra codificación) nos quedamos con los 5 bits de la codificación. Estos bits solo pertenecen a una quincena concreta, es decir, si se hace la misma operación en quincenas distintas, el resultado sería distinto y la subclave no coincidiría.
- Máscara 11110: Sigue el mismo principio que el caso anterior, solo que ahora la validez es de un mes entero, ya que las 2 quincenas del mismo mes comparten los primeros 4 bits, y por tanto la subclave seguiría siendo la misma.

Como la máscara sólo tiene en cuenta los meses, podría darse el caso de que un fichero se pueda abrir cada enero (o el periodo que sea) de cada año. Para evitar eso, la clave generada también tendrá en cuenta el año actual en el momento de cifrado, haciendo que la función que calcula la clave reciba como parámetro el año.

Otra consideración a tener en cuenta es el día del mes en que comienza el tramo de fechas, como por ejemplo cifrar un fichero para un mes a finales de ese mes. Como el tramo es cerrado, solo será válido ese mes concreto, por tanto cuando empiece el mes nuevo, el cifrado dejará de tener validez. Para evitar los

problemas de cifrar un fichero en los últimos días de un periodo y que luego no sea válido, se introduce un offset que se corresponde con el día de creación del fichero, de forma que la codificación de los meses es dinámica. Por ejemplo, si un fichero se cifra el día 4 de enero, una codificación de los meses dinámica debe contemplar que las dos quincenas incluyan del 4 de enero al 4 de febrero. De esta forma, para cada fichero, se haría una codificación distinta dependiendo del día de creación del mismo.

El único inconveniente de aplicar el dinamismo en los tramos de meses es que el procesamiento será mayor al tener que hacer una codificación dinámica cada vez que llega una petición al servidor. Además, se necesita que la petición POST realizada desde el dispositivo incluya como parámetro la fecha de creación del fichero.

C. *Reto de hora*

Otro posible *reto* es la comprobación de la hora en el momento de intentar cifrar/descifrar un fichero. Así, se puede limitar el tramo de horas en las que se permite manejar la información sensible, como por ejemplo abrir solo una serie de archivos en horario de oficina. Este *reto* podría resolverse de manera local, pero dado que el cambio de hora en un dispositivo suele ser muy fácil de llevar a cabo y se podría engañar al sistema, es mejor ejecutarlo en el servidor.

Este *reto* tiene que comprobar que un fichero se puede abrir en una franja de horas predefinidas en el equipo. Esto se podría hacer de diferentes formas, pero al igual que en el *reto* de fecha, se utiliza una máscara que determina la duración del rango de horas válido. La idea es que variando la máscara y haciendo una operación del tipo *Tiempo AND Mascara*, se pueda obtener siempre la misma subclave si se está en el rango de horas correcto. Es decir, cualquier petición a este *reto* siempre dará el mismo resultado si se hace durante el mismo rango de horas.

Para implementarlo, se van a codificar las horas del día en binario según su orden, de forma que las horas cercanas entre sí compartan el mayor número de bits posibles para poder hacer uso de la máscara y así determinar distintos rangos de hora. Es decir, cada hora tendrá una representación en binario, por lo que en total serán 24 horas y se necesitarán 5 bits para poder codificar todas las horas, con la posibilidad de añadir más bits de relleno. Codificando así las horas, se consigue configurar distintos períodos de hora según la longitud de la máscara. Por ejemplo:

- Máscara 11111: Se corresponde con un periodo de una hora, que es la unidad más pequeña codificada. Al realizar la operación lógica *Time AND Mascara* (entendiendo *Time* como la hora actual según nuestra codificación) nos quedamos con los 5 bits de la codificación. Estos bits solo pertenecen a una hora concreta, por lo que si se hace la misma operación en horas diferentes, el resultado sería distinto y la subclave no coincidiría.
- Máscara 11110: Sigue el mismo principio que el anterior caso, solo que ahora la validez es de 2 horas, ya que dos horas consecutivas comparten los primeros 4 bits, y por tanto la subclave seguiría siendo la misma.

En principio, contar con un período mayor de 8 horas de validez no tendría mucho sentido, ya que suele ser la jornada laboral y, además, el siguiente tramo se correspondería con 16 horas, un tramo que no es nada práctico.

Como pasaba en el reto de fecha, los tramos vuelven a ser estáticos, lo que ocasiona problemas similares. Para resolverlo, vamos a seguir la metodología del reto de fecha pero aplicado a variar la hora de inicio, para encontrar un intervalo de tiempo adaptable. Para consiguir una codificación dinámica de las horas es necesario configurar en el equipo la "hora de inicio" y así poder determinar los rangos concretos de horas de cada usuario.

D. *Reto de wifi*

Las redes Wifi que están al alcance del equipo se pueden utilizar para determinar la localización del usuario. Para determinar en qué lugar se puede acceder a los ficheros confidenciales el administrador almacena en la base de datos el SSID, el canal y la potencia mínima de las redes Wifi configuradas para resolver el *reto*. El valor de potencia mínima se utiliza para constatar que se está en el lugar especificado, como por ejemplo el edificio de la empresa y no en la calle a una distancia próxima.

Cuando un equipo necesita cifrar/descifrar un fichero realiza una petición al servidor con el SSID, el canal y la potencia de recepción de las redes Wifi que están a su alcance. Cabe destacar que el equipo no sabe cuáles son las redes configuradas para pasar correctamente el *reto*, por lo que tiene que mandar todas las redes Wifi que están a su alcance. Con estos datos el *reto* genera un trozo de la subclave utilizando cada red Wifi enviada en la petición que concuerde con algún valor de SSID y canal configurado por el administrador, siempre que sea alcanzada a la potencia mínima especificada. De esta forma, si se encuentran todas las redes Wifi necesarias, se obtiene la subclave completa y correcta, mientras que si falta alguna, la subclave generada estará incompleta y, por tanto, no será válida para descifrar el fichero del equipo.

E. *Reto de operador*

Si se tiene una lista de operadores telefónicos por país, se puede comprobar el operador de los dispositivos que tengan conexión telefónica para saber en qué país se encuentra y así tener otro parámetro de localización.

Normalmente, las empresas tienen su servicio de red y móvil con la misma compañía, por lo que el operador siempre es el mismo y puede ser una condición para poder descifrar el fichero.

Por tanto, se configura el reto para que genere una subclave, haciendo una serie de operaciones con el nombre del operador. Es decir, con cada operador se obtendrá una clave distinta y por tanto proporciona un impedimento más para el acceso a los datos sensibles.

F. Robustez

Todos los *retos* planteados tienen como parte común la generación de una subclave y se debe asegurar que a partir de ellas no sea posible conocer el contexto con las que fueron calculadas. De esta forma se evita que, aunque un atacante pueda interceptar la comunicación entre cliente y servidor, sea imposible conseguir el contexto de entorno válido. Esto se consigue con la aplicación de la función hash SHA-256. Por otro lado, cada uno de los *retos* utiliza un número y tamaño de parámetros de entrada distintos sobre los cuales se aplica una serie de operaciones dependiendo del *reto,* a los que se incluirá el valor de identificador del fichero y el identificador de usuario o departamento, asegurando así la aleatoriedad de los valores de subclave obtenidos para cada uno de los ficheros con independencia de que compartan el mismo contexto y que solo puedan ser visualizados por el grupo de usuarios permitido.

Cada uno de los *retos* calculan una subclave y todas ellas son necesarias para que el equipo pueda obtener la clave de cifrado/descifrado. En este sentido, aunque un atacante supiera el valor de una subclave o de todas ellas no sabría cómo obtener la clave final necesaria para el descifrado de la información robada.

Otra opción es que el atacante, ya sea externo o interno de la organización, tenga acceso a un dispositivo autorizado. Por tanto, este dispositivo tiene la herrammienta DLP (Figura 4), con el que es posible calcular la clave final. La principal dificultad radica en que para calcular la clave correcta de un fichero se necesitan cumplir todos y cada uno de los *retos*, por tanto, también sería necesario que el atacante tenga conocimiento del contexto válido y tener la posibilidad de cumplirlo.

Por ejemplo, en el caso de un empleado despedido, cuya empresa utilice este sistema, además de tener que seguir registrado en el servidor como usuario autorizado y de tener a su disposición un dispositvo cliente válido, tendría que generar un escenario que cumpla el contexto correcto para descifrar los ficheros. Esto es, generar la posición geográfica correcta, a la hora y la fecha correctas, con un operador determinado y teniendo a su alcance unas redes wifi determinadas, en el canal adecuado y la potencia necesaria. Y esto teniendo en cuenta solo los retos explicados anteriormente.

VI. Módulo de cifrado

El módulo de cifrado es el encargado de proteger los ficheros confidenciales y de generar la clave de cifrado para cada fichero a partir de los *retos*. Este módulo se encuentra en el dispositivo del cliente y será ejecutado cada vez que un usuario quiera realizar una operación sobre un documento confidencial.

A. Proceso de cifrado

El proceso necesario para el cifrado de un fichero se lleva a cabo mediante una serie de pasos. Primero se genera la clave asociada a los retos a partir de las subclaves obtenidas tras la ejecución de los *retos*. Tras ello, se obtiene el algoritmo de cifrado, su modo de ejecución y el algoritmo de autenticación a partir del fichero de configuración almacenado en el dispositivo cliente. Además, es necesario realizar el cálculo del vector de inicialización (valor aleatorio diferente para cada fichero) y almacenamiento de dicho valor en la cabecera del fichero. Por último, se realiza el cifrado del contenido del documento utilizando la clave final generada y el vector de inicialización.

B. Proceso de descifrado

El proceso de descifrado se realiza siguiendo una serie de etapas. Se inicia con la generación de la clave de cifrado a partir de las subclaves obtenidas tras la ejecución de los *retos*. En segundo lugar se obtiene el algoritmo de cifrado, su modo de ejecución y el algoritmo de autenticación a partir del fichero de configuración almacenado en el dispositivo cliente. Tras ello se procede a la autenticación de la cabecera del fichero si fuera necesario y la obtención del vector de inicialización. Y finliza con el descifrado del contenido del documento.

C. Obtención de clave

La clave final con la que se cifra o descifra el fichero debe calcularse a partir de las subclaves obtenidas de la ejecución de los retos. El objetivo es añadir complejidad al algoritmo que genera la clave final para que sea difícil reproducir su comportamiento.

El algoritmo debe poseer dos propiedades importantes, que sea computacionalmente eficiente y resistente a colisiones. Una colisión ocurre cuando dos entradas distintas producen el mismo resultado. En nuestro caso podría ocurrir si distintas subclaves produgeran la misma clave final. Esto invalidaría en gran parte la funcionalidad de los *retos*. Para evirlarlo, vamos a utilizar una función Hash [10] teniendo en cuenta sus características principales: Aceptan cadenas de cualquier tamaño como entrada, producen una salida con un tamaño fijo, son computacionalmente eficientes, son unidireccionales (difíciles de invertir), resistentes a colisiones (propiedad inyectiva) para un tamaño suficientemente grande de la cadena, deterministas, es decir que para una misma entrada producen un mismo resultado.

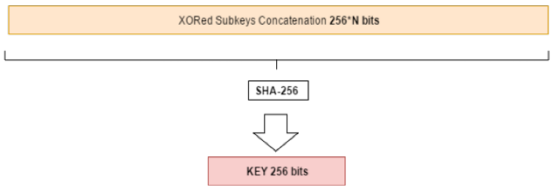


Fig. 6. Obtención de la clave final

Este módulo recibe como entradas las subclaves de los distintos *retos*, todos ellos con un tamaño de 256 bits (suficiente para evitar colisiones). Todas ellas se concatenan obteniendo una única cadena de 256*N bits. Finalmente, se obtiene la clave final de 256 bits haciendo el hash SHA-256 [11] de la cadena, como se puede ver en la figura 6.

D. *Elección de algoritmo de cifrado*

Se ha decidido implementar el algoritmo de cifrado simétrico AES utilizando el modo CTR ya que es el modo de cifrado que tiene mejor rendimiento, independientemente del tamaño del fichero [12][13]. Se utilizarán bloques de 128 bits y claves con tamaño de 256 bits que actualmente se consideran seguras [14]. Con este algoritmo, con un valor aleatorio y con la clave calculada anteriormente se cifran y descifran los ficheros confidenciales.

VII. CONCLUSIONES

Hoy en día la fuga de datos y el uso indebido de información se consideran una amenaza emergente para las organizaciones, especialmente cuando son llevadas a cabo por sus propios empleados. Además, las herramientas de DLP del mercado suelen enfocarse en evitar fugas de información desde atacantes externos y tratan a sus usuarios como si fueran absolutamente de confianza. La aplicación de cifrado basado en contexto en una herramienta DLP es un gran paso adelante para resolver este problema.

La arquitectura de DLP propuesta utiliza una serie de *retos* para obtener la clave de cifrado/descifrado a partir del contexto de entorno del dispositivo cliente que quiere utilizar datos sensibles. De esta forma se asegura que cada usuario solo pueda acceder a los datos críticos a los que tenga autorización dentro del contexto válido configurado por el administrador basado en las políticas de la empresa.

Estos *retos* pueden ser ejecutados de manera local en el dispositivo o de forma remota utilizando un servidor externo. La herramienta DLP integrada en el *Kernel* del dispositivo cliente se encarga de cifrar y descifrar los ficheros con claves generadas a partir de las subclaves obtenidas de la ejecución de los *retos*. Todo este proceso se lleva a cabo de manera totalmente transparente al usuario, el cual no tiene conocimiento del contexto de entorno válido. Además, el rendimiento de las aplicaciones de usuario se ve afectado mínimamente; por ejemplo la lectura de un fichero pasa de 20ms sin la herramienta a 150ms con la herramienta DLP, de los cuales 120ms se deben a la ejecución de los retos, que es independiente del tamaño del fichero. De esta forma, es posible desarrollar una herramienta DLP que realmente proteja a las organizaciones, no sólo de los atacantes externos, sino también de sus propios empleados.

Como trabajo futuro, quedan por realizar y definir más *retos* diferentes de los propuestos, como podrían ser *retos* cuya entrada de contexto sea multimedia, además de hacer estudios sobre la fortaleza de las contraseñas finales generadas por este método. Por último, se estudiará la posibilidad de controlar la seguridad en entornos donde el contexto de lectura y escritura sean distintos, mediante la realización de un conjunto de *retos* que compartan distintas propiedades matemáticas para obtener el mismo valor de subclave.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado con el apoyo del MINECO (proyecto DroneFS), con el código RTC-2015-4064-8 y del MINETUR (proyecto CiberNoid) con el código TSI-100200-2015-035.

REFERENCIAS

[1] Imad M. Abbadi, Muntaha Alawneh, “Preventing insider Information Leakage for Enterprises”, International Conference on Emerging Security Information, Systems and Technologies, pp. 27-31, 2011.

[2] Preeti Raman, Hilmi Güneş Kayacık, Anil Somayaji, “Understanding Data Leak Prevention”, Annual symposium on information assurance (ASIA), pp. 27-31, 2011.

[3] Asaf Shabtai, Yuval Elovici, Lior Rokach ”A Survey of Data Leakeage Detection and Prevention Solutions”, SpringerBriefs in Computer Science, 2012.

[4] Jungyub Lee, Sungmin Oh, Ju Wook Jang, “A Work in Progess: Context based Ecryption Scheme for Internet of Things”, Procedia Computer Science, v. 56, pp. 271-275, 2015.

[5] J. Al-Muhtadi, R. Hill, R. Campbell, M. D. Mickunas, “Context and location-aware encryption for pervasive computing environments”, Pervasive Computing and Communications Workshops, 2006.

[6] Vipul Goyal, Omkant Pandey, Amit Sahai, Brent Waters, “Attribute-based encryption for fine-grained access control of encrypted data”, Proceedings of the 13th acm conference on computer and communications security, pp. 89-98, 2006.

[7] Brent Waters, “Ciphertext-Policy Attribute-Bsed Encryption: An Expressive, Efficient, and Provably Secure Realization”, International Workshop on Public Key Cryptography, pp. 53-70, 2011.

[8] Z. Liu, Z. Cao, and D. S. Wong, “Efficient generation of linear secret sharing scheme matrices from threshold access trees”, IACR Cryptology ePrint Archive, 2010:374, 2010.

[9] Marina González, José J. García, Alberto García, Jorge Roncero, Víctor A Villagrá, “Propuesta de Sistema de Protección de la Información para Vehículos Aéreos no Tripulados”, II Jornadas Nacionales de Investigacion´ en Ciberseguridad (JNIC 2016), 2016, pp. 125–129.

[10] Phillip Rogaway, Thomas Shrimpton, “Cryptographic Hash-Function Basics: Definitions, Implications, and Separations for Preimage Resistance, Second-Preimage Resistance, and Collision Resistance
”, International Workshop on Fast Software Encryption, 2004, pp. 371–388.

[11] Harris E. Michail, George S. Athanasiou, George Theodoridis, Andreas Gregoriades, Costas E. Goutis, “Design and implementation of totally-self checking SHA-1 and SHA-256 hash functions’ architectures”, Microprocessors and MicroSystems, v. 45, Part B, pp. 227-240, 2016.

[12] Masram, Ranjeet, et al. "Analysis and comparison of symmetric key cryptographic algorithms based on various file features." *International Journal of Network Security & Its Applications* 6.4 (2014): 43.

[13] Altigani, Abdelrahman, Muawia Abdelmagid, and Bazara Barry. "Evaluating AES Performance Using NIST Recommended Block Cipher Modes of Operation." (2015).

[14] Paola Ceminari, Ariel Arelovich, Martín Di Federico, “Diseño de tres arquitecturas para un módulo criptográfico AES”, Biennial Congress of Argentina (ARGENCON), 2016 IEEE.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6450

Understanding the detection of view fraud in Video Content Portals

Miriam Marciel*‡, Ruben Cuevas*, Albert Banchs*†, Roberto González‡

Mohamed Amed‡, Stefano Traverso§, Arturo Azcorra*†

*Universidad Carlos III de Madrid, †Institute IMDEA Networks,‡NEC Europe, §Politecnico di Torino

The Interactive Advertisement Bureau (IAB) reported that online advertising generated revenue of $49B in 2014, in the U.S. alone. Of particular interest to this work is video advertising. Online video advertising is estimated to have generated $3.3B in 2014, in the U.S. alone. Given such revenues, it is no surprise that online advertising attracts fraud. Recent studies have estimated that 15-30% of ad-impressions were fraudulent. With respect to online video ads, the media and online advertising industry both report that fraud is endemic. In comparison to *click fraud* in search and display advertising, fraud in online video advertising has received little attention. Typically, the goal of click fraud is to inflate user activity counters at a particular target, such as a webpage. On-line video ads however offer new attack paths, and revenue streams. First, the status and earning from uploading popular online videos commonly attracts fraudulent activity, and online video portals put significant effort in regularly auditing their systems. For example, it was recently reported that YouTube removed more than 2B suspected ”fraudulent” views from accounts associated with the music industry. Second, in contrast to search and banner advertising, where advertisers can collect partial information on their users from clickbacks, online videos advertisers must delegate the detection and auditing of fraud to the portals that host their content. The common attack in online video ad fraud is to inflate the view counters of videos using botnets or crowd sourced users. In fact, it is easy to find paid services that generate tens of thousands of views to videos hosted on popular portals (e.g. YouTube, Dailymotion and Vimeo) at a low price. If the goal of the attacker is simply to increase the popularity and visibility of their videos, then this is enough. If however, the goal is to generate revenue, then the attacker attempts to have ads served to their fake viewers, and collects a share of the revenue. In response to the scale of the video-ad fraud, the media and online advertisers have consistently publicized the need for more effective anti-fraud solutions. Unfortunately, today we lack the tools to understand and independently audit the function and performance of fraud detection mechanism deployed by online portals.

In this paper, we first propose a measurement methodology to aid in filling this gap. Employing a modular active probe, we evaluate the performance of the fraud detection mechanism (for public and/or monetized views) of 5 online video portals, namely YouTube, Dailymotion, Vimeo, Myvideo.de, and TV UOL. Finding that YouTube is the only portal deploying a sufficiently discriminative view audit system, we deepen our analysis to study some of its key parameters. We focus on parameters that are directly accessible to users, and are reported to be manipulated by view-inflation bots in the wild on YouTube. We study the impact of manipulating the behavior of an IP address, such as varying the number of videos visited per day, the views per video, and the duration per view. We then look at the impact of changing the browser-profile of viewers, such whether or not cookies are enabled, and the impact of mixing viewer activity in NATed traffic. Our main findings are: (1) Of the 5 portals investigated, YouTube is the only portal to deploy a significantly discriminative view audit system in the public view counter. All other portals do not sufficiently discount their view counters even under the simplest fake views generation configurations. (2) A deeper analysis reveals that the detection mechanism of YouTube’s public view counter is susceptible to simple fake views generation strategies such as using multiple values in the HTTP connection attributes. (3) We find a consistent discrepancy between the reported views counted by by the public and monetized view counters in YouTube. We find that the monetized view counters report at least 75% more fake views than public view counters.

This work received funding from the EU H2020 (ReCRED (653417)), the Spanish Government (DRONEXT (TEC2014-54335-C4-2-R)), and the Government of Madrid (BRADE (P2013/ICE-2958)).

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6462

Arquitectura De Tiempo-Real para Sistemas Big-Data

Pablo Basanta-Val[1], Neil C. Audsley[2] , Ian Gray[4],
Norberto Fernández[5], Luis Sánchez-Fernández[6]
Departarmento de Ingeniería Telemática, Deparment of Computer Science, Centro Universitario de la Defensa

Universidad Carlos III de Madrid[1,6], Universidad de York[2-4], Universidade de Vigo[5]

pbasanta@it.uc3m.es[1], neil.audsley@cs.york.ac.uk[2], ian.gray@cs.york.ac.uk[4], norberto@cud.uvigo.es[5], luiss@it.uc3m.es[6]

Resumen. **Uno de los retos actuales en la ciencia de la computación se refiere a los sistemas de big-data que describen aplicaciones que no pueden ser soportados con las herramientas de procesado actuales. Estas herramientas resultan insuficientes para analizar, capturar, buscar, almacenar, transferir o visualizar esa gran cantidad de datos. Los sistemas big-data también se refieren a los algoritmos que realizan algún tipo de analítica que extrae información valiosa de esos datos, para encontrar nuevas correlaciones, buscar tendencias en negocios, o para combatir el crimen. Los sistemas big-data están caracterizados por la existencia de una ingente cantidad de información (proveniente de dispositivos móviles, sensores, cámaras, etc.) que necesita ser procesada para alcanzar una meta [1-4][6-7] de negocio.**

Por otro lado, los sistemas de tiempo real se refieren a sistemas sujetos a ciertas restricciones temporales, que se pueden definir de forma más primitiva con plazos temporales necesarios para procesar los eventos de una aplicación, hasta que generan su salida. Estos máximos plazos están derivados de las características del entorno externo de ejecución que impone requisitos estrictos sobre la aplicación. En los sistemas de tiempo real estos están usualmente en el orden de los milisegundos o microsegundos, pero también pueden ser plazos más largos, cuando interactúan con sistemas físicos humanos.

A la hora de caracterizar estas restricciones temporales se han de tener en cuenta formalismos de diferentes ámbitos (como por ejemplo lenguajes de programación [5], sistemas operativos, lenguajes de modelado) que definen mecanismos y algoritmos para la caracterización de una aplicación, estimando y acotando sus tiempos de respuesta.

Sin embargo, a día de hoy, no existe una teoría clara de cómo se debe enfocar las restricciones temporales en los sistemas de big-data, para formar sistemas big-data de tiempo-real. Existen [1-3] algunos trabajos pioneros que intentar explorar diferentes facetas y oportunidades derivadas de las sinergias de estos dos tipos de sistemas. En este contexto, este artículo define una arquitectura de tiempo real para sistemas big-data capaz de ejecutar analíticas de tiempo-real [2]. Estas analíticas están soportadas por la arquitectura que es capaz de ejecutar diferentes aplicaciones en un gran número de máquinas. Esta arquitectura ha sido implementada extendiendo Apache Spark y Apache Storm [1] añadiendo las técnicas necesarias para ofrecer rendimiento de tiempo real. Todos estos resultados son de gran interés para ambas comunidades, la de tiempo real y la de los sistemas big-data.

Palabras Clave- ***tiempo-real, big-data, arquitectura***

AGRADECIMIENTOS

Esta contribución está basada en el trabajo descrito en [2-3], financiado por Programa Jesus Castillejo (CAS14/00118), HERMES-SMARTDRIVER (TIN2013- 46801-C4-2-R), AUDACity (TIN2016-77158-C4-1-R), JUNIPER (FP7-IC6-318763),e-Madrid (S2013/ICE-2715) y por Grid'5000 (ver https://www.grid5000.fr).

REFERENCIAS

[1] N. Marz and J. Warren, Big Data: Principles and Best Practices of Scalable Real-Time Data Systems. London, U.K.: Manning Publications Co., 2015.

[2] Pablo Basanta-Val, Neil C. Audsley, Andy J. Wellings, Ian Gray, Norberto Fernández García: Architecting Time-Critical Big-Data Systems. IEEE Trans. Big Data 2(4): 310-324 (2016)

[3] Pablo Basanta-Val, Norberto Fernández García, Andy J. Wellings, Neil C. Audsley: Improving the predictability of distributed stream processors. Future Generation Comp. Syst. 52: 22-36 (2015)

[4] Zhihan lv et al. .Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. IEEE Transactions on Industrial Informatics Year: 2017, DOI: 10.1109/TII.2017.26502

[5] P. Basanta-Val, P., Anderson, J.S.: 'Using real-time Java in distributed systems: problems and solutions' inHiguera-Toledano, M.T., Wellings, A.J. (Eds): 'Distributed and embedded real-time Java systems' (Springer, February, 2012).

[6] P. Basanta-Val, N Fernandez-Garcia, L Sanchez-Fernandez, JA Fisteus: Patterns for Distributed Real-Time Stream Processing'. IEEE Transactions on Parallel and Distributed Systems DOI: 10.1109/TPDS.2017.2716929

[7] P. Basanta-Val, N. Fernández-García, L. Sánchez-Fernández: Predictable remote invocations for distributed stream processing, Future Generation Computer Systems, Available online 25 August 2017, ISSN 0167-739X, https://doi.org/10.1016/j.future.2017.08.02

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemáica (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6602

Sparse Intra-Flow Network Coding: comportamiento y modelado

Ramón Agüero
Dpto. Ingeniería de Comunicaciones
Universidad de Cantabria
ramon@tlmat.unican.es

Introducción

Network Coding (NC) fue propuesto originalmente por Ahlswede et al. in 2000. Cuestionaban el paradigma tradicional *store-and-forward*, promoviendo que los nodos intermedios tuvieran un papel más activo, procesando y combinando paquetes mientras que atraviesan la red. Desde entonces, son varios los trabajos que han analizado los beneficios que este novedoso enfoque podría aportar.

Inicialmente la idea era que los nodos intermedios mezclaran paquetes pertenecientes a diferentes flujos de tráfico, por lo que estas soluciones reciben el nombre de Inter-flow Network Coding. Una aproximación diferente, aunque no necesariamente excluyente, vino dada por el Intra-Flow Network Coding, en las que se combinan paquetes que pertenecían al mismo flujo, en las que se centraría esta propuesta.

Intra-flow Network Coding y RLNC

Como ya se ha dicho, en las soluciones Intra-flow Network Coding, se combinan (codifican) paquetes que pertenecen al mismo flujo de datos, tanto por parte de la fuente como de los nodos intermedios. Se puede decir, por tanto, que estas soluciones comparten algunas de las características de las técnicas de codificación fuente rateless, como los códigos LT y Raptor, que se han incorporado a varios sistemas de comunicaciones inalámbricas. Las principales ventajas que aporta NC serían: (i) codificación/decodificación *on-the-fly*, (ii) bajo retardo, y (iii) capacidad de recodificación.

Una de las primeras, y la más importante solución Intra-Flow NC es Random Linear Network Coding (RLNC). La información a transmitir se divide en bloques, y el nodo fuente transmite combinaciones lineales de paquetes que pertenecen al mismo bloque, hasta que el receptores capaz de decodificarlo. En ese momento, el transmisor podría pasar al siguiente bloque.

En RLNC, los coeficientes que se utilizan para construir las combinaciones lineas se eligen aleatoriamente de un cuerpo de Galois, de longitud $Q = 2^q$. Éste, junto con el tamaño del bloque, son los dos parámetros de configuración que tiene una clara influencia en el rendimiento de RLNC, tal y como se ha visto en varios estudios previos, en los que estudiaoms este comportamiento de manera exhaustiva, tanto sobre canales ideas como sobre enlaces con errores. Se utilizó una combinación del protocolo UDP y de RLNC para ofrecer un servicio de comunicación fiable, comparando su rendimiento con el mostrado por TCP, que es la solución más utilizada en este tipo de servicios. Los resultados obtenidos mediante simulación sobre la plataforma ns-3 se compararon con los valores proporcionados por un modelo teórico.

De RLNC a Sparse Network Coding

Una de las principales desventajas de RLNC, especialmente cuando $q > 1$ es que las operaciones de decodificación requieren la manipulación (operaciones algebraicas) de matrices de alta densidad. Así, la complejidad de RLNC podría crecer de manera considerable, limitando su aplicabilidad, especialmente en escenarios de aplicación concretos.

Para aliviar ese problema, se han propuesto diferentes alternativas que reducen la complejidad y la sobrecarga del RLNC, Algunas de las más relevantes son los códigos BATS o Fulcrum, aunque la charla se centrará en las soluciones Sparse Network Coding (SNC).

En SNC, cada transmisión se consigue con la combinación de un número pequeño de paquetes, en lugar de emplear todo el bloque (como sería el caso de RLNC). Se ha demostrado que así se consigue reducir de manera notable el número de operaciones (esto es, complejidad) en el decodificador. Sin embargo también podría generar un incremento considerable de la sobrecarga, ya que la probabilidad de generar paquetes linealmente dependientes (de aquellos previamente recibidos) podría ser elevada.

Se presentará un modelo novedoso, que puede utilizarse para comprender el compromiso entre la complejidad y la

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

sobrecarga, basado en una Cadena de Markov Absorbente. Una extensa campaña de simulación puso de manifiesto su gran precisión.

Afinando la densidad de SNC

Como ya se ha dicho anteriormente, el uso de SNC reduce fuertemente la complejidad de RLNC, pero con el coste de una mayor sobrecarga. Esto se puede aliviar permitiendo qe el transmisor modifique de manera dinámica (afine) la densidad (esto es, el número de paquetes que se combinan cada vez) a medida que la transmisión del bloque avanza.

Se han hecho varias propuestas para llevar a cabo dicha modificación. El modelo mencionado previamente puede aprovecharse para promover el uso de un esquema de modificación que mejore otras alternativas. Para ello hay que tener en cuenta el compromiso entre complejidad y rendimiento. Además, se tienen que tener en cuenta las limitaciones que los diferentes esquemas tienen al aplicarse sobre dispositivos reales.

Recodificación

Ya se ha dicho que uno de los elementos que distingue las soluciones Intra-Flow Network Coding de los esquemas de codificación fuente es la capacidad de los nodos intermedios de recodificar paquetes. Sin embargo, al utilizar SNC, aprovechar las posibilidades la recodificación es sensiblemente más complicado, ya que al recodificar se incrementa la densidad de los paquetes, lo que impide aprovechar las bondades de los esquemas de codificación de baja densidad. Hemos llevado a cabo análisis para caracterizar de manera detallada la densidad de los paquetes tras el proceso de recodificación, proponiendo soluciones heurísticas que permiten limitar el incremento de la densidad.

Referencias

[1] David Gómez, Sofiane Hassayoun, Arnaldo Herrero, Ramón Agüero, David Ros. *Impact of Network Coding on TCP Performance in Wireless Mesh Network*. Proceedings of PIMRC'12 doi:10.1109/PIMRC.2012.6362888

[2] David Gómez, Ramón Agüero, Marta García, David Ros. *TCP Acknowledgement Encapsulation in Coded Multi-hop Wireless Networks*. Proceedings of VTC'Spring 2014 doi:10.1109/VTCSpring.2014.7023118

[3] David Gómez, Eduardo Rodríguez, Ramón Agüero, Luis Muñoz. *Reliable communications over Wireless Mesh Networks with Inter and Intra-flow network coding*. Workshop on ns-3, 2014 doi:10.1145/2630777.2630781

[4] David Gómez, Eduardo Rodríguez, Ramón Agüero, Luis Muñoz. *Reliable communications over lossy wireless channels by means of the combination of UDP and Random Linear Coding*. Proceedings of ISCC'2014 doi:10.1109/ISCC.2014.6912516

[5] David Gómez, Eduardo Rodríguez, Pablo Garrido, Ramón Agüero, Luis Muñoz. *Enhanced Opportunistic Random Linear Source/Network Coding with Cross-Layer Techniques over Wireless Mesh Networks*. Proceedings of IFIP Wireless Days 2014 doi:10.1109/WD.2014.7020842

[6] Pablo Garrido, David Gómez, Ramón Agüero, Luis Muñoz. *Performance of Random Linear Coding Over Multiple Error-Prone Wireless Links*. IEEE Communication Letters, June 2015 doi:10.1109/LCOMM.2015.2421448

[7] Pablo Garrido, David Gómez, Ramón Agüero, Joan Serrat. *Combination of Intra-Flow Network Coding and Opportunistic Routing: Reliable Communications over Wireless Mesh Networks*. Proceedings of SIMUTOOLS'2015 doi:10.4108/eai.24-8-2015.2261115

[8] Pablo Garrido, David Gómez, Ramón Agüero, Joan Serrat. *Combination of Random Linear Coding and Cross-Layer Opportunistic Routing: Performance over Bursty Wireless Channels*. Proceedings of PIMRC'15 doi:10.1109/PIMRC.2015.7343571

[9] Pablo Garrido, David Gómez, Frank Fitzek, Ramón Agüero. *When TCP and Network Coding meet Wireless Links*. Proceedings of European Wireless'16

[10] Pablo Garrido, David Gómez, Jorge Lanza, Ramón Agüero. *Exploiting Sparse Coding: A Sliding Window Enhancement of a Random Linear Network Coding Scheme*. Proceedings of ICC'16 doi:10.1109/ICC.2016.7510783

[11] Pablo Garrido, Chres W. Sorensen, Daniel E. Lucani, Ramón Agüero. *Performance and Complexity of Tunable Sparse Network Coding with gradual growing functions over Wireless Networks*. Proceedings of PIMRC'16 doi:10.1109/PIMRC.2016.7794915

[12] Pablo Garrido, David Gómez, Jorge Lanza, Joan Serrat, Ramón Agüero. *Providing Reliable Services over Wireless Networks Using a Low Overhead Random Linear Coding Scheme*. Mobile Netw Appl (2016) doi:10.1007/s11036-016-0731-7

[13] Pablo Garrido, Daniel Lucani, Ramón Agüero. *How to Tune Sparse Network Coding over Wireless Links*. Proceedings of WCNC'17

[14] Pablo Garrido, Daniel Lucani, Ramón Agüero. *A Markov Chain Model for the Decoding Probability of Sparse Network Coding*. IEEE Transactions on Communications, 2017 doi:10.1109/TCOMM.2017.2657621

[15] Pablo Garrido, Daniel Lucani, Ramón Agüero. *Role of intermediate nodes in Sparse Network Coding: characterization and practical recoding*. Proceedings of European Wireless'17

Agradecimientos

Los autores agradecen la financiación del Gobierno de España (Ministerio de Economía y Competitividad, Fondo Europeo de Desarrollo Regional, FEDER) de este trabajo a través del proyecto ADVICE, Dynamic provisioning of connectivity in high density 5G wireless scenarios (TEC2015- 71329-C2-1-R).

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6938

Multimedia Services Distribution Using Adaptive and Cognitive SDNs

Jaime Lloret[1], Jesus Tomas[1], Oscar Romero[1], Jose Miguel Jimenez[1], Albert Rego[1], Belen Carro[2], Antonio Sánchez-Esguevillas[2], Manuel López-Martín[2], Santiago Egea[2]

[1]Instituto de Investigación para la Gestión Integrada de Zonas Costeras
[2]Dpto. TSyCeIT, ETSIT

[1]Universitat Politècnica de Valencia
[2]Universidad de Valladolid

[1]Camino Vera, s/n, 46022, Valencia, España
[2]Paseo de Belén 15, Valladolid 47011, España

jlloret@dcom.upv.es, jtomas@dcom.upv.es, oromero@dcom.upv.es, jojiher@dcom.upv.es, alremae@teleco.upv.es, belcar@tel.uva.es, antoniojavier.sanchez@uva.es, manuel.lopezm@uva.es, santiago.egea@alumnos.uva.es

I. INTRODUCTION

Current networks have much limitation due to their rigidity, which is given by static configurations mainly based on commands or static scripts [1-3]. The resource provisioning is less automatic and the efficiency decreases [4-5]. Moreover, virtualization and cloud are changing radically the traffic patterns of the data center. This is mainly due to the communication between servers, because the applications are split in many virtual machines that must communicate.

Software Defined Networks (SDNs) are able to divide the control plane from the data plane, which allow higher programmable, automatic and flexible networks [6-7]. In SDNs, we do not need to program node by node, but by a centralized manner through software that can be implemented independently of the manufacturer or the model (if they are supporting the same communication protocol). SDNs provide a more open network and allow accessing better to certain intelligent functions, which can contribute higher intelligence to the network operating [8]. These features make SDNs ideal to have a system that is able to adapt with the aim of having higher performance [9], mainly in multimedia delivery [10].

Cognitive networks use the information gathered from the network, such as observing traffic patterns for different network devices or the used protocols, the behavior of the users and servers, and the additional information that can be taken from the wireless networks (user movement, location, etc.) [11], in order to implement a series of procedures. In order to achieve this goal, artificial intelligence and automatic learning is over the available information. This will allow improving a specific objective and achieve higher system performance. In particular, we have applied several advanced machine learning techniques (Ensemble Methods, Decision Trees, Kernel Methods…) and new deep learning algorithms (Variational Autoencoders, Convolutional and Recurrent Neural Networks…) to specific problems of network traffic detection.

In this work, we designed and developed a network architecture and the communication protocol, that use the cognitive information taken from the data frames, the users and servers behavior, and the traffic patterns (traffic changes, quality of service parameters, state of the frames, behaviors in the values of the fields of the frames, etc.) with the aim of improving the multimedia delivery performance. The designed network is able to self adapt in each case. Network devices gather the data of the network parameters and patterns that are used by a smart network algorithm, which is included in a SDN module, take decisions based on the empirical data [12-13]. The cognitive adaptive software defined network can be implemented in a wide range of multimedia applications.

ACKNOWLEDGEMENTS

This work has been partially funded by the Ministerio de Economía y Competitividad del Gobierno de España and the Fondo de Desarrollo Regional (FEDER) within the project "Inteligencia distribuida para el control y adaptación de redes dinámicas definidas por software, Ref: TIN2014-57991-C3-2-P", and the Project "Distribucion inteligente de servicios multimedia utilizando redes cognitivas adaptativas definidas por software", Ref: TIN2014-57991-C3-1-P, in the Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Subprograma Estatal de Generación de Conocimiento.

REFERENCIAS

[1] J. Lloret, A. Canovas, J. Tomas, M. Atenas, A Network Management Algorithm and Protocol for Improving QoE in Mobile IPTV, Computer Communications, Vol. 35, Issue 15, Pp. 1855-1870, Sep. 2012.

[2] J. Lloret, Al. Canovas, J. J. P. C. Rodrigues, K. Lin, A Network Algorithm for 3D/2D IPTV Distribution using WiMAX and WLAN Technologies, Journal of Multimedia Tools and Applications. November 2011.

[3] J. Lloret, M. Garcia, M. Atenas, A. Canovas, A QoE management system to improve the IPTV network, Int. Journal of Communication Systems, Vol. 24 I. 1, Pp. 118-138, Abril 2010.

[4] A. Canovas, F. Boronat, C. Turro, J. Lloret, Multicast TV over WLAN in a University Campus Network, 5th Int. Conf. on Networking and Services, Valencia, April 20-25, 2009.

[5] A. Canovas, D. Bri, S. Sendra and J. Lloret, Vertical WLAN Handover Algorithm and Protocol to Improve the IPTV QoS of the End User, IEEE International Conference on Communications 2012, Ottawa (Canada), June 10–15, 2012.

[6] J. M. Jimenez, O. Romero, A. Rego, A. Dilendra, J. Lloret, Performance Study of a Software Defined Network Emulator, The Twelfth Advanced International Conference on Telecommunications (AICT 2016), May 22 - 26, 2016 - Valencia, Spain

[7] J. M. Jimenez, O. Romero, A. Rego, J. Lloret, Analyzing the Performance of Software Defined Networks vs Real Networks, International Journal On Advances in Networks and Services, volume 9, numbers 3 and 4, 2016, Pp. 107 - 116

[8] J. M. Jimenez, O. Romero, J. Lloret and J. R. Diaz, Energy Savings Consumption on Public Wireless Networks by SDN Management, Mobile Networks and Applications. 30 November 2016. DOI: 10.1007/s11036-016-0784-7

[9] Albert Rego, Sandra Sendra, Jose Miguel Jimenez, Jaime Lloret, OSPF routing protocol performance in Software Defined Networks, SDS 2017, May 8-11, 2017, Valencia (Spain).

[10] J. M. Jimenez, O. Romero, A. Rego, A. Dilendra, J. Lloret, Study of Multimedia Delivery over Software Defined Networks, Network protocols and Algorithms, Vol 7, No 4 (2015). Pp. 37-62.

[11] M. Garcia, S. Sendra, C. Turro, J. Lloret, User's Macro and Micro-mobility Study using WLANs in a University Campus, International Journal On Advances in Internet Technology, Vol. 4, Issue 1&2, Pp. 37-46 July 2011.

[12] S. Sendra, A. Rego, J. Lloret, J. M. Jimenez, Oscar Romero, Including Artificial Intelligence in a Routing Protocol Using Software Defined Networks, May 21-25, 2017, ICC 2017, Paris (France)

[13] M. Taha, L. Garcia , J. M. Jimenez, J. Lloret, SDN-based Throughput Allocation in Wireless Networks for Heterogeneous Adaptive Video Streaming Applications, IWCMC 2017, June 26-30 , 2017, Valencia (Spain)

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6951

On the Dimensionality Reduction of Markov Chains for Networks Modeling

V. Casares-Giner, L. Tello-Oquendo, V. Pla, J. Martínez-Bauset.
Instituto Universitario de Tecnologías de la Información y Comunicaciones
Universitat Politècnica de València
Camino de vera, s/n, 46022 Valencia. España.
vcasares@itaca.upv.es, luiteloq@upv.es, vpla@itaca.upv.es, jmartinez@itaca.upv.es,

***Resumen*—Tras los iniciales trabajos de Erlang y de Engset, las cadenas y los procesos de Markov se han utilizado para el modelado y análisis de los sistemas y redes de telecomunicación. A menudo, la enorme dimensión del espacio de estados impide un cómputo de los parámetros relevantes en tiempos razonablemente acotados. En muchos escenarios tales inconvenientes se han abordado mediante la utilización de técnicas de agregación-agrupación de estados. En esta presentación se aportan un número de ejemplos que ofrecen una visión metodológica y cronológica de algunas de las técnicas propuestas en la literatura especializada.**

***Palabras Clave*—Cadenas de Markov, procesos de Markov, teletráfico, espacio de estados, agregación.**

I. INTRODUCCIÓN

Tras las pioneras contribuciones de A. Markov (1906) de A. K. Erlang (1917) [1] y de T. O. Engest (1918) [2] las cadenas y procesos de Markov han sido ampliamente utilizadas como herramientas de modelado de diversos sistemas y redes de telecomunicación, tales como las redes de voz, de datos, de distribución de contenidos, y también en otras disciplinas de la ciencia. Muy a menudo las cadenas y procesos resultantes alcanzan dimensiones difícilmente manejables lo que supone inconveniencias en los desarrollos analíticos y de computación. Para paliar tales dificultades, tradicionalmente se ha recurrido a reducir el número de estados de la cadena/proceso de Markov mediante técnicas de agregación-agrupación de estados. Las agregaciones obedecen a factores diversos. Por ejemplo, si únicamente estamos interesados en una observación parcial del sistema, podemos "ocultar" los estados ajenos a la observación. En otras situaciones el modelo simultáneamente contiene varios procesos los cuales exhiben escalas temporales diferentes. Por ejemplo, tal es el caso en que el modelo capta tanto el nivel de sesión como el de ráfaga como el de paquetes de datos. Mientras la duración de una sesión o llamada es del orden de minutos, el tiempo de transmisión de un paquete de datos es del orden de mili-segundos - dependiendo del medio de transmisión. En estos casos, si por ejemplo la observación deseada es a nivel de sesiones, resulta conveniente "enmascarar" el resto de niveles. A título ilustrativo, en la siguiente sección se describen dos primeros ejemplos clásicos del uso de técnicas de agregación, y a continuación otros dos ejemplos elaborados por los autores de esta comunicación y que recientemente fueron publicados.

II. ALGUNOS EJEMPLOS DE REDUCCIÓN DE LA DIMENSIONALIDAD

A. *Tráfico de desbordamiento*

Los trabajos de Erlang y de Engset básicamente se focalizan en el modelado del servicio telefónico, considerándose como la semilla del teletráfico. En 1956, Wilkinson [3] estudió la caracterización del tráfico en un grupo secundario que se ocupa por desbordamiento de un grupo primario de tamaño M circuitos al cual se le ofrece tráfico de Poisson y servicio exponencial. En [3], Wilkinson aproxima la variable aleatoria del número de circuitos ocupados en el grupo secundario con una distribución exponencial negativa. Ya en 1973, y como alternativa a la teoría del azar equivalente de Wilkinson, A. Kuczura [5] propuso el proceso IPP (*Interrupted Poisson Process*) como aproximación del tráfico de desbordamiento anteriormente descrito. El proceso IPP es un proceso *on-off*; en el estado *on*, respectivamente en estado *off*, la tasa de generación de eventos (llamadas) es $\lambda_{on} > 0$, respectivamente $\lambda_{off} = 0$. A diferencia de la aproximación de Wilkinson, en la de Kuczura tres son los parámetros que definen un proceso IPP, la tasa λ_{on} y las dos tasas de transición entre ambos estados, lo que supone un grado de libertad adicional, por lo que cabe alcanzar una mejor aproximación al proceso de desbordamiento.

Sea (i, j) el estado que define el proceso de Markov (i (j) identifica el número de llamadas en curso en el grupo primario (secundario)), por lo que $0 \leq i \leq M$ y $0 \leq j < \infty$. En ambas aproximaciones subyace la técnica de agrupamiento de estados consistente en "ocultar" el

número de circuitos ocupados en el grupo de primera elección. En particular, para un valor dado de $j = J$ se agrupan los $M + 1$ estados (i, J) $i = 0, 1, \ldots, M$ en un único estado (Wilkinson) o en dos estados (Kuczura). En ELprimer caso, la definición de las nuevas tasas de transición se sustenta en la identificación de los dos primeros momentos (media y varianza) del número de circuitos ocupados de segunda elección (ver apéndice de [3] elaborado por J. Riordan). En el caso de Kuczura se utliza el tercer momento de la v.a. indicada, si bien caben otras aproximaciones.

B. Sistema de colas con prioridades

Sea un sistema con dos servidores al que se le ofrece dos tráficos de Poisson, de tasas λ_H y λ_L y servicio exponencial de tasas, respectivamente μ_H y μ_L. El tráfico H es prioritario sin expulsión con respecto al trafico L. Esto es, cuando una petición se genera, ésta será inmediatamente atendida si alguno de los dos servidores está libre, ocupándose sin interrupción alguna hasta la finalización de su servicio. Si por el contrario los dos servidores están ocupados, la petición esperará en una cola (*buffer*) de capacidad infinita. No se contemplan abandonos por impaciencia durante la espera. Tras finalizar un servicio el servidor liberado será ocupado por una petición del tipo H, si la hubiese en cola; caso de no haberla el servidor será ocupado por una unidad del tipo L, si la hubiese en cola, y caso de no haberla el servidor quedará disponible para la siguiente petición.

Se trata de un sistema $M/M/2$ (notación de Kendall), al cual se le ofrece dos flujos de Poisson con prioridad sin expulsión del flujo H con respecto al flujo L. El proceso queda definido por dos v.a., i y j, y que respectivamente indican el número de unidades del tipo H y del tipo L. La dimensión del espacio es $0 \leq i, j < \infty$, y su resolución entraña bastante complejidad la cual puede reducirse al utilizar técnicas de agrupación. En particular, para un $I \geq 2$ cualquiera y fijado un valor de $j = J$, el colectivo de infinitos estados (i, J) con $I \leq i < \infty$ puede agruparse y reemplazarse por un único estado $(\underline{I}, J)$ con las siguientes tasas de transición salientes (las entrantes quedan sin modificarse). Una primera transición saliente hacia el estado $(\underline{I}, J+1)$ de tasa λ_L y una segunda tasa de transición saliente hacia el estado $(I-1, J)$ que denotamos por $\mu_{bp_{2\mu_H}}$. En buena lógica $\mu_{bp_{2\mu_H}}$ se iguala a la inversa del tiempo medio de la duración del período cargado (*busy period*) de una cola $M/M/1$ con tasas de llegada y de servicio, respectivamente λ_H y $2\mu_H$ [4]. El proceso resultante ya puede resolverse de forma más directa al tratarse de un proceso QBD (*Quasi Birth Death)* [9].

Obviamente el anterior procedimiento puede generalizarse al reemplazar la agrupación de infinitos estados indicada por una distribución PH (*PHase type)* [9] con un número de estados mayor que la unidad. En tal caso, las tasas de transición entre los estados transitorios de la distribución PH y las asociadas a las transiciones hacia estados absorbentes se obtendrían mediante algún procedimiento de ajuste, por ejemplo con el algoritmo EM (*Expectation-Maximization*) [8].

C. Sistema celular con macro células y pico células

Células de diverso tamaño, macro, micro, nano, pico, y femto células dan cobertura en los actuales sistemas radio celulares. Sin entrar en los pormenores de cada tipo de célula, describimos aquí un análisis de teletráfico correspondiente a un escenario de macro y pico células, en donde el segundo tipo alberga las micro, nano, pico, y femto células. Bajo la cobertura de una macro célula existen varias (muchas) pico células. La cobertura de cualquier pico-célula se solapa al 100 por 100 en una única macro-célula y no existen áreas de solape común entre dos o más pico-células. La residencia en macro y pico células sigue un modelo exponencial de tasas, respectivamente μ_{rm} y μ_{rp}. El tráfico de llamadas o sesiones se genera siguiendo la ley de Poisson, de tasa λ y con servicio exponencial de tasa μ. Para una descripción más detallada del escenario en estudio ver [12].

La asignación de recursos es como sigue. Tras la generación de una llamada por parte de un móvil, si éste percibe doble cobertura (de su propia macro-célula y de una pico-célula) la llamada se ofrecerá en primera opción a la pico-célula siendo admitida siempre y cuando haya canales libres en la pico-célula; en caso contrario será ofrecida a su macro-célula (*directed retry*, [6]). Si el móvil únicamente tiene cobertura simple, de su propia macro-célula, la llamada se ofrecerá a su macro-célula. Una llamada que se ofrezca a una macro-célula bien en primera opción (cobertura simple) bien en segunda opción (cobertura doble) será admitida siempre y cuando haya canales libres en la macro, caso contrario la llamada será bloqueada (modelo de pérdidas). Para llamadas en curso los traspasos o *handovers* se gestionan según se indica. Supongamos un móvil que reside en una macro (m) y su llamada es atendida por la macro (m). Diremos que su estado es $(R, S) = (m, m)$ (R =residencia, S = servicio). Cuando el móvil visite una pico-célula con algún canal libre o disponible, el servicio en curso se transferirá al mencionado canal y su estado evolucionará a $(R, S) = (p, p)$; y en caso contrario el servicio seguirá siendo dado por la macro-célula, y el estado de la llamada pasará a ser $(R, S) = (p, m)$. Mientras el móvil se encuentre en el estado (p, m) si se libera un canal de la pico-célula el sistema ejecutará un traspaso dirigido (*directed handover*) [7] y el estado del móvil evolucionará a (p, p). Un móvil que se halle en el estado (p, m) y abandona la residencia de la pico-célula su estado evolucionará a $(R, S) = (m, m)$. Cuando un móvil se encuentre en el estado (p, p) y salga de la cobertura de la pico-célula, entonces se gestionará un traspaso de su llamada hacia un canal libre de su macro-célula, en caso exitoso el estado evolucionará a (m, m), caso contrario se producirá una terminación forzosa. Finalmente cuando un móvil en el estado (m, m) visite una macro-célula vecina su estado evolucionará a $(R, S) = (m, m)$ si hay canal alguno disponible en la nueva macro-célula; caso contrario se producirá una terminación forzosa.

Como parámetros de prestaciones básicos tenemos la probabilidad de bloqueo de las llamadas nuevas, la proba-

bilidad de un *directed retry*, la probabilidad de un *directed handover* y la probabilidad de terminación forzosa, entre otros. La evaluación de los mismos a partir del modelo Markoviano requiere de una enorme esfuerzo de cómputo con altísimo consumo de tiempo, pues en escenarios con pocas macro y pico células el número de estados del proceso Markoviano resultante es enorme. En [12] se estudia el sistema radio celular descrito utilizando técnicas de agrupamiento de estados que permiten reducir su cardinal y en consecuencia calcular varios de los parámetros indicados con resultados muy próximos a los obtenidos por simulación.

D. Red de servicios integrados

En [11] se describe el siguiente sistema de servicios integrados. Una única estación base ofrece servicio a tráfico real (RT) y tráfico no real (NRT). Un enlace de capacidad total igual a C Mbps se comparte entre las comunicaciones, llamadas o sesiones, RT y NRT. El tráfico RT, tiene estricta prioridad sobre el tráfico NRT. Denotemos por N_{rt} el máximo número de canales que el tráfico RT puede ocupar. Cuando una llamada RT se genera ocupa un canal disponible con tasa c; en caso de no disponibilidad la llamada se bloquea. Cabe observar que una llamada RT que se admite ocupa 1 canal durante el servicio requerido. N_{rt} es tal que satisface $N_{rt} \times c << C$. c es lo suficientemente pequeño tal que permita que el tráfico NRT pueda ocupar recursos de manera razonable. La capacidad no utilizada por el tráfico RT se reparte equitativamente ente los flujos NRT acorde con la disciplina PS (*Processor Sharing*). Los flujos RT y NRT se caracterizan por llegadas de Poisson con tasas, respectivamente λ_{rt} y λ_{nrt}. El servicio requerido por cada petición RT admitida es exponencial de tasa μ_{rt}. El tamaño del flujo generado por las sesiones NRT también se rigen por una ley exponencial de valor medio L (bits) y su tasa correspondiente μ_{nrt} se indica más adelante dado que el servicio de un flujo NRT depende de los recursos disponibles.

La v.a. número de llamadas en progreso en la red del tipo RT y la v.a. para las del tipo NRT definen un proceso de Markov. Identifiquemos un estado genérico por (i, j), identificando i y j las llamadas RT y NRT, respectivamente. Cuando el sistema alcanza el estado (i, j) la capacidad disponible para el tráfico NRT viene dado por $C_{nrt}(i) = C - i.c$. La tasa binaria asignada a cada flujo NRT admitido resulta ser $c_{nrt} = C_{nrt}/j$ y se actualiza cada vez que hay una llegada admitida o se produce una salida, bien sea del tipo RT como NRT. Es claro que la tasa de servicio de un flujo NRT, considerado como elástico, viene dado por $\mu_{nrt} = C_{nrt}(i)/L$ (llamadas por segundo).

Con el fin de satisfacer el QoS del flujo NRT admitido, el máximo número de llamadas (flujos) NTR se limita a N_{nrt} por lo que un flujo NRT se bloqueará si en el instante de llegada tenemos que $j = N_{nrt}$. Así pues, el conjunto factible de estados es tal que $0 \leq i \leq N_{rt}$, $0 \leq i + j \leq N_{rt} + N_{nrt}$ y la cardinalidad es igual a $(N_{rt} + 1)(N_{nrt} + 1)$. En [11] se identifica el índice i como nivel y el índice j como fase puesto que el proceso resultante es un QBD (*Quasy Birth-Death*). Basado en la separación de escalas temporales entre niveles y fases, y para resolver los parámetros más relevantes del sistema descrito, en [11] se propone un nuevo algoritmo iterativo denominado AMCA (*Absorbing Markov Chains Approximation*). En el algoritmo se analiza el régimen transitorio de cada nivel y se determina la fracción de tiempo que el sistema está visitando cada una de sus fases hasta que un cambio de nivel ocurre. Una vez se han obtenido las fracciones de tiempo para todas las fases de cada nivel, se computa una nueva distribución estacionaria del sistema completo. El procedimiento se repite hasta que una predeterminada exactitud se alcanza. Para iniciar el mismo se utiliza alguna de las soluciones obtenidas en algoritmos previos como el QSA (*Quasi-stationary approximation*) [10].

III. CONCLUSIONES

Se han presentado varios ejemplos ilustrativos de las técnicas de reducción de la dimensión del número de estados de una cadena y/o proceso de Markov. En un futuro inmediato se piensa extender el trabajo a otros casos que en la literatura abierta han merecido notable atención.

AGRADECIMIENTOS

Trabajo desarrollado en el entorno temático de los proyectos nacionales TIN2013-47272-C2-1-R y TEC2015-71932-REDT.

REFERENCIAS

[1] A. K. Erlang, “The solution of some problems in the theory of probabilities of significance in automatic telephone exchanges”, First published in Electroteknikeren, Vol. 13, p. 5, 1917.

[2] T.O. Engset, “Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern. Elektrotechnische Zeitschrift”, 1918, Heft 31. Translated to English in Telektronikk (Norwegian), June 1991.

[3] R. I. Wilkinson, “Theories fo toll traffic engineering in the USA”, Bell System Tehcnical Journal, Vol. 35, n. 2, pp. 421-514, 1956.

[4] L. Tackas, “Introduction to the theory of queues”, Oxford Univ. Press. New York, 1962.

[5] A. Kuczura, “The interrupted Poisson process as an overflow process”, Bell System Tehcnical Journal, Vol. 52, n. 3 , pp. 437-448, 1973.

[6] B. Eklundh, “Channel utilization and blocking probability in a cellular mobile telephone system with directed retry”. IEEE Transactions on Communications, Vol. 34, n. 4 , pp. 329-337, 1986.

[7] J. Karlsson, B. Eklundh, “A cellular telephone system with load sharing - An enhancement of directed retry”, IEEE Transactions on Communications, Vol. 37, n. 5, pp. 530-535, 1989.

[8] S. Asmussen, O. Nerman, M. Olsson “Fitting phase-type distributions via the EM algorithm”, Scandinavian Journal of Statistics, Vol. 23, No. 4, pp. 419-441, December 1996.

[9] G. Latouche, V. Ramaswami, “Introduction to matrix analytic methods in stochastic modeling”, ASA-SIAM, 1999 .

[10] J. Martinez-Bauset , V. Pla , J. Vidal , L. Guijarro , “Approximate analysis of cognitive radio systems using time-scale separation and its accuracy, IEEE Commun. Lett. Vol. 17, n. 1 pp. 35-38, 2013.

[11] L. Tello-Oquendo, Vicent Pla, J. Martinez-Bauset , V. Casares-Giner, “Performance analysis of wireless networks based on time scale separation: A new iterative method”, Computer Communications, Vol. 86, pp. 40-48, July 2016.

[12] V. Casares-Giner, J. Martinez-Bauset, X. Ge, “Performance model for two-tier mobile wireless networks with macrocells and small cells”, Wireless Networks. Published on line: 26 november 2016, DOI 10.1007/s11276-016-1407-8.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6494

Intelligent Traffic Light Management using Multi-Behavioral Agents

Luis Cruz-Piris, Diego Rivera, Ivan Marsa-Maestre, Enrique de la Hoz y Susel Fernandez
Departamento de Automática,
Universidad de Alcalá
Edificio Politécnico, 28805 Alcalá de Henares (Madrid), Spain
{luis.cruz, diego.rivera, ivan.marsa, enrique.delahoz, susel.fernandez}@uah.es

***Abstract*—One of the biggest challenges in modern societies is to solve vehicular traffic problems. In this scenario, our proposal is to use a Multi-Agent Systems (MAS) composed of three types of agent: traffic light management agents, traffic jam detection agents, and agents that control the traffic lights at an intersection. This third type of agent is able to change its behaviour between what we have called a *selfish* mode (the agent will try to influence the other neighbour agents of its type to achieve its goal) or an *altruistic* mode (the agent will take into consideration the other neighbour *selfish* agents indications). To validate our solution, we have developed a MAS emulator which communicates with the *Simulation of Urban MObility* (SUMO) traffic simulator using the Traci tool to realize the experiments in a realistic environment. The obtained results show that our proposal is able to improve other existing solutions such as conventional traffic light management systems (static or dynamic) in terms of reduction of vehicle trip duration.**

***Keywords*—Multi-Agents System, Intelligent Transportation System, Smart Cities, Sensor Networks and Traffic Simulations**

I. INTRODUCTION

Intelligent Transportation Systems (ITS) have experimented a fast improvement over the last years thanks to the evolution in the technologies that they rely on. Specifically, subsystems such as Advanced Traffic Management Systems (ATMS) have benefited from the rise of technologies like those used in devices derived for the Internet of Things (IoT) paradigm. These technologies have allowed to increase the data volume used to make decisions.

A side effect of the evolution of the ATMS in the last years has become a great challenge in modern systems: how to process such a high information volume so that the decision-making process is correct and efficient.

Our proposal consists of defining a MAS for intersection management and coordination between intersections about traffic lights management. This kind of system has shown its utility to solve problems within distributed environments.

In section II we review existent systems that modify traffic light phases in order to solve traffic congestion problems, and we discuss why MAS have been proven as an effective tool in this topic. In our case, the main goal of the system will be to control the traffic lights scheduling in a road network, so they can adapt it to the environment. Three types of agent can be distinguished in the system. The first type will manage the variation of the duration of phases in a traffic light, guaranteeing its correct operation. The second type will be in some of the elements of the traffic scenario (i.e. vehicles or sensors in strategic points of the vial network), and will use the information of their environment to make decisions about whether there is a possible congestion situation. Finally, the third type of agent will be located at intersections and will make decisions about the variation of traffic lights phases using the information generated by the other agents. This agent will be responsible for notifying the changes on the duration of phases of all the traffic lights in a single intersection. This third type of agent will vary its own operation per the congestion degree of the intersection where it is located. For low traffic loads, it will work in what we call *altruistic* or *collaborative* mode, in which the agent will help other agents of the same type in different intersections to achieve their goal. However, when high traffic loads are detected, the agent will work in a *selfish* or *isolated* mode, in which the agent will only react to reports from the agents located in the same intersection.

We have modelled and developed the agents described before to validate the solution and emulate the MAS in a simulation environment. We have connected this development with the widely recognized traffic simulator SUMO [1]. Furthermore, we have selected a portion of the well-known traffic scenario “TAPAS Cologne” (Travel And Activity PAtterns Simulation Cologne) [2], [3] for the validation experiments. The portion we have selected reproduces the traffic in a portion of the map of the city of Cologne (Germany), as shown in Figure 1, for a whole

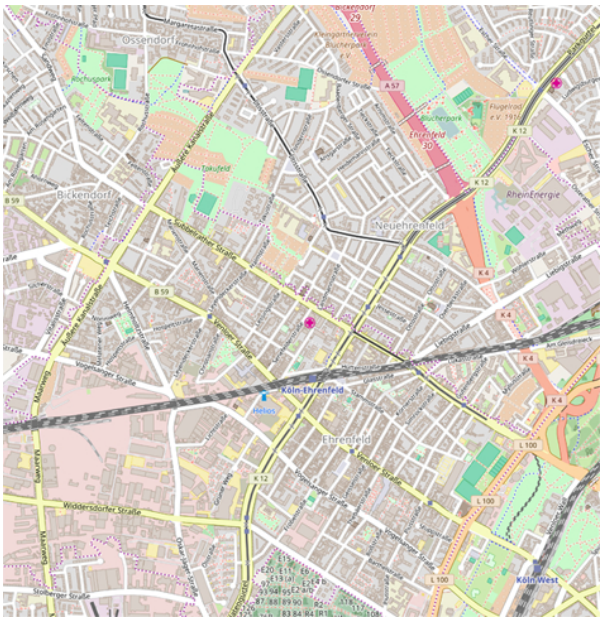


Fig. 1: OpenStreetMap capture of the portion of the scenario used in the system validation.

day.

One of the biggest challenges in the traffic management systems is being able to process all the information that is produced in their environment and make effective decisions using it. The main advantage of using a MAS to solve this kind of scenarios is the possibility to divide the problem in smaller sub-problems, and therefore improve scalability and efficiency of the system. This paper contributes to this goal in the following ways:

- By defining the agents and mechanisms needed to develop the proposed system, applying different operation modes to them depending on the circumstances of the environment (section III).
- By developing a MAS emulator, capable of interacting with the traffic simulator (section IV).
- By validating the proposed solution and evaluating it (section V). We have used a scenario based on real-life data, giving more validity to the obtained results. In addition, the data-set used is open source, allowing reproducibility of the results.

Finally, in sections B and VI we discuss this results and we expose the conclusions and future work lines derived from our work.

II. RELATED WORK

There are many factors involved in the correct circulation of vehicles. In fact, problems related to vehicle traffic are widely studied, as it is one of the most relevant challenges in modern societies.

The intelligent traffic management systems aim to have a global overview of the problem so they can make the right decisions in each case. We will focus here specifically in the solutions that are based on the optimization (static or dynamic) of the traffic light behaviour, as it is the goal that our proposed system pursues too.

One way of addressing the problem is to improve the traffic light scheduling, allowing a correct traffic light synchronization while trying to optimize the time that each vehicle should wait at the traffic light. It is possible to find many works where evolutionary algorithms (EA) are used to solve this problem. [4] investigate the potential of EA for the optimization of traffic light controllers using a Genetic Algorithm (GA), one of the most popular algorithms of this category. They conclude that to solve this type of problems it is useful to use evolution strategies. In [5] authors show the use of an iterative optimization algorithm (specifically a Particle Swarm Optimization (PSO) algorithm) to find successful cycle programs of traffic lights. They validate their proposal using the SUMO microscopic traffic simulator, obtaining an improvement in terms of total trip times and number of vehicles that arrive at their destination in a predefined simulation time.

[6] present a review on agent-based technology for traffic. The authors approach several topics and they have grouped them in two main categories: modelling and simulation, and control and management.

If we briefly analyse some of the challenges of traffic related problems, it is safe to say that it is geographically distributed, the environment is dynamic, and there is a strong interaction between the elements that compose them. [7] remark that the challenges that raise these properties can be well addressed with agent-based technology.

Therefore, we are going to focus in applications of agent technology used to improve traffic related problems. The work of [8] shows a review about this kind of solutions, where we are going to pay special attention to those which interact with traffic lights.

There are other Intelligent agent-based urban signal control systems like the ones presented by [9] or [10]. The first paper defines in a generic way, the necessary elements that should compose this kind of systems. The second one proposes a hierarchical multiagent architecture defined with some depth, including a description of the internal operation of the system. The most remarkable proposal in both works is the demonstration of the viability of the usage of MAS applied to intelligent traffic light management. Another approach to Intelligent agent-based transportation systems is shown in the work of [11], where an ontology-driven architecture is defined aiming to improve the driving environment through a traffic sensor network. This paper presents satisfactory results, but has only been validated in a small simple scenario.

ACTAM (Adaptive and Cooperative Traffic light Agent Model) [12] aims to reduce traffic congestion in urban roads. It proposes a complete agent system, able to synchronize and improve the efficiency in traffic light management in cities. It shows the improvement of using multiagent systems instead of a static traffic light scheduling. Unfortunately, it has only been validated in a small traffic network with about 30 intersections, which is far from a real-life scenario. Furthermore, the results are only compared with fixed-time signal control, and are not compared with other available management systems like

actuated control traffic.

III. PROPOSED SYSTEM

The main goal of our proposal is to reduce the duration of vehicle trips. To achieve this goal, we will modify the intervals of the traffic lights (phases) in a traffic network using a MAS. In the following sections, we are going to describe the types of agents involved in the system and the general operation of the system itself.

A. Agents

The development of our MAS is based on the definition of three main types of agents where each agent will be defined to perform a set of specific tasks.

1) "TLAgent" Traffic Light Management Agent: These agents are located at each traffic light. Their goal is to manage the light changes. At system initialization, there are predefined static phases to change the states of the traffic light that the agent manages. This type of agent is also aware of the existence of the rest of agents of the same type located in the same intersection.

If the agent does not receive any message indicating otherwise, the lights will be set following the static predefined preferences. In case that another agent sends a request asking to change the duration of its phases, the agent will store and use the new phase duration if the new duration is between two predefined thresholds, and the rest of agents of the same type have confirmed the phase variation.

2) "TJamAgent" Traffic Jam Detection Agent: These agents are located both in vehicles and in sensors installed across the map. They receive information from the environment and decide, per their preferences, if they need to communicate their decision about the state of the environment (if the agent considers that there is a traffic jam or not). Agents located in vehicles know their geographic position, and their instantaneous velocity. If the obtained values for the velocity are below the threshold value defined in preferences during an established period, the agent will communicate its situation.

When this type of agent is in a sensor (induction loops, video-vehicle detections, etc.), the operation is similar. The only variation consists on the way it obtains the information to determine possible traffic jams, that will depend on the sensor.

3) "IntersectionAgent": These agents are in the communications system of the traffic lights of a certain intersection. They receive the information from the nearest TJamAgents and decide whether they should change the phases of the traffic lights to prioritize the traffic flow in one of the ways of the intersection.

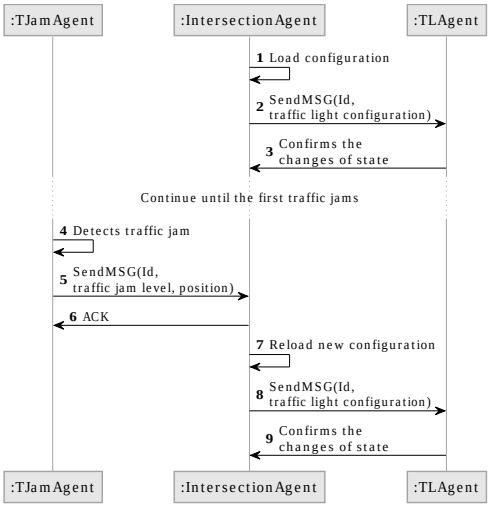
These agents can manage from 2 to N traffic lights, depending on the number of lanes that end in the intersection. This allows to reduce the amount of data used in the decision-making process by grouping the traffic lights in those which have the same state (red light or green light), as a variation in one group will affect the other.

When the system is being initialized, the IntersectionAgent will request the TLAgents that are under its management to send reports about the current phases of their traffic lights. Using this information, it will create a state machine that will be used later to perform the needed changes.

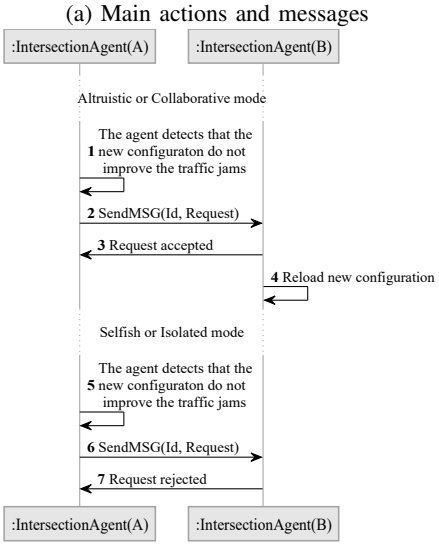
These agents have two operating modes:

- **Altruistic or Collaborative**: Additionally to the information received by the agent from TJamAgents, the agent accepts requests from other IntersectionAgents for the prioritization of some traffic flows.
- **Selfish or Isolated**: In this mode, the agent determines that the map zone under its management is congested enough not to cooperate with other agents and make its own decisions.

By default, every IntersectionAgent starts operating in *altruistic* mode. To avoid a possible system block, every time the agent starts operating in *selfish* mode, a timer is started, so the time an agent can operate in that mode is always limited. This behavior, together with the main actions and exchange of messages that the agents perform, is shown in Figure 2.



(a) Main actions and messages



(b) Altruistic vs. Selfish mode

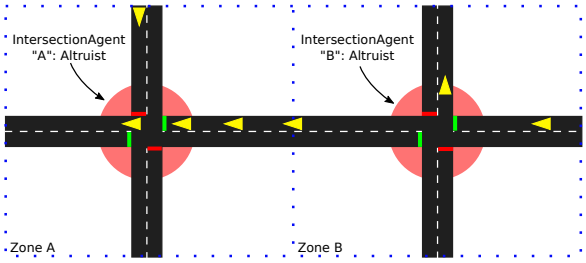
Fig. 2: Communication between agents.

B. Agent behaviour

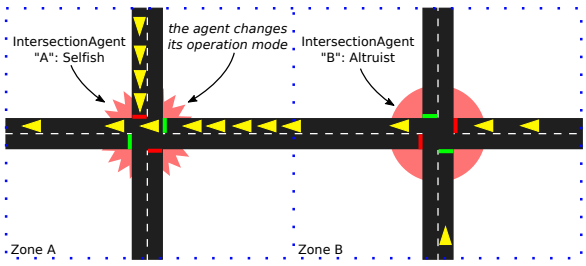
The operation of the system is based in the existence of a set of agents able to obtain information from the environment, and to decide when there is a congestion situation, agents able to change the traffic light states, and agents able to manage the external and internal synchronization of the phases of each traffic light in an intersection.

In this section, we are going to show the operation of the proposed MAS by using a basic use case formed by the two intersections shown in Figure 3. The different elements that compose the system are represented in the following way: The traffic lights (TLAgents) are represented by red lines or green lines, depending on their state, the yellow triangles represent the vehicles moving in each moment (TJamAgents), and the IntersectionAgents appear as red areas.

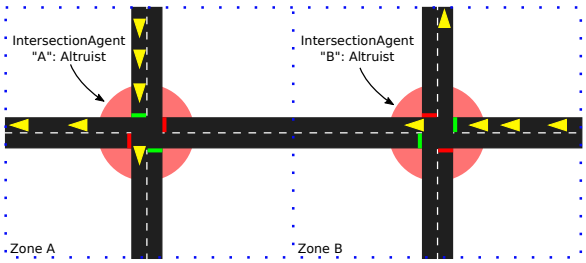
When we talk about horizontal traffic flow we are referring to the flow generated by vehicles going from the right to the left of the figures. On the other hand, vertical traffic flows are those generated from the top to the bottom of the figures and vice versa.



(a) Initial state



(b) Zone "A" Congested



(c) Zone "A" normal flow

Fig. 3: MAS operation use case.

1) **Initial state** (Figure 3a): The IntersectionAgent "A" and the IntersectionAgent "B" start in the *altruistic* operation mode, so they will collaborate to improve the traffic flow where the traffic volume is higher.
   - TJamAgents that are in Zone A have a velocity below the maximum velocity value due to the high traffic volume. Each agent decides individually that it must report this situation to the IntersectionAgent "A".
   - IntersectionAgent "A", using the received information, decides to prioritize the horizontal traffic flow and makes two requests:
     - Reports to the TLAgents of its intersection that they must increase the duration of the traffic lights green state of their phases.
     - Reports to the IntersectionAgent "B" this change so it can adjust the synchronization of the phases of the traffic light it manages.
2) **Zone "A" Congested**: IntersectionAgent "A" changes its operation mode to the *selfish* mode. In Figure 3b it is shown traffic jams both in the vertical traffic flow of the Zone "A" and the horizontal traffic flow. IntersectionAgent "A" stops collaborating with the rest of IntersectionAgents, performing the following actions:
   - Reports to the IntersectionAgent "B" that it must limit the horizontal traffic flow. Given that the IntersectionAgent "B" is still operating in *altruistic* operation mode, it will prioritize this request over the reports received from the TJamAgents. Therefore, it will request his TLAgents that they must reduce the duration of the traffic lights green state of their phases in the horizontal traffic flow.
   - Reports to the TLAgents that they should change the traffic lights phases so red and green states are of the same duration (for vertical and horizontal flows).
3) **Zone "A" normal flow**; After the actions of the previous step, The Zone "A" returns to normal traffic flow values, and therefore, the IntersectionAgent "A" changes its operation mode to the *altruistic* mode. Figure 3c shows how the horizontal traffic flow has been reduced. Each IntersectionAgent manages again the duration of its traffic lights using the information received from the TJamAgents and other IntersectionAgents.

IV. IMPLEMENTATION

Once the proposal has been described, it is important to validate it using an implementation where it can be confronted with a realistic situation. In our case, we have developed a simulation platform composed basically by two modules: Traffic simulation module (SUMO traffic simulator + TAPASCologne simulation scenario) and MAS emulation module implemented using Python.

The communication between both modules is performed using the TraCI (*Traffic Control Interface*) tool [13], included in the SUMO package. This tool provides a TCP-based client/server architecture that allows to control

and modify the SUMO simulations through an external application.

A. Simulation scenario

Using a realistic simulation scenario is essential to perform the validation of the system, as it will provide conditions like the real-life scenarios.

In our proposal we have chosen the scenario called "TAPAS Cologne" that is referenced in the SUMO documentation. It is a complete simulation scenario of the German city of Cologne. It was created by the Institute of Transportation Systems at the German Aerospace Center (TIS-DLR), and its goal is to reproduce, with the maximum possible realism, the urban traffic of Cologne. It defines a map of 400 km^2 and 24 hours of traffic.

The original simulation scenario is composed by a road network with 68642 edges, 30354 nodes and 1547333 routes. The size of this scenario causes very high simulation times. Therefore, for the validation of our proposal, we have decided to crop the scenario in a smaller portion that, still being representative of the original scenario, will yield lower simulation times. The solution proposed by [14] is tested on a simulated network of the Lower Downtown Toronto network. Analysing the features of that network, we have cut the scenario of "TAPAS Cologne" to obtain a new sub-scenario.

This sub-scenario has 1416 edges and 716 nodes (73 of those are intersections managed by traffic lights and the rest are managed by priority rules). Equally, the routes of the scenario have been reduced to a more manageable number, using just the routes that are related to the chosen portion of the scenario. The total selected routes are 246374.

Moreover, in the "TAPAS Cologne" documentation it is said that for a proper simulation, the parameter *scale* must be set at 0.3. This means that only the 30% of the routes will be actually inserted during the simulation. For our sub-scenario, this scale value must be calibrated again. We have done this calibration by executing consecutive simulations increasing the value of *scale* in 0.01 for each new simulation. For each simulation, the number of teleports have been measured (a teleport is an event that happens in SUMO simulations when vehicles are blocked for a given time. It consists on the automatic disappearing and appearing of the vehicle, in order to unlock the simulation). The result of these simulations is shown in Figure 4.

Using the results of these tests, we have decided to increase the value of the parameter *scale* to 0.40. Because at this value is where we detect that the number of teleports starts to raise (although it is still a reasonable low value: 241), and allows us to use a high volume of traffic. Using this value for *scale* means that the actual number of routes inserted during the simulation is 98550.

B. System operation

As we have pointed before, the MAS proof of concept module has been completely written in Python. In the

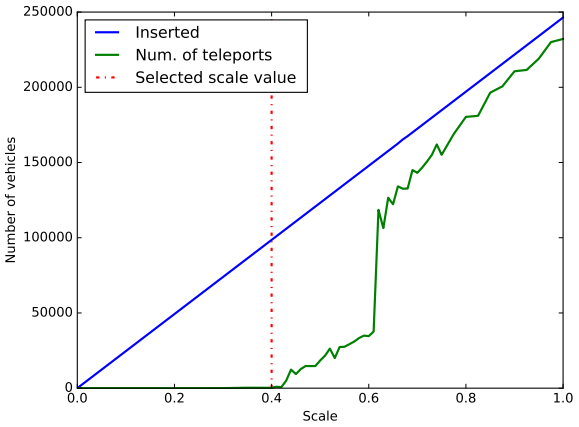


Fig. 4: Evolution of the teleports number according to the scale parameter.

Fig. 5: Representation of network edges (black lines) and intersection agents (red dots).

Figure 6, we show a block diagram of the whole simulation system.

The main steps followed while the system is being executed are:

1) System initialization: The application reads the data from the sub-scenario, generates the map division, initializes the data structures that will be used later, and launches a subprocess that starts the SUMO simulator.
2) Loop until end of the simulation: There is a parameter in the application configuration that states the duration in seconds of each simulation step. Specifically, we have chosen a value of 30 seconds for this parameter. The MAS module will perform the following tasks including requests to the simulation module (via TraCi):

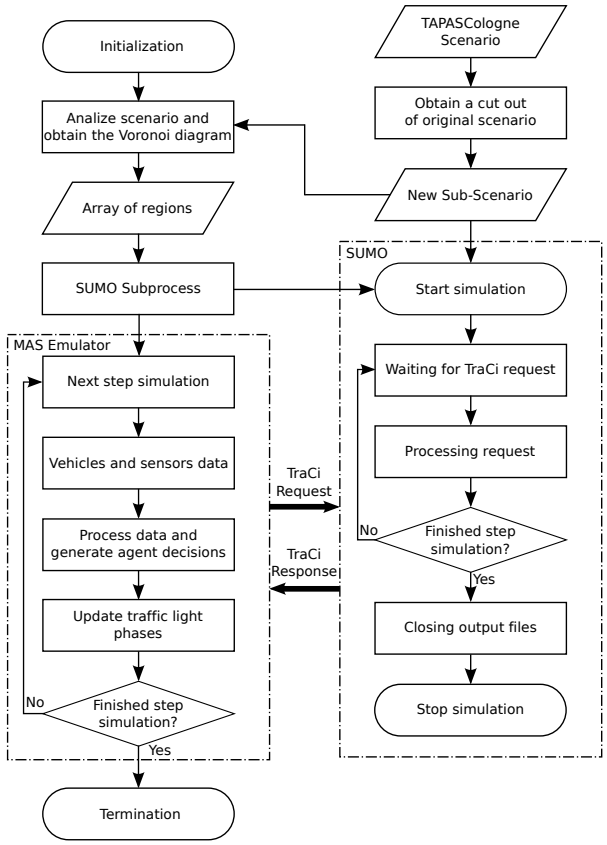


Fig. 6: Block diagram of the Multiagent System and simulation platform.

a) Requests one step simulation (30 seconds) and then waits for the end of the step.
b) Requests the current parameters of each vehicle that was active in that step of simulation, and also the information from the net induction loops.
c) Processes the information, and models it for the agents to make decisions.
d) Requests the variation of traffic light phases.

3) End of the simulation.
4) Analysis of the data contained in the SUMO output file, containing the simulation results. This results are compared with those obtained from a previous simulation performed in the same scenario but without the intervention of our MAS.

V. EVALUATION

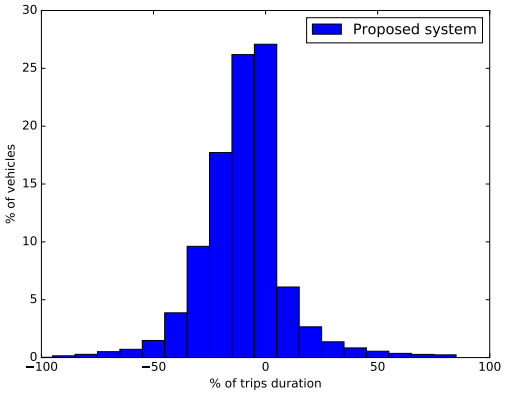
A. Experiment settings

The evaluation of the proposed system has been carried out by defining different sets of simulations over the same sub-scenario. In these experiments, the input data have been the same, and the results obtained have been evaluated over the same set of routes. These are the carried-out experiments:

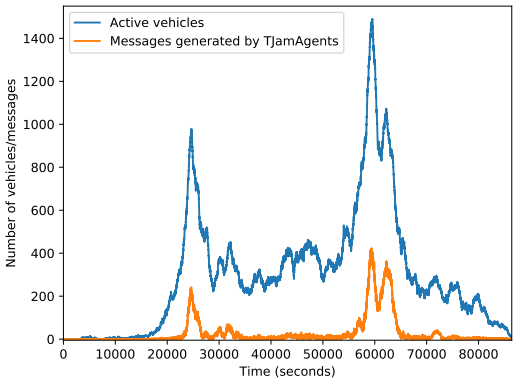
- Experiment 1 (DEFAULT): Simulation executed using the default configuration offered by the TAPAS-Cologne scenario. The traffic light phases are statically defined and do not change during the simulation.
- Experiment 2 (ACTUATED): Simulation executed using the SUMO Actuated Traffic Lights system.
- Experiment 3 (MAS): Simulation executed using the MAS emulation module.

After finishing all the experiments, we have focused on the evaluation of the duration of each simulated trip. Therefore, a decreasing duration of trip will show the improvement in traffic efficiency introduced by the variations in traffic light phases.

In Figure 7a, we show the results of comparing the experiment "DEFAULT" with the experiment "MAS". The x axis represents the percentage of increase or decrease in the trip durations (using steps of 10%) and the y axis represents the percentage of vehicles inserted over the total vehicles in the simulation. We also measured the number of active vehicles at each time step and the messages generated by the TJamAgents. These results are shown in Figure 7b.



(a) % of increase or decrease in the trip durations over the % of vehicles.



(b) Active vehicles during the simulation vs. messages generated by the TJamAgents

Fig. 7: Experiment results.

B. Results discussion

To allow for an easier analysis of the obtained results, the values have been grouped in three sets: one set for

the duration of trips that are lower than the ones in the experiment 1, other for the duration of trips that are equal, and a third one for the trips that are higher in duration. Table I shows the percentage of vehicles obtained in each category for experiments 2 and 3.

Tabla I: Results summary (% of total vehicles)

Trip durations	Lower	Equal	Higher
Actuated Traffic Lights	58.70	28.41	12.89
MAS	60.52	27.08	12.41

Given that the system has been defined to prioritize some traffic flows over others, it is expected that not every vehicle in the simulation is able to reduce the duration of its trips. The obtained results show that the percentage of vehicles suffering an increase in the duration of its trip is low. The value is lower than the 13% in the experiment 2 and 3. It is also worth mentioning that half of those vehicles only experienced a 10% of increase in trip duration.

The improvement between using the "actuated traffic lights" system and our MAS may seem not too remarkable, but it must be contextualized. The simulation using the first method needs an induction loop in every edge of the network that ends in an intersection with traffic lights. Besides that, it must evaluate in every simulation step the registered values by each sensor so it can modify the duration of the traffic lights accordingly. Conversely, the MAS can perform the same task without using fixed sensors distributed along the network, as it is able to obtain information form the vehicles themselves (and incidentally from the possible induction loops installed in some roads), it also limits the amount of data needed to make decisions, as each agent decides if it is necessary to communicate its state or not.

On the other hand, it is important to highlight that, in very high traffic congestion situations, such as the ones simulated in our experiments, there are certain intersections that can reach blocking states where the variations in the actuated traffic lights are not enough to solve them. In that kind of situations, the proposed MAS has been able to "unlock" 1790 vehicles, that have been able to reduce the duration of their trip in Experiment 3.

Finally, the values shown in Figure 7b validate that the MAS is capable of reacting to situations of serious traffic jams, regardless of the number of active vehicles.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have addressed traffic congestion situations which is one of the most relevant challenges in the most of big cities around the world. More specifically, the main goal of our proposal was to reduce the traffic congestion situations by changing duration of traffic light phases. After analysing the results, and comparing them with other solutions like the static definition of traffic light scheduling or the actuated traffic lights, it is possible to say the use of a MAS is effective.

A secondary goal, but mandatory for the validation of the proposal, has been the implementation of the process that emulates the MAS and managing to communicate it with a widely-used traffic simulator such as SUMO. This goal has been accomplished, and has allowed us to perform experiments over a simulation scenario, which provides higher guarantees about the usefulness of the solution.

Although the conducted experiments yield satisfactory results, there are some avenues for further research in this topic. The decision-making rules of the agents used in this first proof of concept implementation have been very simple but effective, just defining some triggering values for the agents from which the TJamAgents report the situation, and a weighted value applied to the data received by each InteserctionAgent. Those values are used to decide if the traffic is prioritized in one way or another. Once the viability of the system has been proven, and the connection with the simulation platform has been developed, the future work will be related to make more complex decision-making algorithms, that allow better and more effective results. Finally, it would be important to study which is the minimum percentage of agents that should participate in the MAS without deteriorating the system effectiveness.

ACKNOWLEDGMENT

This work has been supported by the Spanish Ministry of Economy and Competitiveness grants TIN2016-80622-P, TIN2014-61627-EXP and TEC2013-45183-R, and by the University of Alcalá through CCG2016/EXP-048.

REFERENCES

[1] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO - Simulation of Urban MObility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 128–138, December 2012.

[2] S. Uppoor and M. Fiore, "Large-scale urban vehicular mobility for networking research," in *Vehicular Networking Conference (VNC), 2011 IEEE*. IEEE, 2011, pp. 62–69.

[3] "TAPASCologne scenario," http://sumo.dlr.de/wiki/Data/Scenarios/TAPASCologne, online; accessed 14 July 2017.

[4] H. Taale, T. Bäck, M. Preuss, A. Eiben, J. De Graaf, and C. Schippers, "Optimizing traffic light controllers by means of evolutionary algorithms," in *Proceedings of the 6th European Congress on Intelligent Techniques and Soft Computing*, vol. 3, 1998, pp. 1730–1734.

[5] J. Garcia-Nieto, E. Alba, and A. C. Olivera, "Swarm intelligence for traffic light scheduling: Application to real urban areas," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 2, pp. 274 – 283, 2012, special Section: Local Search Algorithms for Real-World Scheduling and Planning. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197611000777

[6] A. L. C. Bazzan and F. Klugl, "A review on agent-based technology for traffic and transportation," *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 375–403, 006 2014.

[7] J. L. Adler and V. J. Blue, "A cooperative multi-agent transportation management and route guidance system," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 5, pp. 433–454, 2002.

[8] B. Chen and H. H. Cheng, "A review of the applications of agent technology in traffic and transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 485–497, 2010.

[9] D. A. Roozemond, "Using intelligent agents for pro-active, real-time urban intersection control," *European Journal of Operational Research*, vol. 131, no. 2, pp. 293–301, 2001.

[10] M. C. Choy, D. Srinivasan, and R. L. Cheu, "Cooperative, hybrid agent architecture for real-time traffic signal control," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 33, no. 5, pp. 597–607, Sept 2003.

[11] S. Fernandez, R. Hadfi, T. Ito, I. Marsa-Maestre, and J. R. Velasco, "Ontology-based architecture for intelligent transportation systems using a traffic sensor network," *Sensors*, vol. 16, no. 8, p. 1287, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/8/1287

[12] C. Ruey-Shun, C. Duen-Kai, and L. Szu-Yin, "Actam: Cooperative multi-agent system architecture for urban traffic signal control," *IEICE transactions on Information and Systems*, vol. 88, no. 1, pp. 119–126, 2005.

[13] A. Wegener, M. Piórkowski, M. Raya, H. Hellbrück, S. Fischer, and J.-P. Hubaux, "Traci: An interface for coupling road traffic and network simulators," in *Proceedings of the 11th Communications and Networking Simulation Symposium*, ser. CNS '08. New York, NY, USA: ACM, 2008, pp. 155–163. [Online]. Available: http://doi.acm.org/10.1145/1400713.1400740

[14] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): Methodology and large-scale application on downtown toronto," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1140–1150, Sept 2013.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6501

Evaluación de equipamiento de bajo coste para realizar medidas de red en entornos domésticos

Eduardo Miravalls-Sierra, David Muelas, Jorge E. López de Vergara, Javier Ramos, Javier Aracil
High Performance Computing and Networking Research Group
Departamento de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Fco. Tomás y Valiente, 11, 28049 Madrid, Spain
{eduardo.miravalls, dav.muelas, jorge.lopez_vergara, javier.ramos, javier.aracil}@uam.es

***Resumen*—En la actualidad, la proliferación de dispositivos móviles y accesos a Internet utilizando tecnologías inalámbricas en los entornos domésticos obliga a cambiar las metodologías para la realización de medidas de red. Para que éstas representen fidedignamente las condiciones ofrecidas a los usuarios, las prestaciones del equipamiento de medida y el número de dispositivos empleados deben adaptarse a las condiciones reales de un despliegue. Para facilitar y abaratar el desarrollo de medidas en estas condiciones, este trabajo presenta una evaluación de las capacidades de varias plataformas de propósito general y bajo coste. Nuestros resultados muestran que, aunque aparecen limitaciones relacionadas con cómo son conectadas a la red y los protocolos empleados, son aptas para medir una gran variedad de situaciones.**

***Palabras Clave*—Medidas de red, indicadores de calidad, calidad de servicio, WiFi, COTS**

I. INTRODUCCIÓN

A lo largo de los últimos años, el auge de las tecnologías WiFi para acceder a Internet ha transformado el uso que los usuarios finales hacen de los sistemas de telecomunicaciones. Por ello, ha aumentado el interés por estudiar la variabilidad de las prestaciones de este tipo de enlaces, ya que en muchos entornos se utilizan puntos de acceso WiFi como mecanismo para proveer conectividad a los usuarios.

Las características particulares de este tipo de medio de transmisión y sus interacciones con los protocolos de comunicaciones de Internet ocasionan la aparición de problemas de rendimiento que limitan la Calidad de Servicio (*Quality of Service*, QoS) y Experiencia (*Quality of Experience*, QoE) de dichos usuarios finales. Por otro lado, el usuario común realiza medidas con los dispositivos que tiene a su alcance, habitualmente un portátil o un dispositivo tipo Raspberry Pi u ODROID, que suelen tener pocos recursos y *hardware* de gama media o baja.

Teniendo en cuenta la importancia capital que tienen estos factores para los Proveedores de Servicios de Internet (*Internet Service Providers*, ISPs), para las Instituciones Públicas y para los usuarios finales, resulta indispensable el desarrollo de sistemas que permitan medir con suficiente precisión este tipo de conexiones, manteniendo el coste controlado.

En esta dirección, se han impulsado diversas iniciativas, como el Proyecto Bismark [1], el Atlas de RIPE o SamKnows [2], que persiguen el objetivo de caracterizar el comportamiento de las conexiones a Internet desde los accesos domésticos. Estos experimentos han utilizado una estrategia de emplazamiento de pequeñas sondas *hardware* basadas en arquitecturas de bajo coste (routers domésticos con OpenWRT, o dispositivos como Raspberry Pi u ODROID) para realizar medidas activas y pasivas de red.

Además, el dimensionado de un acceso inalámbrico que garantice un cierto nivel de QoS a un número determinado de usuarios no es fácilmente extrapolable a partir de medidas con un número reducido de dispositivos. Las características de los protocolos de acceso al medio ocasionan que al aumentar el número de usuarios las prestaciones de la red decaigan de manera significativa, incrementando el *jitter* y las interferencias en la red [3]. Esto motiva el desarrollo de sistemas de medida con numerosos dispositivos, que evalúen la QoS en el punto de trabajo típico del despliegue.

Para facilitar y abaratar la realización de medidas en estas condiciones, presentamos una evaluación de los resultados que pueden obtenerse empleando dispositivos de gama económica. Las principales aportaciones del artículo son (i) la identificación de las características del equipamiento que tienen un efecto observable en los resultados de medida, y (ii) la extracción en base a ellas de una serie de recomendaciones.

El resto del artículo tiene la siguiente estructura. En la Sección II definimos las principales métricas indicadoras de la calidad (*Key Performance Indicators*, KPIs) que hemos utilizado. En la Sección III describimos el entorno experimental considerado, en el que hemos medido con herramientas y técnicas de estado del arte para aumentar

la generalidad de nuestros resultados. En la Sección IV comentamos y analizamos los resultados obtenidos para, finalmente, extraer las principales conclusiones y definir líneas de trabajo futuro en la Sección V.

II. MEDIDAS DE PRESTACIONES DE RED

En esta sección incluimos algunos preliminares referentes a las medidas de prestaciones de red. En primer lugar, revisamos diversas propuestas de metodologías y buenas prácticas, así como sus limitaciones tanto en términos de medidas en general como en el contexto de conexiones inalámbricas. Por otro lado, proporcionamos las definiciones que vamos a manejar de los KPIs utilizados habitualmente por las operadoras, extraídas de este primer bloque de propuestas. Específicamente, por su relación con la QoE, vamos a considerar el ancho de banda, la latencia, el *jitter* y las pérdidas.

A. Metodologías de medida

Un problema recurrente a la hora de medir las prestaciones de una red, es que las medidas son muy sensibles de la metodología utilizada. Es por ello que surgieron iniciativas como *IP Performance Metrics (IPPM)* [4] o el *Two Way Active Measurement Protocol* (TWAMP) [5], que intentan fijar buenas prácticas y definiciones para las métricas, de manera que sean claramente objetivas y repetibles. Así, se trata de evitar resultados no repetibles debido a, entre otras cosas, componentes estocásticas no controladas en los modelos, errores de medida debidos a la metodología y diseño experimental, etc.

Pese a estos intentos, varios estudios muestran que se pueden obtener diferentes resultados para la misma métrica, no siempre equivalentes, dependiendo de la herramienta utilizada —por ejemplo, el reciente estudio incluido en [6] muestra las divergencias entre algunas herramientas populares, ampliamente utilizadas por usuarios finales. Además, diversos factores relacionados con el propio sistema desde el que se realizan las medidas (p.e. carga de CPU, política del planificador, etc.) tienen efectos significativos sobre los resultados de las medidas [7].

Cuando se trata de medir las capacidades de un canal inalámbrico la tarea se complica, ya que este tipo de enlaces tienden a introducir pérdidas o latencias aleatorias que son difíciles de predecir, y que dependen de muchos factores. Con frecuencia se publican nuevos estudios en los que se intenta modelar el retardo de enlaces inalámbricos como en [8], [9]. En ambos estudios recurren a modelos de tipo árbol de decisión (*decision tree*) o *random forest*, que son interpretables.

Otra complicación que pueden introducir los accesos inalámbricos es una pérdida notable de rendimiento cuando hay muchos usuarios [10] debido a las interacciones entre las peculiaridades del nivel de enlace y los protocolos de nivel superior.

Por ejemplo, TCP puede aumentar las retransmisiones como consecuencia de la variabilidad de la latencia por las colisiones a nivel de enlace, sin que dichas retransmisiones fuesen realmente necesarias. Como consecuencia, la evaluación de accesos que vayan a ser compartidos por un gran número de usuarios deben considerar estos casos a la hora de evaluar las prestaciones esperadas, lo que motiva el abaratamiento de los dispositivos de medida utilizados.

B. Ancho de banda

El ancho de banda de un enlace mide la cantidad de información por unidad de tiempo que es capaz de transportar (p.e. en bits por segundo). En la práctica el término "ancho de banda" puede referirse a uno de varios conceptos [11].

Dado un camino extremo a extremo formado por un conjunto ordenado de n enlaces $i = 1, ..., n$, se define la capacidad del enlace i como la máxima tasa de transmisión a nivel IP, B_i. Por tanto, la capacidad B^* del camino se define como el mínimo de las capacidades $\{B_i\}_1^n$ de cada uno de los enlaces que forman el camino

$$B^* = \min_{i=1...n} \{B_i\} \tag{1}$$

En un camino los enlaces i_k tales que $B_{i_k} = B^*$ son denominados *narrow link*s o enlace angosto.

El ancho de banda disponible, *available bandwidth*, de un camino se define como la capacidad no usada en ese instante de tiempo. Es la métrica complementaria del ancho de banda consumido. Depende de B^* y de la cantidad de tráfico que esté circulando en ese momento por la red.

Por otro lado, el *Bulk Transfer Capacity* (BTC) es la máxima cantidad de información que un protocolo que implemente control de la congestión, p.e. TCP, puede enviar por unidad de tiempo.

Cuando se realizan medidas activas, según el protocolo de transporte utilizado se medirán cosas diferentes. La técnica de trenes de paquetes utiliza UDP, que permite medir la capacidad B^* del camino [7]. En cambio si se utiliza el protocolo TCP se mide el BTC de un camino.

Finalmente, se puede medir el ancho de banda consumido de manera pasiva contabilizando cuántos bytes por unidad de tiempo están transmitiendo por un enlace una aplicación o conjunto de equipos específicos.

C. Latencia

La latencia es el tiempo que transcurre desde que un byte es enviado hasta que llega al otro extremo.

Se distinguen la latencia en un solo sentido (*One Way Delay*, OWD) [12]; y la suma de las latencias en los dos sentidos, el tiempo de ida y vuelta (*Round Trip Time*, RTT). Estimar el OWD de una comunicación es difícil, ya que requiere que los relojes de los dos extremos se mantengan sincronizados. Por esta razón, se suele medir el RTT, que sortea el problema de sincronización de relojes ya que la base de tiempos es la misma. Asumiendo que las dos latencias en un solo sentido sean iguales, el OWD se estima en ocasiones como la mitad del tiempo que transcurre entre una petición y su respuesta, como se sugiere en protocolos como Q4S [13].

La forma más sencilla de estimar el RTT es realizando medidas activas con un mecanismo similar a un `ping` [14]. Estimar la latencia de manera pasiva monitorizando conexiones TCP requiere conocer los protocolos de

nivel superior que se están utilizando para poder identificar peticiones con respuestas. Además, algoritmos como el de Nagle o el asentimiento retrasado (*delayed ACK*) dificultan la estimación del RTT debido a que las respuestas pueden no ser inmediatas, sobreestimando la latencia. Una forma inequívoca de identificar una petición y su respuesta en TCP es durante el establecimiento de la conexión, donde se puede estimar el RTT como la diferencia de tiempos entre el SYN y el ACK en el lado del servidor:

$$RTT = t_{ACK} - t_{SYN} \tag{2}$$

y de SYN,ACK y el ACK en el cliente:

$$RTT = t_{SYN,ACK} - t_{SYN} \tag{3}$$

D. Variación de la latencia

La variación de la latencia o *jitter* es un parámetro importante de la calidad, ya que afecta gravemente a aplicaciones multimedia tales como Voz sobre IP (VoIP).

Existen multitud de definiciones de *jitter*. La definición que vamos a utilizar es la dada en [15] y de manera equivalente en [13]: dadas $N+1$ medidas de latencia unidireccionales, $\{l_i\}_{i=0}^{N}$ podemos calcular N diferencias, $\{\Delta_j\}_{j=1}^{N}$

$$\Delta_j = |l_j - l_{j-1}| \tag{4}$$

y definir el *jitter* mediante un estadístico de centralidad de los $\{\Delta_j\}$ como la media aritmética o la mediana. Otras alternativas son estimar el *jitter* como el resultado de aplicar un filtro exponencial [15] de parámetro 1/16 a los $\{\Delta_j\}$; o estimarlo como la desviación estándar de las $\{l_i\}$.

E. Pérdidas

Las pérdidas son un indicador de saturación de la red entre otros, ya que los *routers* y los equipos finales descartan paquetes cuando reciben paquetes más rápido de lo que pueden procesar. También pueden perderse paquetes cuando los datos se corrompen, ya sea por un problema de *hardware* o por ruido en la comunicación, muy habitual en enlaces inalámbricos.

Los protocolos de medida Q4S o IPPM [16] utilizan números de secuencia para estimar las pérdidas. Esta estimación se calcula como el ratio de paquetes que no se recibieron frente al total esperado.

III. DESCRIPCIÓN DEL ENTORNO EXPERIMENTAL

Para la realización de los experimentos hemos utilizado el siguiente equipamiento:

- Dos routers TP-Link Archer C7 AC 1750 con una configuración MIMO $3 \times 3 \times 3$. Uno de ellos se ha dejado con el *firmware* de fábrica[1], que denotaremos como "TP-Link"; y en el segundo se ha instalando una distribución Linux dd-wrt[2], y nos referiremos a éste como "dd-wrt". El precio de estos dispositivos ronda los 80 euros cada uno.

[1] 3.15.1 Build 160616 Rel.44182n

[2] http://dd-wrt.com/site/index

Tabla I
HERRAMIENTAS UTILIZADAS

Herramienta	Protocolo	Ancho de banda	Latencia	*Jitter*	Pérdidas
iperf3	TCP	✓	✗	✗	✓
iperf3	UDP	✓	✗	✓	✓
ping	ICMP	✗	✓	✗	∼

- PC1: con un procesador Intel Core i7 CPU 860 a 2.80GHz, 8GB DDR3 a 1333MHz, y dos interfaces de red: la integrada en la placa de intel Gigabit Ethernet (conectada a Internet) y una segunda interfaz utilizada para conectarse al router que proporciona una tarjeta de red PCI Broadcom Corporation NetXtreme II BCM5709 Gigabit Ethernet (rev 20). El sistema operativo era Ubuntu 14.04 Mate.
- PC2: tiene un Intel Core i7-2600 CPU a 3.40GHz, 8 GB de RAM DDR3 a 1333MHz, y se ha utilizado la interfaz de red integrada Intel Gigabit Ethernet. El sistema operativo era CentOS 7.
- Y por último, se ha utilizado una placa ODROID C2[3] con un adaptador de red WiFi 802.11n USB Edimax EW-7811UN que alcanza hasta los 150Mb/s. El precio del kit es de 100 euros aproximadamente.

Para realizar las medidas hemos utilizado varias herramientas del estado del arte: `iperf3` para medir el ancho de banda, las pérdidas y el *jitter*; y `ping` para medir la latencia, ya que `iperf3` no proporciona esa métrica.

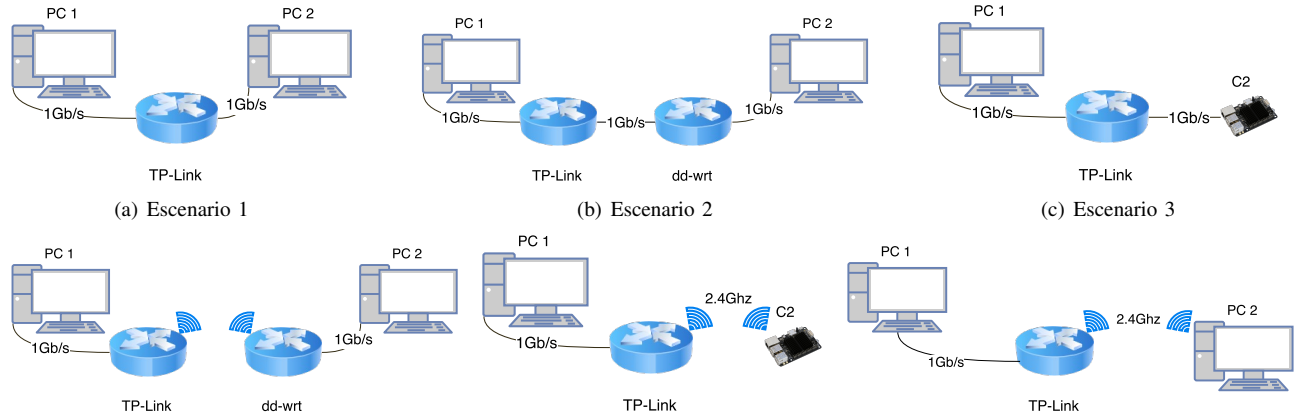
`iperf3` es la herramienta estándar *de facto* para medir la capacidad de un canal. `iperf3` tiene diferentes comportamientos en función del protocolo de transporte que se quiera utilizar. Utilizando TCP, es capaz de estimar el ancho de banda útil máximo; sin embargo, cuando se utiliza UDP es necesario conocerlo, ya que al no tener retroalimentación del otro extremo, `iperf3` envía siempre a máxima tasa, y solo el servidor es capaz de detectar cuántos paquetes le llegan realmente. Por ello, para medir el ancho de banda con UDP en cada escenario, se ha buscado primero de manera iterativa el ancho de banda máximo que se podía lograr sin pérdidas antes de tomar las medidas.

Por tanto, es importante observar los valores calculados en el servidor siempre, ya que nos interesa estudiar los efectos que haya podido tener la red sobre el tráfico.

Para evitar los efectos del planificador, se ha fijado la afinidad de `iperf3` a un mismo *core* tanto en el cliente como en el servidor, aislado en el arranque para evitar que la competición por los recursos de CPU pudieran interferir en las medidas.

Se tomaron 10 medidas con `iperf3` por cada sentido en cada escenario, lo que nos proporciona un total de 100 muestras por sentido gracias a que cada ejecución de `iperf3` proporciona 10 muestras de 1 segundo. Del mismo modo, hemos realizado 100 medidas con `ping` por sentido para estimar la latencia de cada enlace. El uso de CPU fue siempre inferior al 20 %, por lo que no fue un factor limitante en nuestras pruebas. En todos los casos las pérdidas fueron nulas o menores del 1 % de los paquetes.

[3] http://www.hardkernel.com/main/products/prdt_info.php



(a) Escenario 1 (b) Escenario 2 (c) Escenario 3

(d) Escenarios 4 y 5 con red a 2.4GHz, 8 y 9 a 5GHz (e) Escenario 6 (f) Escenario 7

Figura 1. Topologías de red para cada escenario

En la Tabla I se muestran las distintas configuraciones utilizadas y los datos que proporcionan cada una de ellas. Cabe destacar que aunque `ping` se podría utilizar como estimador de pérdidas, en este caso se utiliza como una medida de conectividad [17]. Esto se debe a que la tasa configurada (un paquete por segundo) fuerza a considerar duraciones experimentales extensas para obtener estimadores significativos de pérdidas en las condiciones habituales de red. Por ejemplo, con una tasa de un 1 % de pérdidas, sería necesaria una duración del experimento de 100 segundos para tener suficientes muestras para la estimación [13].

Hemos diseñado varios escenarios experimentales en el que los únicos equipos conectados a esa red son un equipo que hace las veces de cliente y otro que hace las veces de servidor. Las medidas se han realizado en un laboratorio universitario, en escenarios cableados, WiFi en la banda de 2.4GHz y WiFi en la banda de 5GHz. Cuando se realizan experimentos con WiFi, éstos se pueden ver afectados por otras redes inalámbricas y por otras aplicaciones que usen esa banda de frecuencias. En nuestro caso se han realizado los experimentos en presencia de otras redes WiFi de 2.4Ghz y una red de 5GHz. Como nuestro objetivo es simular entornos domésticos reales, y estos fenómenos son factores habituales que pueden afectar a la QoS, no hemos aislado nuestros equipos contra estos fenómenos. Los escenarios son:

- Escenario 1: este escenario consiste en conectar los dos equipos al mismo router directamente con cable, como se muestra en la Fig. 1(a), como escenario de referencia de un caso ideal con conexión directa con el servidor sin saltos inalámbricos.
- Escenario 2: en este escenario comprobamos las características de red que el router dd-wrt es capaz de medir. En este escenario, ilustrado en la Fig. 1(b), los resultados se ven claramente afectados por las limitadas capacidades computacionales del router.
- Escenario 3: como ya hemos mencionado, con la proliferación de los dispositivos tipo ODROID, es interesante investigar las capacidades de red que son capaces de aprovechar. Por ello, en este escenario sustituimos el PC1 del escenario 1 por una ODROID C2 como se muestra en la Fig. 1(c).

y los mismos escenarios pero sustituyendo un enlace cableado por un enlace WiFi:

- Escenario 4: en este escenario conectamos los dos equipos a través de una red WiFi, en la que el router TP-Link hace las veces de AP (*Access Point*, Punto de Acceso) y el router dd-wrt actúa de pasarela entre el PC2 y la red WiFi en la banda de 2.4GHz con un ancho de canal de 40MHz (MCS 22 en tx y MCS 23 en rx). Este escenario se ilustra en la Fig. 1(d).
- Escenario 5: este escenario utiliza la misma topología que el escenario 4, pero ahora el equipo PC2 sólo se utiliza para establecer una sesión SSH al router dd-wrt, que es el que se utilizará para realizar las medidas. Esto evita mantener la sesión SSH por el enlace WiFi, que puede alterar las medidas realizadas.
- Escenario 6: en este escenario medimos las características de red que un usuario experimentaría si decide conectar su ODROID mediante el adaptador de red WiFi USB (MCS 7).
- Escenario 7: en este escenario conectamos el PC2 con el mismo adaptador de red USB del escenario anterior para ver el efecto que tiene sobre las métricas en este equipo.
- Escenario 8: en este escenario conectamos el PC2 con el PC1 manteniendo la topología de la Fig. 1(d) pero con un enlace WiFi en la banda de 5GHz con un ancho de canal de 80MHz (MCS 8).
- Escenario 9: en este escenario mantenemos la topología anterior pero medimos desde dd-wrt.

En todos los escenarios anteriormente descritos, hemos fijado el PC1 y el router TP-Link, que hacen las veces de "servidor" en nuestras pruebas a las que se conectan diferentes clientes mediante distintos tipos de enlace. Por ello, en nuestros resultados nos referimos siempre con "Bajada" al sentido desde el PC1 al otro equipo; y nos referimos con "Subida" al sentido contrario.

Tabla II
MEDIDAS DE ANCHO DE BANDA EN MBITS/S

Protocolo	Sentido	Escenario 1 $\hat{\mu} \pm \hat{\sigma}$	Escenario 2 $\hat{\mu} \pm \hat{\sigma}$	Escenario 3 $\hat{\mu} \pm \hat{\sigma}$	Escenario 4 $\hat{\mu} \pm \hat{\sigma}$	Escenario 5 $\hat{\mu} \pm \hat{\sigma}$	Escenario 6 $\hat{\mu} \pm \hat{\sigma}$	Escenario 7 $\hat{\mu} \pm \hat{\sigma}$	Escenario 8 $\hat{\mu} \pm \hat{\sigma}$	Escenario 9 $\hat{\mu} \pm \hat{\sigma}$
TCP	Bajada	935.24± 7.99	384.80± 7.98	934.25± 7.33	164.82± 9.81	177.91±13.94	96.63± 9.56	33.79± 7.08	274.93±12.81	434.87±25.82
	Subida	930.74±11.25	179.72± 1.99	941.32± 7.03	111.77±10.20	77.21±28.78	52.93±18.82	9.72± 6.68	213.44±12.44	170.93± 6.76
UDP	Bajada	953.17±11.48	10.15± 0.33	943.30±40.15	53.47± 1.56	8.32± 0.28	45.00± 4.64	42.48± 2.21	246.79± 7.80	8.61± 0.29
	Subida	947.22±11.40	269.83± 8.40	19.05± 0.40	44.45± 1.46	96.98±17.38	62.00±17.06	37.97± 7.84	42.93± 1.27	208.71± 6.41

Tabla III
MEDIDAS DE RTT EN MILISEGUNDOS

Sentido	Escenario 1 $\hat{\mu} \pm \hat{\sigma}$	Escenario 2 $\hat{\mu} \pm \hat{\sigma}$	Escenario 3 $\hat{\mu} \pm \hat{\sigma}$	Escenario 4 (*) $\hat{\mu} \pm \hat{\sigma}$	Escenario 5 $\hat{\mu} \pm \hat{\sigma}$	Escenario 6 $\hat{\mu} \pm \hat{\sigma}$	Escenario 7 $\hat{\mu} \pm \hat{\sigma}$	Escenario 8 $\hat{\mu} \pm \hat{\sigma}$	Escenario 9 $\hat{\mu} \pm \hat{\sigma}$
Bajada	0.221±0.069	0.204±0.012	0.6 ±0.038	5.491±10.939	3.318±4.535	75.281±90.705	16.045± 75.715	1.098 ± 0.449	0.926 ± 0.063
Subida	0.192±0.044	0.317±0.065	0.629±0.029	3.042± 3.305	2.987±4.281	17.347±90.427	92.718±302.146	1.059 ± 0.474	1.029 ± 0.292

Tabla IV
MEDIDAS DE *jitter* EN MILISEGUNDOS SEGÚN `IPERF3`

Sentido	Escenario 1 $\hat{\mu} \pm \hat{\sigma}$	Escenario 2 $\hat{\mu} \pm \hat{\sigma}$	Escenario 3 $\hat{\mu} \pm \hat{\sigma}$	Escenario 4 $\hat{\mu} \pm \hat{\sigma}$	Escenario 5 $\hat{\mu} \pm \hat{\sigma}$	Escenario 6 $\hat{\mu} \pm \hat{\sigma}$	Escenario 7 $\hat{\mu} \pm \hat{\sigma}$	Escenario 8 $\hat{\mu} \pm \hat{\sigma}$	Escenario 9 $\hat{\mu} \pm \hat{\sigma}$
Bajada	0.113±0.012	0.018±0.007	0.013±0.004	0.020±0.016	0.035±0.033	0.238±0.156	0.114±0.052	0.036±0.005	0.023±0.005
Subida	0.109±0.012	0.004±0.001	0.016±0.002	0.056±0.130	0.202±0.180	0.554±1.112	3.293±7.383	0.038±0.012	0.094±0.021

Tabla V
MEDIDAS DE *jitter* EN MILISEGUNDOS A PARTIR DE `PING`

Sentido	Métrica	Escenario 1	Escenario 2	Escenario 3	Escenario 4 (*)	Escenario 5	Escenario 6	Escenario 7	Escenario 8	Escenario 9
Bajada	$\hat{\mu}$	0.035	0.013	0.035	5.339	3.624	64.866	15.488	0.332	0.383
	$\hat{\mu}_c$	0.024	0.008	0.024	1.255	1.580	22.500	0.840	0.142	0.028
	$\hat{\sigma}$	0.065	0.012	0.031	11.495	5.738	114.916	75.423	0.466	0.806
	$\hat{\sigma}_c$	0.030	0.014	0.046	2.943	4.720	57.850	2.920	0.398	0.392
Subida	$\hat{\mu}$	0.051	0.071	0.023	2.386	2.627	21.052	82.314	0.285	0.200
	$\hat{\mu}_c$	0.045	0.057	0.012	0.560	0.352	1.450	0.390	0.038	0.033
	$\hat{\sigma}$	0.038	0.074	0.026	4.100	5.307	127.741	266.506	0.614	0.379
	$\hat{\sigma}_c$	0.049	0.052	0.027	2.360	2.090	3.910	2.345	0.228	0.230

Hemos medido en el escenario en el que solo hay un cliente, y los distintos canales de la comunicación se encuentran completamente disponibles para él, lo que nos da cotas máximas de rendimiento que podríamos experimentar. Estos resultados sirven como medida inicial, antes de adquirir más dispositivos (lo que aumentaría el coste) con los que realizar medidas en paralelo y estudiar los fenómenos de competición de acceso al medio que éstos experimentarán. Sin embargo, no podemos olvidar que se están realizando avances mejorando las tecnologías de acceso WiFi, aumentando el ancho de banda disponible utilizando bandas y protocolos de acceso al medio diferentes [18], y se publican estudios sobre su impacto sobre los protocolos de transporte [19].

IV. ANÁLISIS DE LOS RESULTADOS

En la Tabla II se muestran las medidas de ancho de banda para cada escenario. Fijándonos en el escenario 2 se aprecia que `iperf3` ejecutado en el router dd-wrt es capaz de procesar el tráfico TCP el doble de rápido del que es capaz de generarlo. En bajada no es capaz de procesar más de 10Mb/s de tráfico UDP. Esto se debe a un *bug* conocido de `iperf3` que parece afectar a la versión 3.1.X[4][5] que justifica que en el resto de escenarios inalámbricos (5 y 9) solo sea capaz de medir hasta 8Mb/s en bajada. Destaca la abultada reducción de ancho de banda cuando se utiliza UDP al medir con la ODROID. Este resultado fue muy similar en todas nuestras pruebas y se tuvo especial cuidado en que no hubiese fragmentación IP, lo que indica que en subida `iperf3` está generando el tráfico UDP de forma poco eficiente, probablemente relacionado con el *bug* anteriormente mencionado.

Comparando el escenario 6 con el 7 concluimos que el ancho de banda medido con el adaptador Edimax depende de cómo se realizan las interconexiones con el USB y cómo lo gestiona el sistema operativo, ya que a pesar de disponer el PC2 de muchos más recursos, el ancho de banda percibido por éste es muy inferior al medido por la ODROID con exactamente la misma antena. De estos resultados se puede observar que independientemente de la tecnología de acceso inalámbrico, el ancho de banda máximo percibido por un usuario se verá notablemente reducido frente al ancho de banda en bajada utilizando TCP. Comparando el escenario 8 frente al 4 vemos un aumento de más del 50 % en todas las medidas salvo en subida con UDP. Se esperaría un aumento mayor del ancho de banda medido, pero debido a pérdidas esporádicas, TCP no es capaz de alcanzar mayor *throughput*, ya que el enlace en la banda de 5GHz es capaz de alcanzar tasas de hasta 440Mb/s.

En la Tabla III mostramos los resultados de medidas de latencia obtenidos utilizando `ping`. Hemos marcado con un asterisco las medidas del escenario 4 ya que de manera sostenida de cada 100 sondas de `ping` se experimentaba un 10 % de pérdidas, de modo que los resultados mostrados son solo para las respuestas que llegaron a las solicitudes de eco.

En las Tabla IV mostramos el `jitter` calculado por `iperf3` frente al que calculamos utilizando la Ec. 4

[4]https://github.com/esnet/iperf/issues/234
[5]https://github.com/esnet/iperf/issues/296

con los resultados de `ping` en la Tabla V en la que mostramos tanto la media como la mediana $\hat{\mu}_c$, y tanto la desviación típica como el rango intercuartílico $\hat{\sigma}_c$. Estos resultados indican una gran dispersión de los valores, la presencia de valores atípicos (*outliers*) como es el caso de los escenarios 4 a 6 y de pérdidas ocasionales en el enlace. Los resultados tan dispares entre las medidas de `iperf3` y las calculadas con `ping` no son contradictorios, ya que miden situaciones diferentes. `iperf3` envía todos los paquetes de manera que en cuanto se termina de transmitir un paquete se comienza inmediatamente a transmitir el siguiente (*back-to-back*). Es decir, el *jitter* en el caso de estar transmitiendo de manera sostenida y sin que el otro extremo transmita nada. Sin embargo, el *jitter* medido por `ping` se corresponde con la variabilidad en la latencia que sufren paquetes enviados periódicamente, lo que significa que aplicaciones que transmitan de manera periódica y determinista pero no sostenida sufrirán de mucha mayor variabilidad en la latencia percibida. Por otro lado, es interesante ver que el `jitter` en los escenarios inalámbricos es siempre superior en subida frente a la bajada según `iperf3` aunque según la Tabla V la situación es la contraria.

Finalmente, las medidas de TCP se han validado con otras herramientas del estado del arte, como `nttcp`, con la que se obtienen resultados similares a los mostrados.

V. CONCLUSIONES

En este artículo hemos explorado varios escenarios domésticos en los que hemos realizado medidas con equipamiento de gama baja-media, y hemos medido las prestaciones de red que estos ofrecen utilizando los principales indicadores de calidad.

Hemos comprobado que el ancho de banda que un dispositivo puede medir varía en gran medida según la capacidad de cómputo del mismo y del mecanismo de acceso a la red WiFi que utilice. Pese a que los dispositivos embebidos de bajo coste tipo ODROID no aprovechan todo el ancho de banda de la red inalámbrica, son capaces de alcanzar hasta 100Mb/s con TCP en una red de 2.4GHz con un adaptador económico. No obstante, un PC de sobremesa es capaz de medir anchos de banda muy superiores a los 100Mb/s con TCP en la banda de 2.4GHz y de 200Mb/s en la banda de 5GHz conectado a través de otro router que gestione la comunicación inalámbrica. De hecho, se obtienen resultados aún mejores midiendo desde el propio router.

Las medidas de latencia dependen de la forma de conectarse a la red, ya que hemos obtenido resultados cercanos a 1ms en los escenarios con una red WiFi en la banda de 5GHz frente a la gran variabilidad en la banda de 2.4GHz.

Por último, incluso con una utilización mínima del canal y en entornos controlados, los enlaces inalámbricos presentan pérdidas esporádicas, *jitter* elevado, y un ancho de banda limitado; lo que puede afectar gravemente a la calidad de experiencia del usuario.

Estos resultados muestran la importancia capital a la hora de desplegar una red inalámbrica considerar estos retos a la hora de implantar aplicaciones con requisitos de latencia y de variabilidad acotadas.

Algunas líneas de trabajo futuro que estamos explorando actualmente son evaluar el efecto de la concurrencia de múltiples usuarios sobre las medidas; y los efectos causados por uno o varios usuarios compitiendo por el mismo canal, unos tratando de subir datos mientras otros tratan de bajar datos.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y del Fondo Europeo de Desarrollo Regional a través de los proyectos TRÁFICA (MINECO / FEDER TEC2015-69417-C2-1-R) y RACING DRONES (MINECO / FEDER RTC-2016-4744-7). Los autores también agradecen al Ministerio de Educación Cultura y Deporte por la beca de colaboración del primer autor.

REFERENCIAS

[1] S. Sundaresan, S. Burnett, N. Feamster, and W. De Donato, "BISmark: A testbed for deploying measurements and applications in broadband access networks." in *USENIX Annual Technical Conference*, 2014.

[2] M. Bagnulo, T. Burbridge, S. Crawford, P. Eardley, J. Schoenwaelder, and B. Trammell, "Building a standard measurement platform," *IEEE Communications Magazine*, May 2014.

[3] S. Choi, K. Park, and C.-k. Kim, "Performance impact of interlayer dependence in infrastructure WLANs," *IEEE Transactions on Mobile Computing*, 2006.

[4] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis, "RFC 2330: Framework for ip performance metrics," 1998.

[5] H. Hedayat, R. Krzanowski, A. Morton, K. Yum, and J. Babiarz, "RFC 5357: A two-way active measurement protocol," 2008.

[6] E. Atxutegi, F. Liberal, E. Saiz, and E. Ibarrola, "Toward standardized internet speed measurements for end users: current technical constraints," *IEEE Communications Magazine*, 2016.

[7] J. Ramos, P. S. del Río, J. Aracil, and J. L. de Vergara, "On the effect of concurrent applications in bandwidth measurement speedometers," *Computer Networks*, 2011.

[8] C. Pei, Y. Zhao, G. Chen, R. Tang, Y. Meng, M. Ma, K. Ling, and D. Pei, "Wifi can be the weakest link of round trip network latency in the wild," in *35th Annual IEEE International Conference on Computer Communications (INFOCOM)*, 2016.

[9] Z. Hu, Y.-C. Chen, L. Qiu, G. Xue, H. Zhu, N. Zhang, C. He, L. Pan, and C. He, "An in-depth analysis of 3G traffic and performance," in *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*. ACM, 2015.

[10] M. Maity, B. Raman, and M. Vutukuru, "Tcp download performance in dense wifi scenarios," in *2015 7th International Conference on Communication Systems and Networks (COMSNETS)*, 2015.

[11] J. Strauss and M. F. Kaashoek, "Estimating bulk transfer capacity."

[12] G. Almes, M. Zekauskas, S. Kalidindi, and A. Morton, "RFC 7679: A one-way delay metric for IP performance metrics (IPPM)," 2016.

[13] J. Salvachua, J. Garcia, J. Quemada, M. Cortes, L. Vizcaino, G. Fernandez, J. Lajo, and C. Barcenilla, "Internet Draft: The quality for service protocol," 2015.

[14] G. Almes, M. J. Zekauskas, and S. Kalidindi, "RFC 2681: A round-trip delay metric for IPPM," 1999.

[15] C. Demichelis and P. Chimento, "RFC 3393: IP packet delay variation metric for IP performance metrics (IPPM)," 2002.

[16] G. Almes, S. Kalidindi, and M. Zekauskas, "RFC 2680: A one-way packet loss metric for IPPM," 1999.

[17] J. Mahdavi and V. Paxson, "RFC 2678: IPPM metrics for measuring connectivity," 1999.

[18] E. Charfi, L. Chaari, and L. Kamoun, "PHY/MAc enhancements and QoS mechanisms for very high throughput WLANs: A survey," *IEEE Communications Surveys & Tutorials*, 2013.

[19] R. Karmakar, S. Chakraborty, and S. Chattopadhyay, "Impact of IEEE 802.11 n/ac PHY/MAC high throughput enhancements over transport/application layer protocols-a survey," *arXiv preprint:1702.03257*, 2017.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6554

Optimización de la eficiencia en la conducción para rutas predeterminadas

Roberto García[1], Alejandro G. Tuero[1], Laura Pozueco[1], Xabiel G. Pañeda[1], Victor Corcoba[1], José A. Sanchez[1], David Melendi[1], Abel Rionda[2].

[1] Departamento de Informática, Universidad de Oviedo

[2]ADN Mobile Solutions. Parque Tecnológico. Gijón, Asturias, España

{garciaroberto, garciatalejandro, pozuecolaura, xabiel, corcobavictor, sanchezsjose, melendi}@uniovi.es, abel.rionda@adnmobilesolutions.com

***Resumen*- The transport sector is one of the main causes of the emission of pollutants to the environment. Among the different alternatives for reducing consumption and emission of harmful particles the most attractive for professional transport companies is the use of efficient driving techniques as it allows to take advantage of the existing fleet without the need to invest in new vehicles and technology . In this work we determine the optimal driving technique to minimize fuel consumption in a route. The results can be applied in training courses in efficient driving. For the calculations we use real data of a professional bus fleet. Then, we develop a consumption model and, using route optimization, we determine the driving technique that minimizes the consumption in the analyzed route. The results indicate that efficient driving has a very significant influence on fuel consumption. With optimum driving, reductions in consumption of up to 15 liters/100km (28% of reduction) could be achieved in the analyzed route. For the whole company, this is a significant reduction of consumption, considering the large distances covered by the professionals of the transport sector.**

***Palabras Clave*- Conducción eficiente, optimización de rutas, transporte profesional**

I. INTRODUCCIÓN

Las técnicas de formación en conducción eficiente en flotas de transporte profesional están recibiendo mucha atención en los últimos tiempos. El consumo de combustibles fósiles hace del sector del transporte uno de los principales causantes de la emisión de CO2 y otras partículas contaminantes al ambiente. Para mitigar esta situación se puede mejorar la tecnología de los vehículos o hacer un uso más eficiente de la ya existente mediante programas de mantenimiento y técnicas de conducción eficiente.

En las flotas de transporte profesional la opción más atractiva, y económica, es la utilización de técnicas de conducción eficiente ya que permite aprovechar la flota de vehículos existente sin necesidad de invertir en nuevos vehículos y tecnología. Para lograr una disminución relevante en el consumo de combustible es necesario formar adecuadamente a los conductores. Éstos controlan la aceleración, frenado, velocidad, rpms, marcha engranada, posición del vehículo en la calzada, por lo que su acción es determinante para lograr mejoras notables en el consumo de combustible y en la eficiencia medioambiental.

Los programas de conducción eficiente orientados a la formación deben conseguir que el conductor sea capaz de acometer las acciones de conducción necesarias, adaptadas al vehículo y la vía, para realizar una conducción segura y con un menor consumo de combustible. La evaluación del programa de conducción eficiente se llevará a cabo mediante el rendimiento de los conductores que lo integran, bien mediante comparaciones simples de consumos de combustible o analizando las acciones de conducción. Como el consumo de combustible depende de muchos factores ajenos a la acción de conducción (temperatura, tráfico, carga, etc), la manera más justa de evaluar el rendimiento de un conductor es mediante el análisis de su conducción (aceleración, deceleración, conducción en inercia, ralentí).

Asimismo, para evaluar de forma objetiva el rendimiento del conductor es necesario registrar todos los eventos del vehículo mediante la lectura de parámetros recogidos de la ECU (Engine Control Unit) a través del bus CAN [1] y almacenados en una base de datos para su análisis posterior.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

El objetivo de este artículo es determinar la técnica óptima de conducción en una ruta y vehículo determinados para minimizar el consumo de combustible. Así, se podrán mejorar las estrategias de formación y evaluación de los profesionales al posibilitar la comparación de la acción de conducción desarrollada con el óptimo para esa ruta y vehículo.

Aplicando técnicas de regresión lineal se ha obtenido un modelo de consumo para el vehículo analizado. Con el modelo de consumo se ha diseñado un algoritmo de optimización de rutas que permite realizar el mismo trayecto en el mismo tiempo, pero minimizando el consumo de combustible. Finalmente, a partir de la ruta óptima se determina la acción de conducción que posibilita el menor consumo de combustible en el trayecto analizado. Los resultados indican que, en el trayecto de 11.1 km, se obtienen ahorros medios de 15 l/100km (-28% de consumo). Para el mes completo, solamente contabilizando las rutas realizadas por el vehículo analizado, se habrían ahorrado 1008 litros de combustible.

Para la realización del trabajo se dispone de los datos reales de 16 compañías del transporte en España y Marruecos, donde se han recogido los datos de 880 conductores. Para el estudio que se presenta se ha seleccionado una compañía de autobuses urbanos debido a que el alto consumo de combustible de estos vehículos y el elevado número de horas diarias de funcionamiento generan un gran impacto en las condiciones medioambientales de la ciudad en la que prestan servicio. La compañía de autobuses opera en una ciudad española de más de 250,000 habitantes. Se ha analizado la información de un autobús urbano no articulado en una línea de transporte en el mes de diciembre de 2015, procesando 1,048,576 muestras con 21 parámetros cada una de ellas.

En la sección II se comentan los trabajos relacionados. En la sección III se describe el sistema de recogida y análisis de la información que será utilizada para el modelo de regresión de la sección IV. El algoritmo de optimización de rutas se implementa en la sección V que se aplicará al caso de estudio de una ruta real en la sección VI. Finalmente, la sección VII muestra las conclusiones y líneas de trabajo futuras.

II. TRABAJOS RELACIONADOS

Actualmente existe un gran número de investigaciones relacionadas con la conducción eficiente y sus beneficios a la hora de reducir el consumo de combustible en flotas profesionales y, por consiguiente, disminuir las emisiones de CO2 al ambiente. Así, Rutty et al. [2] evaluaron el efecto de la formación en conducción eficiente, concluyendo que, gracias a la formación de los conductores en *eco-driving*, las emisiones medias de CO2 se redujeron en 1.7kg por vehículo y día. De la misma forma, los autores en [3], a través de un programa de cursos de formación en Grecia, demostraron que se puede ahorrar hasta un 10% de combustible. La evaluación llevada a cabo por Strömberg and Karlson [4] sobre programas de formación en conducción eficiente en Suecia demostró que las reducciones de consumo alcanzaban el 6,8% para una flota de autobuses. También, mediante técnicas de monitorización continua, Vagg et al. [5] consiguieron reducir el consumo el 7.6%. Otros trabajos en la misma línea, como los de Ferreira et al. [6] para una flota de autobuses en Lisboa, muestran reducciones de consumo de 3 a 5 l/100km aplicando formación en técnicas de conducción eficiente. En España, Rionda et al. [7] obtuvieron reducciones de consumo del 10% aplicando diferentes técnicas de formación a los conductores de una empresa de autobuses urbanos.

Todos los trabajos mencionados presentan ahorros significativos en el consumo de combustible, aplicando técnicas de eco-driving. Sin embargo, el punto de comparación es respecto al consumo antes de la formación. No existe ningún método que permita estimar el máximo ahorro posible ni comparar la técnica de conducción con la ideal. En nuestro trabajo se ha realizado una estimación de la técnica ideal de conducción que permitiría alcanzar el mismo destino en el mismo tiempo de servicio con el mínimo consumo de combustible. Estos resultados pueden aplicarse en los cursos de formación para mejorar las técnicas de eco-driving, además de proporcionar un límite inferior de comparación para el consumo de combustible en unas condiciones determinadas.

Para desarrollar el algoritmo de optimización de rutas es necesario elaborar un modelo de consumo basado en datos reales de funcionamiento del vehículo bajo estudio. En el trabajo de Delgado et al. [8] para una flota de cinco autobuses se demostró que la velocidad, la aceleración y la distancia entre paradas eran suficientes para predecir el consumo de combustible con una precisión razonable. Otros factores como tamaño del vehículo y gradiente de la vía también influyen en el consumo [9]. En el caso presentado en este artículo se va analizar el mismo vehículo en la misma ruta, por lo que el tipo de vehículo y vía se mantienen invariables a lo largo del estudio. Otros factores como la carga del vehículo, tráfico, condiciones meteorológicas o la temperatura también afectarían al consumo. Al no disponer de esta información, en la elaboración del modelo se suponen las mismas condiciones en todos los casos.

Respecto a la optimización del consumo de un vehículo basado en un perfil de velocidad, se han utilizado varios métodos en la literatura [10], [11], [12]. La mayoría de estudios están basados en optimizar la utilización de energía, aunque hay otros en los que el objetivo es minimizar el tiempo de la ruta [13]. En este trabajo hemos seguido la técnica de programación dinámica tomando como función de coste a minimizar el consumo de combustible. De forma similar a Mensing et al. [14], se ha optimizado el consumo en

una ruta en la que el vehículo recorre la misma distancia con las mismas paradas en el mismo tiempo, con un menor consumo de combustible. Otra aportación respecto a otros artículos es que para estimar y minimizar la función de coste se han utilizado los datos reales de consumo de un autobús urbano capturados a través de la ECU del vehículo.

III. Sistema de recogida y análisis de la información

Para la captura de datos de los vehículos, almacenamiento y análisis se ha utilizado el dispositivo CATED BOX[1], desarrollado por la compañía ADN Mobile [15], [16], [17]. La arquitectura general de CATED BOX tiene dos subsistemas con diferentes componentes hardware y software: un sistema empotrado y el sistema central.

Un sistema empotrado lee la información del vehículo a través de la ECU (*Engine Control Unit*) y la envía a un sistema central para ser almacenada y procesada, como se indica en la Fig. 1. El sistema captura una serie de variables de la ECU que dependen de las características particulares del vehículo, entre las que se incluyen velocidad, revoluciones por minuto, aceleración longitudinal, ditancia total, consumo instantáneo, carga motor, peso total, marcha engranada, etc. Esta información recogida de la ECU se complementa con los datos de posición GPS y del acelerómetro.

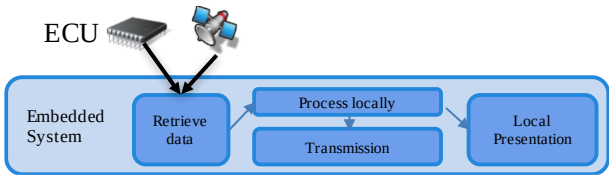


Fig. 1. Sistema empotrado de CATED BOX

El sistema central se encarga de recibir la información del sistema empotrado en los vehículos y almacenarla para su posterior análisis. Como se indica en la Fig. 2, el núcleo del sistema está compuesto por dos bases de datos. Una base de datos no relacional (CouchDB) para almacenar la información proveniente del sistema embarcado y una base de datos relacional (PostgreSQL) donde se almacenan los resúmenes de ruta y otra información complementaria. De forma periódica, se ejecutan varias tareas para leer datos de la base CouchDB, realizar operaciones y almacenar los resultados en la base de datos PostgreSQL. También se incluyen otros detalles proporcionados por las compañías de transporte, como información de vehículos, conductores y las rutas que realizan. Toda esta información complementaria permite correlar la información del rendimiento del vehículo con un conductor o ruta particulares, como en el caso del trabajo que se presenta en este artículo.

El sistema central también incluye una serie de aplicaciones diseñadas para los diferentes perfiles de usuario, como pueden ser:

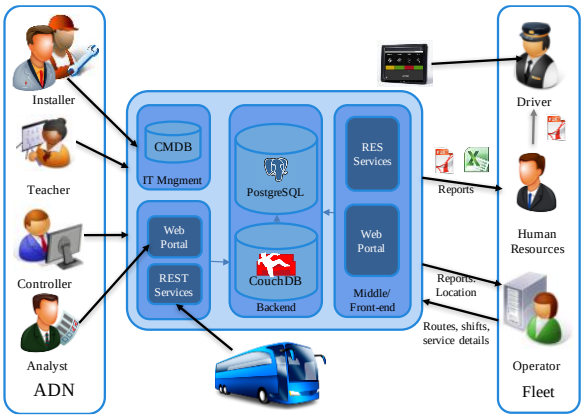


Fig. 2. Sistema central de CATED BOX

- Gestión y control de los dispositivos embarcados e infraestructura
- Generación de informes e indicadores de rendimiento para las compañías de transporte
- Generación de informes para proporcionar soporte a los formadores en conducción eficiente

Debido a las funcionalidades de la herramienta, CATED BOX puede utilizarse en las compañías de transporte para formar y evaluar el rendimiento de los conductores en lo referencte a eficiencia en la conducción. Los estudios realizados en este artículo están enfocados a proporcionar un soporte para la formación en conducción eficiente al indicar cuál sería la acción de conducción óptima par realizar un trayecto, minimizando el consumo de combustible y, por tanto, la emisión de partículas contaminantes al medioambiente. También sería de ayuda en los procesos de evaluación, ya que proporciona un umbral con el que comparar la eficiencia de cada ruta y vehículo.

IV. Modelo de regresión para el cálculo del consumo de combustible

Como paso previo a la construcción del algoritmo que optimiza el consumo en la ruta analizada se va a definir el modelo que permite determinar el consumo en función de la velocidad y la aceleración. Se ha analizado la información de un autobús urbano no articulado en una línea de transporte, procesando 1,048,576 muestras con 21 parámetros.

El análisis comienza con la recogida de datos de la ECU del vehículo mediante un dispositivo on-board diseñado a tal efecto, que se complementa con los datos

[1] http://www.adnmobilesolutions.com/en/catedbox.html

de posición GPS y del acelerómetro. Con toda esta información se compone una traza que está formada por las variables que contienen toda la información sobre el rendimiento del vehículo, recogidas con una periodicidad de 1.5 segundos. El volumen de datos generado es de más de un millón de trazas.

Estos datos sufren un primer procesamiento para obtener los parámetros de cada una de las rutas. A continuación, se hace un filtrado de la información, con el objetivo de identificar aquellas trazas y rutas que no deben entrar a formar parte del modelo por presentar datos anómalos. Así, se eliminan del modelo las rutas y trazas que presenten valores no válidos como consumos instantáneos negativos, rpm fuera del rango de operación del vehículo, velocidades y marchas engranadas anormales, etc.

Para el algoritmo de optimización de la ruta es necesario determinar el consumo instantáneo de combustible en función de la velocidad actual, la velocidad anterior y la aceleración.

$$Consumo_k = f(v_k, v_{k-1}, a_k) \quad (1)$$

Antes de comenzar con la especificación del modelo hay que identificar si existe multicolinealidad, lo que indica relación de dependencia entre las variables predictoras. Cuando hay relación entre las variables predictoras algunas de ellas tienen que ser eliminadas para simplificar el modelo y mejorar su explicación. Una forma de medir la multicolinealidad es por medio de la matriz de correlación. Sin embargo, la multicolinealidad también puede aparecer cuando la variación de una variable predictora se explica por una combinación lineal de otros predictores. En este caso, Variance Inflation Factor (VIF) es el test de multicolinealidad recomendado para detectar relaciones significativas entre las variables independientes. Como indican los autores en [18], los problemas de multicolinealidad aparecen si VIF>4. Se ha calculado el factor VIF para las variables predictoras y no se ha detectado multicolinealidad.

Se han probado diferentes modelos de regresión, llegando a la conclusión que el modelo que mejor representa el consumo de combustible es el siguiente:

$$Consumo_k = \beta_0 + \beta_1 \cdot v_k + \beta_2 \cdot v_{k-1} + \beta_3 \cdot a_k + \beta_4 \cdot (v_k - v_{k-1}) + \beta_5 \cdot v_k^2 + \beta_6 \cdot v_{k-1}^2 + \beta_7 \cdot a_k^2 + \beta_8 \cdot (v_k - v_{k-1})^2 \quad (2)$$

Se ha utilizado el *toolbox* estadístico de MATLAB[2] para implementar las regresiones. El modelo original utiliza 1,039,762 observaciones que se corresponden con el número total de trazas después del proceso de filtrado. La salida del modelo contiene el coeficiente constante β_0 y los coeficientes de las variables predictoras. Los coeficientes se estiman utilizando MLE (*Maximum Likelihood Estimation*) para minimizar el error cuadrático medio entre el vector de predicción $\hat{y}=\beta \cdot f(X)$ y la respuesta real, $\hat{y}-y$. El error RMS del modelo es 1.87.

Los términos del modelo de regresión se indican en la Tabla I. La primera columna indica el término incluido en el modelo, de acuerdo a la ecuación (2). La segunda columna es el valor del coeficiente (βi), con el error estándar (SE) en la siguiente. La columna pValor es uno de los términos más importantes ya que representa el p-valor del estadístico F para el test de hipótesis. En todos los casos es menor que 0.01, indicando que todos los términos considerados son significativos en el modelo, para un nivel de significancia del 1%.

Una vez eliminados los *outliers*, el modelo se ha ajustado utilizando 763,532 observaciones. Una medida de la bondad de ajuste del modelo es el coeficiente de determinación (R^2). Este estadístico indica el nivel de ajuste del modelo a los datos reales a predecir. En este caso, el coeficiente de determinación es de 0.972, próximo a la unidad, lo que sugiere que el modelo explica el 97.2% de la variabilidad de la variable dependiente.

Concluyendo, a partir de los resultados de los diferentes estadísticos puede afirmarse que el modelo representa con exactitud la variabilidad en el consumo de combustible en función de la velocidad y la aceleración consideradas.

Tabla I
COEFICIENTES DEL MODELO DE REGRESIÓN

Coeficiente	Estimación	SE	pValor
β_0	2.495	0.0029	<0.01
β_1	-1.065	0.0027	<0.01
β_2	1.356	0.0028	<0.01
β_3	0.261	0.0084	<0.01
β_4	0.000	0.0000	NaN
β_5	-0.094	5.5309	<0.01
β_6	0.094	5.4540	<0.01
β_7	1.136	0.0066	<0.01
β_8	0.537	0.0004	<0.01

V. ALGORITMO DE OPTIMIZACIÓN DE RUTAS

Para optimizar el consumo de una ruta se ha diseñado un algoritmo para cada ciclo de *eco-driving*. Cada uno de estos ciclos comienza y finaliza en una situación de velocidad cero. De esta forma, una ruta está compuesta por la concatenación de varios ciclos de *eco-driving*. El objetivo del algoritmo es minimizar la función de coste consumo de combustible en cada ciclo, optimizando el perfil de velocidad y aceleración.

El sistema se rige por las siguientes ecuaciones, considerando la distancia recorrida (x), la velocidad (v), aceleración (a) y tiempo (t):

$$x_{k+1} = x_k + v_k \cdot \Delta t + \frac{1}{2} \cdot a_k \cdot \Delta t^2 \quad (3)$$

[2]The Mathworks, Inc. Matlab R2015b: Statistics and Machine Learning Toolbox

$$v_{k+1} = v_k + a_k \cdot \Delta t \quad (4)$$

Se va a utilizar una aproximación tridimensional [14] para conseguir realizar el mismo recorrido en el mismo tiempo minimizando el consumo, de acuerdo con el principio de optimización de Bellman [19]. Así, en el algoritmo cada uno de los estados (X) considerados está compuesto de tres variables, con los estados inicial X_0 y final X_f indicados en la Ec. 5:

$$X = \begin{bmatrix} t \\ x \\ v \end{bmatrix}, X_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, X_f = \begin{bmatrix} t_f \\ x_f \\ 0 \end{bmatrix} \quad (5)$$

Para implementar el algoritmo se sigue el esquema indicado en la Fig. 3, donde se muestran el tiempo final, la distancia final y la limitación de velocidad máxima para ese tramo.

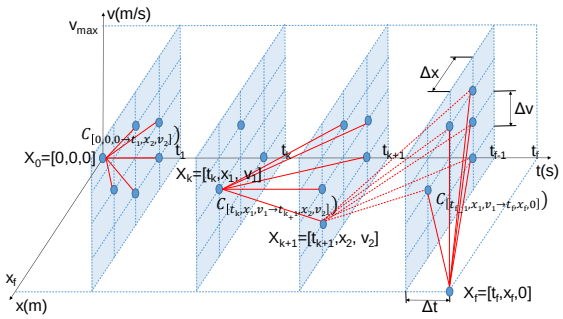


Fig. 3. Cálculo de costes en el método tridimensional

En primer lugar, se recorre el trayecto en sentido inverso, partiendo del estado final X_f hasta llegar al estado inicial X_0, analizando todos los posibles estados en el gráfico tridimensional y almacenando el índice que minimiza el coste energético. Una vez conocida la trayectoria a seguir se recorre nuevamente el trayecto desde el estado inicial al final siguiendo el camino de mínimo consumo y calculando los perfiles de velocidad, aceleración y consumo obtenidos.

Partiendo del estado final, en cada iteración del algoritmo se calcula la función de coste (consumo, de acuerdo a las Ec. 1, 2) en base a las velocidades v_2, v_1 y la aceleración a_1. Se analizan todas las opciones posibles $[t_{k+1},x_2,v_2]$ para alcanzar el estado $[t_k,x_1,v_1]$ y se almacena el índice que minimiza la función de coste, de acuerdo con la Ec. 6.

$$C_{[t_k,x_1,v_1]} = \min_{x2,v2}(C_{[t_{k+1},x_2,v_2]} + C_{[t_k,x_1,v_1 \to t_{k+1},x_2,v_2]}) \quad (6)$$

Se indican a continuación los resultados de la ejecución del algoritmo para un tramo ejemplo de 200m, en un tiempo de 30s, con límite de velocidad máxima de 10m/s y límite de aceleración de $\pm 3 m/s^2$. En la Fig.4 se muestra una representación tridimensional la evolución de las variables tiempo, distancia y velocidad. En las siguientes figuras puede apreciarse cómo para recorrer el trayecto con el menor consumo posible se incrementa progresivamente la velocidad con una aceleración suave, que no supere los límites establecidos para el confort y seguridad de los pasajeros, hasta llegar a una zona en la que la velocidad se mantiene constante. Posteriormente, se aprovecha la inercia del vehículo, para obtener un consumo nulo hasta la detención. En el tramo analizado, el consumo total ha sido de 0.0862 litros, con una media de 43.133 litros/100km.

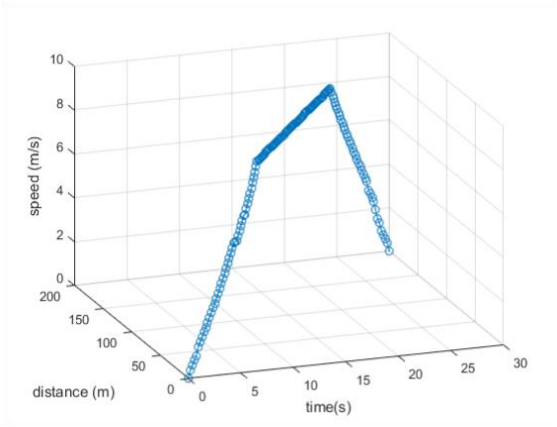


Fig. 4. Evolución del estado [t,x,v] en el tramo analizado

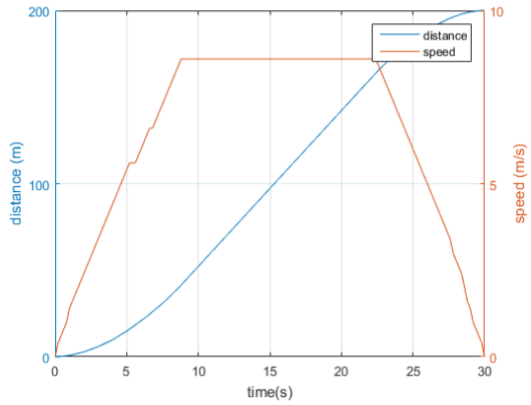


Fig. 5. Distancia recorrida y perfil de velocidad

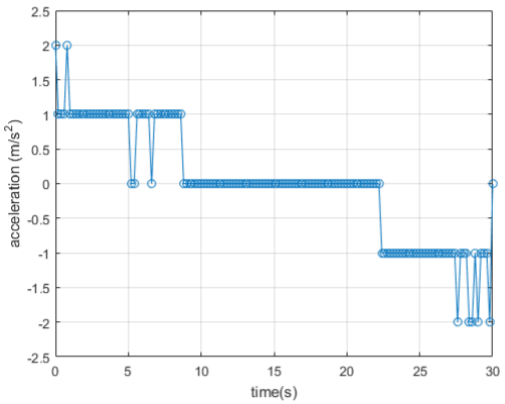


Fig. 6. Perfil de aceleración

VI. CASO DE ESTUDIO: COMPARACIÓN CON RUTA REAL

Se ha seleccionado aleatoriamente una ruta para hacer la comparación con el trayecto óptimo que optimizaría el consumo de combustible. La ruta tiene una longitud de 11.1Km y está compuesta de un total de 34 ciclos de *eco-driving* debidos a las paradas del servicio, semáforos, stops y otras posibles detenciones provocadas por las características del tráfico. En la Fig. 7 se indican los perfiles de distancia y velocidad para los ciclos de *eco-driving* en la ruta seleccionada.

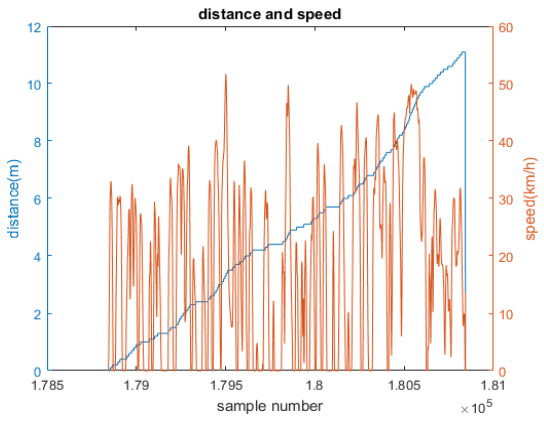


Fig. 7. Perfiles de distancia y velocidad en la ruta seleccionada

Se han identificado los ciclos de *eco-driving* a partir del perfil de velocidad y se ha ejecutado el algoritmo de optimización para cada uno de ellos. La duración de los ciclos y la distancia de cada uno de ellos puede apreciarse en la Fig. 8. Los ciclos tienen una duración media de 50,33 segundos y la distancia media es de 336,36 metros.

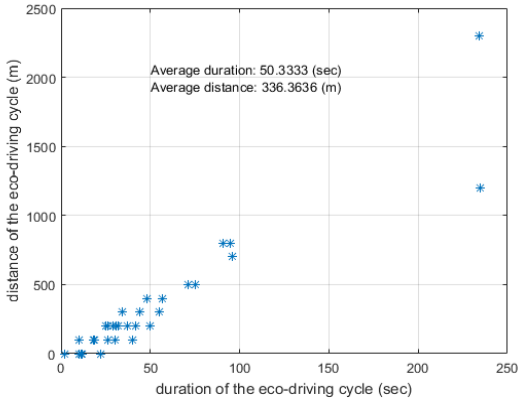


Fig. 8. Duración y distancia de los ciclos de *eco-driving*

En las figuras Fig. 9-12 se muestra una comparación entre los datos reales y los resultados de la simulación para los perfiles de velocidad y aceleración. Puede apreciarse una menor variabilidad en los perfiles de conducción óptima lo cual es indicador, evidentemente, de una conducción más moderada, tratando de mantener una velocidad constante una vez alcanzado el máximo en el ciclo correspondiente.

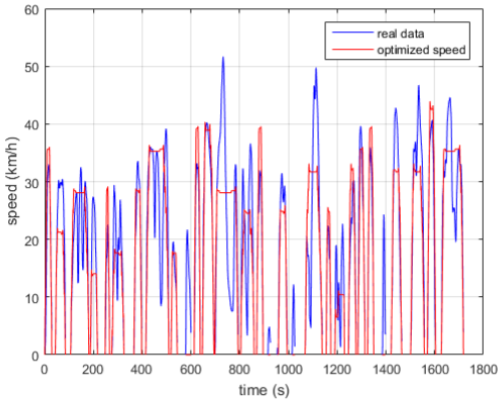


Fig. 9. Comparación de velocidades de ruta real y optimizada

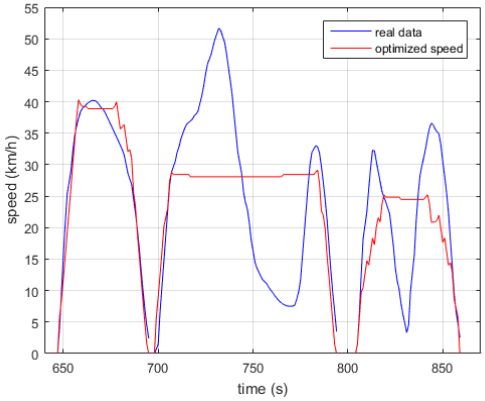


Fig. 10. Ampliación de velocidad en los ciclos de *eco-driving* #11,12,13

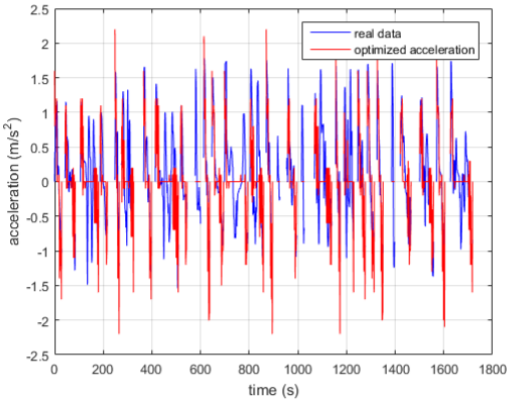


Fig. 11. Comparación de aceleraciones de ruta real y optimizada

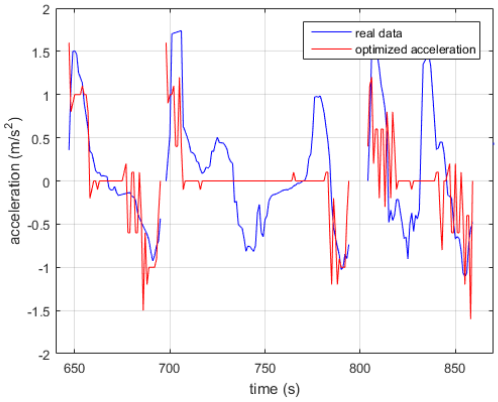


Fig. 12. Ampliación de aceleración en los ciclos de *eco-driving* #11,12,13

En el total de la ruta de 11.1 Km, el consumo real ha sido de 5,78 litros, con un promedio de 52,14 litros/100km. Con la ruta optimizada el consumo total habría sido de 4,13 litros, con un promedio de 37,22 litros/100km. Estos datos indican que se podría haber reducido el consumo de combustible el 28,61%. Ha de tenerse en cuenta que los resultados de la simulación se dan en condiciones óptimas de conducción, sin tráfico ni otros condicionantes externos como las condiciones meteorólogicas, estado de la vía, carga del vehículo, etc.

Extrapolando los resultados a todo el mes, en el que el vehículo analizado completó 611 rutas, la compañía podrían haber ahorrado 1008,2 litros de combustible. Para el total de la flota, donde los diferentes conductores de la empresa de transporte recorren miles de kilómetros, el ahorro de combustible sería de gran importancia para la compañía ya que supondría una importante reducción de costes y, por supuesto, redundaría en una mayor calidad medioambiental en la ciudad en la que presta servicio. Estos resultados avalan el uso de técnicas de conducción eficiente que, sin requerir de importantes inversiones en nuevos vehículos y tecnología, pueden conseguir resultados notables y mejoras medioambientales mediante cursos y estrategias de formación que optimicen el uso de la flota existente.

Para completar la formación de los conductores y servir de soporte a gerentes de flotas y formadores en conducción eficiente se dispone de la herramienta VAT-ED[3] en la que, entre otras funcionalidades, se muestran gráficamente los patrones de conducción efiente e ineficiente en los diferentes puntos de la ruta, como se muestra en la Fig. 13.

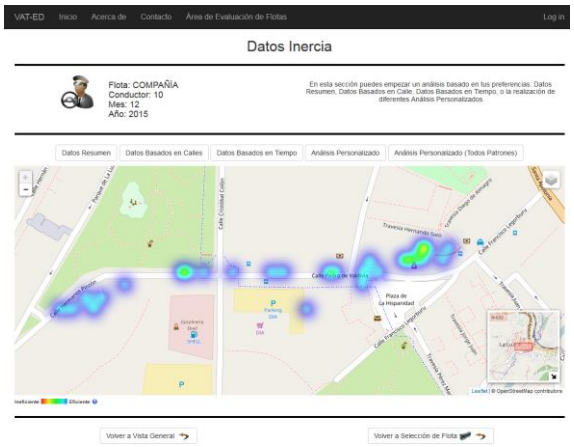


Fig. 13. Patrón de conducción en una zona del recorrido

Mediante esta herramienta gráfica de análisis, que permite seleccionar la flota, el conductor, la línea, el vehículo y el patrón de conducción se puede determinar la acción de conducción real ejecutada en un punto del recorrido. Complementando la información con los resultados de optimización mostrados en este artículo se potenciarían las técnicas de formación y con ello mejorarían los resultados en conducción eficiente de la compañía.

VII. Conclusiones

En este artículo se ha demostrado que la conducción eficiente es una alternativa válida para reducir el consumo de combustible y, por consiguiente, la emisión de partículas nocivas al medioambiente. Esta reducción se puede lograr sin necesidad de realizar fuertes inversiones para renovar y/o mejorar la flota de vehículos existente.

El estudio realizado se ha centrado en el conductor y en él se pretende reforzar la importancia de una buena formación y motivación de los conductores para adoptar técnicas de conducción eficiente. Los algoritmos utilizados en este trabajo permiten determinar la acción óptima de conducción, de manera que se minimiza el consumo en una ruta determinada. Los resultados son de aplicabilidad directa en las estrategias de formación de conductores en conducción eficiente, ya que completan las ya existentes proporcionando una referencia para comparar la acción de conducción real con la óptima para esa ruta y vehículo. Asimismo, los resultados de este trabajo permiten disponer de un umbral óptimo de referencia para evaluar objetivamente las prestaciones relacionadas con la eficiencia de un determinado conductor. Otro campo en el que puede resultar interesante el estudio es el de la conducción automática. Así, los futuros vehículos sin conductor dispondrían de un patrón de aceleración/deceleración que optimizaría el perfil de velocidad haciendo mínimo el consumo en el trayecto a cubrir.

[3] *Visual Analytics Tool for Evaluation of Drivers*, Copyright O-71-2017

Con los análisis realizados se ha demostrado que el ahorro de combustible puede alcanzar el 28%, cifra significativa si se tienen en cuenta las grandes distancias recorridas por todos los conductores de la compañía a lo largo del año. El vehículo analizado completó 611 rutas a lo largo del mes con lo que, extrapolando los resultados obtenidos, la compañía podría haber ahorrado 1008,2 litros de combustible. No solo es importante el ahorro de combustible, sino que también esa reducción redunda en una atmósfera menos contaminada y en una mejor calidad de vida de la población donde se presta el servicio.

Por otro lado, ha de tenerse en cuenta que los resultados de la simulación se dan en condiciones óptimas de conducción, sin tráfico ni otros condicionantes externos como las condiciones meteorológicas, estado de la vía, carga del vehículo, etc que tienen influencia en el consumo y no se han considerado en esta investigación. No es fácil disponer de toda esta información ya que depende de diferentes organismos e incluso el propio vehículo no dispone en ocasiones de los sensores necesarios. Los trabajos futuros irán encaminados a mejorar el modelo de consumo incluyendo todos estos factores.

Aunque el vehículo analizado es de cambio automático, se está trabajando en la mejora del modelo de consumo incluyendo las marchas engranadas y umbrales de cambio, de manera que se obtenga un modelo para cada marcha, probablemente mucho más preciso que el genérico presentado en este trabajo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la Comisión Europea (proyecto GlobalBLED SMEInst-10-2016-2017), el Plan Nacional de Investigación (proyecto MINECO-13-TIN2013-41749-R) y el Plan PCTI del Principado de Asturias (proyecto GRUPIN-14-065).

REFERENCIAS

[1] M. Di Natale, H. Zeng, P. Giusto, y A. Ghosal, *Understanding and Using the Controller Area Network Communication Protocol*. New York, NY: Springer New York, 2012.

[2] M. Rutty, L. Matthews, J. Andrey, y T. D. Matto, «Eco-driver training within the City of Calgary's municipal fleet: Monitoring the impact», *Transp. Res. Part Transp. Environ.*, vol. 24, pp. 44-51, oct. 2013.

[3] M. Zarkadoula, G. Zoidis, y E. Tritopoulou, «Training urban bus drivers to promote smart driving: A note on a Greek eco-driving pilot program», *Transp. Res. Part Transp. Environ.*, vol. 12, n.º 6, pp. 449-451, ago. 2007.

[4] H. K. Strömberg y I. C. M. Karlsson, «Comparative effects of eco-driving initiatives aimed at urban bus drivers – Results from a field trial», *Transp. Res. Part Transp. Environ.*, vol. 22, pp. 28-33, jul. 2013.

[5] C. Vagg, C. J. Brace, D. Hari, S. Akehurst, J. Poxon, y L. Ash, «Development and Field Trial of a Driver Assistance System to Encourage Eco-Driving in Light Commercial Vehicle Fleets», *IEEE Trans. Intell. Transp. Syst.*, vol. 14, n.º 2, pp. 796-805, jun. 2013.

[6] J. C. Ferreira, J. de Almeida, y A. R. da Silva, «The Impact of Driving Styles on Fuel Consumption: A Data-Warehouse-and-Data-Mining-Based Discovery Process», *IEEE Trans. Intell. Transp. Syst.*, vol. 16, n.º 5, pp. 2653-2662, oct. 2015.

[7] A. Rionda *et al.*, «Blended learning system for efficient professional driving», *Comput. Educ.*, vol. 78, pp. 124-139, sep. 2014.

[8] O. F. Delgado, N. N. Clark, y G. J. Thompson, «Modeling Transit Bus Fuel Consumption on the Basis of Cycle Properties», *J. Air Waste Manag. Assoc.*, vol. 61, n.º 4, pp. 443-452, abr. 2011.

[9] E. Demir, T. Bektaş, y G. Laporte, «A comparative analysis of several vehicle emission models for road freight transportation», *Transp. Res. Part Transp. Environ.*, vol. 16, n.º 5, pp. 347-357, jul. 2011.

[10] Y. Saboohi y H. Farzaneh, «Model for developing an eco-driving strategy of a passenger vehicle based on the least fuel consumption», *Appl. Energy*, vol. 86, n.º 10, pp. 1925-1932, oct. 2009.

[11] E. Hellström, J. Åslund, y L. Nielsen, «Design of an efficient algorithm for fuel-optimal look-ahead control», *Control Eng. Pract.*, vol. 18, n.º 11, pp. 1318-1327, nov. 2010.

[12] B. Saerens, J. Vandersteen, T. Persoons, J. Swevers, M. Diehl, y E. Van den Bulck, «Minimization of the fuel consumption of a gasoline engine using dynamic optimization», *Appl. Energy*, vol. 86, n.º 9, pp. 1582-1588, sep. 2009.

[13] E. Velenis y P. Tsiotras, «Optimal Velocity Profile Generation for Given Acceleration Limits; The Half-Car Model Case», en *Proceedings of the IEEE International Symposium on Industrial Electronics, 2005. ISIE 2005.*, 2005, vol. 1, pp. 361-366.

[14] F. Mensing, R. Trigui, y E. Bideaux, «Vehicle trajectory optimization for application in ECO-driving», en *2011 IEEE Vehicle Power and Propulsion Conference*, 2011, pp. 1-6.

[15] A. Rionda Rodriguez, D. Martinez Alvarez, X. G. Paneda, D. Arbesu Carbajal, J. E. Jimenez, y F. Fernandez Linera, «Tutoring System for the Efficient Driving of Combustion Vehicles», *IEEE Rev. Iberoam. De Tecnol. Aprendiz. RITA*, vol. 8, n.º 2, pp. 82-89, may 2013.

[16] A. Rionda *et al.*, «UrVAMM #x2014; A full service for environmental-urban and driving monitoring of professional fleets», en *2013 International Conference on New Concepts in Smart Cities: Fostering Public and Private Alliances (SmartMILE)*, 2013, pp. 1-6.

[17] A. G. Pañeda *et al.*, «An Archictecture for a Learning Analytics System Applied to Efficient Driving», *IEEE Rev. Iberoam. Tecnol. Aprendiz. RITA*, vol. 11. n.º 3, pp. 137-145, july 2016.

[18] K. Mokhtar y M. Z. Shah, «A regression model for vessel turnaround time», *Tokyo Academic Industry & Culture Integration Tour*, pp. 10-19, 2006.

[19] D. E. Kirk, *Optimal Control Theory: An Introduction*. Mineola, N.Y: Dover Publications Inc., 2004.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6580

Modelo de colas con *vacations* aplicado a un sistema de captura de paquetes

Luis Zabala, Armando Ferro, Ander Nieva.
Departamento de Ingeniería de Comunicaciones.
Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU).
ESI Bilbao. Alameda de Urquijo s/n. 48013 Bilbao.
luis.zabala@ehu.eus, armando.ferro@ehu.eus, ander.nieva@ehu.eus.

***Resumen*—La mejora de sistemas de captura de paquetes de red ha sido extensamente cubierta como tema de investigación en los pasados años. La mayoría de estas iniciativas han sido respaldadas por evaluaciones experimentales; sin embargo, ha habido pocas propuestas de modelado. Este trabajo presenta el modelado y el análisis de un sistema de cola finito con *vacations* aplicado a la etapa de captura de paquetes de un sistema de monitorización de red. Se plantean dos modelos con disciplina de servicio diferente (exhaustiva y limitada) y se evalúan sus rendimientos, principalmente en forma de throughput de captura, para distintos escenarios. Éstos contemplan diferentes tasas de entrada de paquetes y tiempos de *vacation*. Los resultados teóricos, derivados de un estudio analítico basado en ecuaciones de balance y su desarrollo en forma matricial, también son comparados con los de una sonda real de tráfico de red que captura paquetes.**

***Palabras Clave*—motor de captura, vacation, teoría de colas**

I. INTRODUCCIÓN

Un sistema finito de colas con *vacations* puede ser útil para modelar y analizar el rendimiento de un sistema de captura de tráfico situado en un entorno de monitorización de red. En un modelo de colas clásico, los servidores siempre están disponibles. Sin embargo, puede haber sistemas de colas reales donde los servidores pueden dejar de estar disponibles durante un cierto periodo de tiempo debido a razones varias. Para analizar estos sistemas, se introduce el estado de *vacation* [1], que representa el periodo de la ausencia temporal del servidor. Por tanto, en este tipo de sistemas, después de cada periodo activo en el que el servidor atiende a los elementos de la cola principal, el servidor pasa a ejecutar tareas adicionales que no están relacionadas con los clientes de la cola principal [2].

En todo sistema de monitorización de red se tiene una primera etapa de captura de paquetes donde se recoge de la red, "en bruto", los datos de medida [3]. Posteriormente esos datos pueden ser analizados en detalle y procesados para extraer de ellos interpretaciones de nivel superior. Finalmente, los resultados extraídos del análisis son presentados en diferentes formatos a los operadores de red. Las compañías de telecomunicaciones invierten grandes cantidades de dinero en monitorización de tráfico con el objetivo de garantizar la satisfacción de sus clientes y, al mismo tiempo, hacer crecer su cuota de mercado [4].

El objeto de estudio de este trabajo es el modelado de la etapa de captura de paquetes, teniendo presente que el sistema puede tener funciones adicionales de monitorización que realizar. Por ello, se propone un modelado de colas con *vacation*, donde la tarea de captura se representa mediante el servicio de la cola principal y las otras funciones que podría llevar a cabo el sistema son ejecutadas por el mismo procesador durante sus tiempos de *vacation*. Estos modelos ayudan a estimar el rendimiento de la etapa de captura y ver cómo influye sobre él los tiempos de *vacation*, es decir, la dedicación a otras tareas. Este modelado cobra mayor interés cuando el sistema se encuentra en condiciones de saturación, por ejemplo, cuando las tasas de transmisión de la red crecen.

En el ámbito de la captura de paquetes, el incremento de las tasas de transmisión de datos es continuo. Esto hace que la implementación de sistemas de monitorización capaces de afrontar este ritmo creciente de las redes sea una ardua labor, incluso para el caso en el que las aplicaciones que se ejecutan en la capa superior del sistema de monitorización llevan a cabo un procesamiento mínimo. La monitorización de tráfico en el rango entre 100 Mbps y 1 Gbps fue considerada todo un reto hace muy pocos años, mientras que los routers comerciales actuales ya tienen interfaces de 10, 40 ó incluso 100 Gbps.

Por otro lado, es un hecho que, en los últimos años, se ha producido una mejora en el rendimiento de los sistemas que procesan tráfico de red implementados sobre hardware de propósito general. Esta mejora se debe tanto a las mejoras de software como a la evolución de hardware (aumento de velocidades de las tarjetas de red, aumento de la frecuencia de reloj de la CPU, aparición de CPUs multinúcleo, nuevas características de las tarjetas de red que hacen posible ahorrar ciclos de CPU, etc.) [5].

Así, ha habido varias propuestas de motores de captura basados en este tipo de arquitecturas como PF_RING [6], PacketShader [7], Netmap [8], PFQ [9], OpenOnLoad [10] y HPCAP [11]. Todo esto hace que los sistemas basados en hardware genérico sean muy atractivos para la monitorización de tráfico de redes de alta velocidad, puesto que su rendimiento puede ser comparado con el de hardware especializado (FPGAs, *network processors*, o soluciones comerciales proporcionadas por fabricantes de routers) y, además, su precio es significativamente menor.

Sin embargo, existen pocas propuestas de modelado para estas soluciones de captura. Entre ellas, se pueden destacar algunas relacionadas con la caracterización de sistemas basados en Linux [12] y otras con sistemas de captura y análisis de tráfico [13].

En definitiva, la existencia de numerosas propuestas de motores de captura, pero la escasez de propuestas de modelado y de análisis del caso en el que otras tareas pueden influir en el rendimiento del sistema de captura, motivan este trabajo donde se pretende aplicar un modelo de cola finita con *vacations* a un sistema de captura de paquetes. Además, tal y como se verá, también se propone un modelo para el caso de un sistema Linux, debido a su carácter abierto y por ser la base de muchos motores de captura sobre hardware de propósito general. Parece de interés construir modelos matemáticos que permitan estudiar teóricamente el rendimiento de sistemas de captura de tráfico, con objeto de contribuir en futuras mejoras de diseño destinadas a este tipo de sistemas.

El resto del artículo está estructurado de la siguiente forma. La sección II presenta un modelo para un sistema de captura genérico con disciplina de servicio exhaustiva, mientras que el modelo expuesto en la sección III es más específico ya que recoge las características de un sistema de captura de paquetes basado en Linux. La sección IV proporciona resultados de los modelos. Finalmente, la sección V expone las conclusiones.

II. MODELO CON *VACATIONS* Y DISCIPLINA DE SERVICIO EXHAUSTIVA

El primer modelo que se presenta, M1, se basa en un modelo de cola finita con una serie de hipótesis markovianas. La llegada de paquetes al sistema de captura sigue un proceso de Poisson con tasa λ. Estos paquetes son los que provienen de la tarjeta de red y son transferidos, vía DMA (Direct Memory Access), a un área de memoria de tamaño finito. A continuación, el tratamiento de paquetes por parte del motor de captura se representa mediante un único servidor que se dedica a esta tarea en su periodo activo y a otras en su periodo de *vacation*. El tiempo dedicado a capturar paquetes sigue una distribución exponencial de media $1/\mu_S$, mientras que los tiempos de *vacation* una distribución exponencial de media $1/\mu_V$. Se denominará N al número máximo de paquetes que puede haber en la etapa de captura. Esto tiene en cuenta el paquete que está siendo atendido por el procesador más los que están en la cola de espera. Si hay N paquetes en el sistema de captura, los nuevos paquetes entrantes serán bloqueados, es decir, dichos paquetes no podrán introducirse en el búfer, ya que éste está completo, y serán rechazados. Por último, el proceso de captura no terminará hasta que se vacíe el búfer, momento en el que comenzará un periodo de *vacation*. Por tanto, se considera una disciplina de servicio exhaustiva.

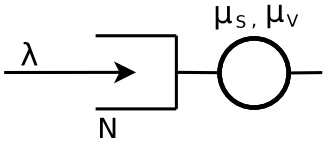


Fig. 1. Modelo de cola finita con *vacations*

A. Cadena de Markov asociada a la etapa de captura

Este primer modelo de cola finita con *vacations* se basa en una cadena de Markov con un espacio de estados $S = \{(n,m), 0 \leq n \leq N, 0 \leq m \leq 1\}$. La variable n indica el número de paquetes que hay en la etapa de captura; m identifica si el servidor está en un periodo activo capturando paquetes (en cuyo caso, $m = 1$), o si está en un periodo de *vacation* ($m = 0$).

La Fig. 2 muestra el diagrama de transiciones de estado en función de las tasas λ, μ_S y μ_V. Primeramente, hay que indicar que si el sistema está vacío, estado $(0,0)$, y llega un paquete nuevo, debido a la política del planificador de tareas, el inicio del proceso de captura no es inmediato, sino que se produce después de una *vacation*. Por este motivo, se da la transición de $(0,0)$ a $(1,0)$ con tasa λ. Durante la *vacation* el proceso está en los estados $(n,0)$, donde el número de paquetes aumenta con tasa λ, salvo en $(N,0)$ por haber bloqueo, y también es posible terminar la *vacation* con tasa μ_V cuando el planificador lo decida.

En los estados $(n,1)$, el procesador está capturando paquetes. Entonces, pueden llegar nuevos paquetes con tasa λ, salvo en $(N,1)$, y salir paquetes del sistema de captura con tasa μ_S. La finalización del proceso de captura sólo se da cuando se vacía el búfer: transición de $(1,1)$ a $(0,0)$ en la Fig. 2.

B. Ecuaciones de balance y probabilidades de estado

Sea $\pi_{n,m}$, la probabilidad del estado (n,m) en régimen estacionario. Se plantean las siguientes ecuaciones de balance [14] asociadas a los estados de la Fig. 2.

$$\lambda\pi_{0,0} = \mu_S\pi_{1,1} \tag{1a}$$

$$(\lambda+\mu_V)\pi_{n,0} = \lambda\pi_{n-1,0}, \quad 1 \leq n \leq N-1 \tag{1b}$$

$$(\lambda+\mu_S)\pi_{n,1} = \lambda\pi_{n-1,1} + \mu_V\pi_{n,0} + \mu_S\pi_{n+1,1}, \quad 1 \leq n \leq N-1 \tag{1c}$$

$$\mu_V\pi_{N,0} = \lambda\pi_{N-1,0} \tag{1d}$$

$$\mu_S\pi_{N,1} = \mu_V\pi_{N,0} + \lambda\pi_{N-1,1} \tag{1e}$$

Por otro lado, utilizando el principio de balance global [15], por el que el flujo saliente de un conjunto de estados es igual al flujo entrante a dicho conjunto

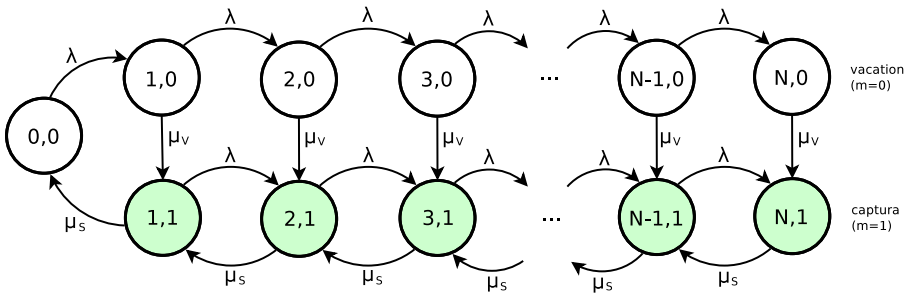


Fig. 2. Cadena de Markov del modelo M1

$$\lambda(\pi_{n,0}+\pi_{n,1})=\mu_S\pi_{n+1,1},\quad 0\le n\le N-1 \tag{2}$$

Y dado que la suma de probabilidades debe ser 1.

$$\sum_{n=0}^{N}(\pi_{n,0}+\pi_{n,1})=1 \tag{3}$$

Tras agrupar y reescribir las ecuaciones (1a)-(1e) y (2), se llega a una solución geométrica matricial

$$\begin{pmatrix}\pi_{n,0}\\ \pi_{n,1}\end{pmatrix}=R^n\begin{pmatrix}\pi_{0,0}\\ 0\end{pmatrix}\quad 1\le n\le N \tag{4}$$

$$\text{con}\quad R=\begin{pmatrix}\lambda/(\lambda+\mu_V) & 0\\ \lambda/\mu_S & \lambda/\mu_S\end{pmatrix}$$

Imponiendo la condición de normalización (3), se obtiene el valor de $\pi_{0,0}$.

$$\pi_{0,0}=\frac{1}{S};\quad S=\begin{pmatrix}1 & 1\end{pmatrix}\left(\sum_{n=0}^{N}R^N\right)\begin{pmatrix}1\\ 0\end{pmatrix} \tag{5}$$

$$\text{siendo}\quad \sum_{n=0}^{N}R^N=(I-R)^{-1}\left(I-R^{N+1}\right)$$

Después de calcular $\pi_{0,0}$ ya se pueden determinar todas las probabilidades de estado $\pi_{n,m}$ aplicando (4).

C. Parámetros de rendimiento

Una vez conocidas las probabilidades de estado en régimen estacionario, $\pi_{n,m}$, ya es posible calcular los parámetros de rendimiento de la etapa de captura: probabilidad de bloqueo, throughput, utilización de CPU en captura, disponibilidad de CPU, etc.

Probabilidad de bloqueo (P_B). Es la probabilidad de perder paquetes cuando el búfer está lleno. Esto sucede en los estados $(N,0)$ y $(N,1)$.

$$P_B=\pi_{N,0}+\pi_{N,1} \tag{6}$$

Throughput de captura, (X_C). Se define como el número medio de paquetes capturados por segundo. La captura tiene lugar en los estados $(n,1)$.

$$X_C=\mu_S\sum_{n=1}^{N}\pi_{n,1}=\lambda\left(1-P_B\right) \tag{7}$$

Frecuencia de *captura* (f_C) Mide el número de procesos de captura que se ejecutan por unidad de tiempo. La frecuencia f_C está relacionada con el ciclo compuesto por un periodo activo de captura y una *vacation*. Teniendo en cuenta el carácter poissoniano de las llegadas de paquetes, llamando T_C a la duración media de un proceso de captura y T_{ciclo} al tiempo medio del ciclo

$$T_{ciclo}=\frac{1}{\lambda}+\frac{1}{\mu_V}+T_C \tag{8}$$

Y la frecuencia de captura queda

$$T_{ciclo}\cdot\pi_{0,0}=\frac{1}{\lambda}\Rightarrow f_C=\frac{1}{T_{ciclo}}=\lambda\pi_{0,0} \tag{9}$$

Tiempo medio de captura (T_C). Es la duración media de un periodo activo del procesador responsable de la etapa de captura. Su expresión se obtiene de la siguiente manera.

$$T_C=T_{ciclo}\sum_{n=1}^{N}\pi_{n,1}=\frac{\sum_{n=1}^{N}\pi_{n,1}}{f_C} \tag{10}$$

III. MODELO CON *VACATIONS* Y DISCIPLINA DE SERVICIO LIMITADA

El segundo modelo propuesto, M2, tiene como objeto representar un sistema más específico y se ha tomado como ejemplo el sistema de captura de paquetes de Linux. Aquí, el proceso que lleva a cabo esta tarea se denomina *softirq* y tiene un límite denominado *budget* que establece el número máximo de paquetes que pueden ser tratados en una *softirq*. Por ello, si se alcanza el *budget*, el procesador deja la *softirq* y pasa a ejecutar otra tarea [18]. Para representar este comportamiento, se propone el modelo M2 basado en una disciplina de servicio limitada, en contraposición a la exhaustiva del modelo M1. El resto de suposiciones coinciden con las del modelo M1. Por tanto, se mantienen las tasas λ, μ_S y μ_V, así como el tamaño de buffer N y se añade el parámetro B, que denota al valor del *budget*.

Un segundo elemento diferencial de M2 es que se tienen en consideración los tiempos consumidos en en rutinas de atención a *hardirq* de Linux [18]. Se supone que el tiempo de ejecución de una rutina asociada a *hardirq* sigue una distribución exponencial con media $1/\mu_H$.

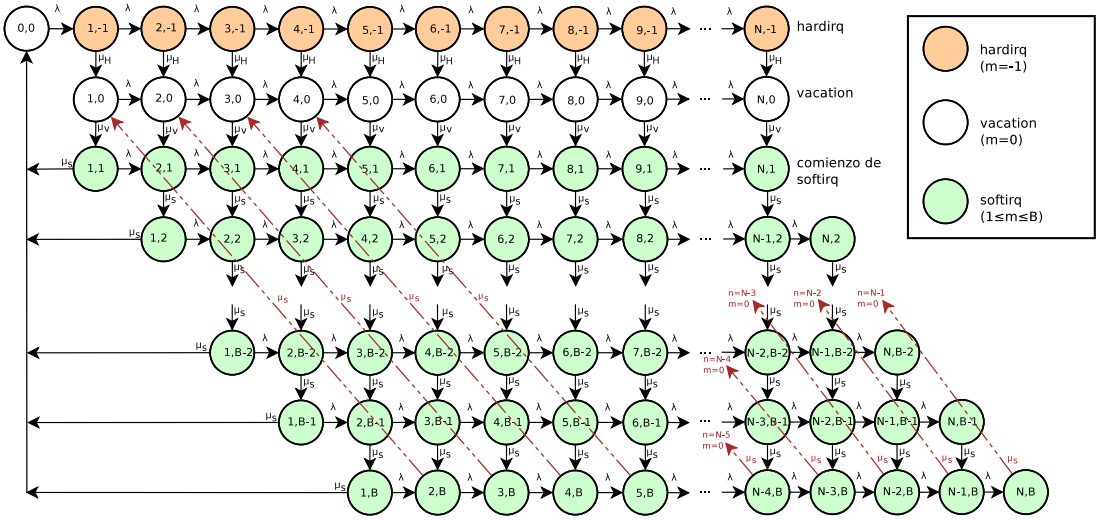


Fig. 3. Diagrama de estados del modelo M2

A. Cadena de Markov asociada a la etapa de captura

M2, modelo de cola finita con *vacations* y disciplina de servicio limitada, se fundamenta en una cadena de Markov con $S = \{(n,m), 0 \leq n \leq N, -1 \leq m \leq B\}$, donde n indica el número de paquetes de la etapa de captura y m identifica diferentes estados del procesador. Así, $m = -1$ indica que el procesador está ejecutando la rutina de servicio de una *hardirq*; $m = 0$ indica que está en *vacation*; y si $m \in [1, B]$, ello indica que el procesador está ejecutando una *softirq* y, concretamente, que se está capturando el paquete número *m* en dicha *softirq*. El valor máximo de m es el valor del *budget*, B.

El diagrama de estados y transiciones de la Fig. 3 refleja el procedimiento del sistema de captura de Linux. Primero, si el sistema está vacío, estado $(0,0)$, y llega un paquete, según el procedimiento de Linux, la llegada de este primer paquete provoca la ejecución de la rutina de atención de *hardirq*. Por ello, en la Fig. 3 se tiene la transición de $(0,0)$ a $(1,-1)$ con tasa λ. La rutina de *hardirq* suele ser corta y, en ella, se deshabilitan las *hardirq* de red, para evitar que llegadas posteriores de paquetes vuelvan a interrumpir al procesador, y se planifica la ejecución de una *softirq*, que tendrá lugar más adelante. Durante la *hardirq*, el sistema está en los estados $(n,-1)$. Pueden llegar nuevos paquetes con tasa λ, exceptuando en $(N,-1)$ o puede finalizarse la propia *hardirq* con tasa μ_H. Esto último implica una transición hacia un estado de *vacation* $(n,0)$. El comportamiento de las *vacation* de M2 es idéntico a las de M1, por lo que, cuando el planificador de tareas lo determina, acaba la *vacation* con tasa μ_V y comienza la ejecución de una *softirq*. Esto último conlleva la transición de un estado $(n,0)$a un estado $(n,1)$.

La ejecución de una *softirq* está representada por el conjunto de estados (n,m) donde $1 \leq n \leq N$ y $1 \leq m \leq B$. Durante la *softirq*, pueden llegar nuevos paquetes con tasa λ y, por tanto, producir transiciones de (n,m) a $(n+1,m)$, salvo en los estados de bloqueo (N,m). También es posible la salida de paquetes procesados, lo que conlleva transiciones de (n,m) a $(n-1,m+1)$. Otro aspecto importante a considerar es la finalización de la *softirq* para la que hay dos opciones. Una es que la cola se vacía y, según el procedimiento de Linux, se habilitan de nuevo las *hardirq*. En el modelo, este caso corresponde a la transición de un estado $(1,m)$ al estado $(0,0)$. La segunda opción es que se alcanza el *budget*, estado (n,B), caso en el que Linux planifica una nueva *softirq* sin deshabilitar las *hardirq* y, en el modelo, se pasa de un estado (n,B) a un estado $(n-1,0)$, es decir, a un estado de *vacation*.

B. Ecuaciones de balance y probabilidades de estado

En régimen estacionario, las probabilidades $\pi_{n,m}$ satisfacen las ecuaciones de balance de los estados de la Fig. 3. En primer lugar, en el estado $(0,0)$,

$$\lambda \pi_{0,0} = \mu_S \sum_{m=1}^{B} \pi_{1,m} \tag{11}$$

En los estados $(1,m)$,

$$(\lambda + \mu_H)\,\pi_{1,-1} = \lambda \pi_{0,0} \tag{12a}$$
$$(\lambda + \mu_V)\,\pi_{1,0} = \mu_H \pi_{1,-1} + \mu_S \pi_{2,B} \tag{12b}$$
$$(\lambda + \mu_S)\,\pi_{1,1} = \mu_V \pi_{1,0} \tag{12c}$$
$$(\lambda + \mu_S)\,\pi_{1,m} = \mu_S \pi_{2,m-1} \tag{12d}$$
$$2 \leq m \leq B-1$$
$$(\lambda + \mu_S)\,\pi_{1,B} = \mu_S \pi_{2,B-1} \tag{12e}$$

Para los estados (n,m) con $2 \leq n \leq N-1$,

$$(\lambda+\mu_H)\,\pi_{n,-1} = \lambda\pi_{n-1,-1} \tag{13a}$$
$$(\lambda+\mu_V)\,\pi_{n,0} = \lambda\pi_{n-1,0} + \mu_H\pi_{n,-1} + \mu_S\pi_{n+1,B} \tag{13b}$$
$$(\lambda+\mu_S)\,\pi_{n,1} = \lambda\pi_{n-1,1} + \mu_V\pi_{n,0} \tag{13c}$$
$$(\lambda+\mu_S)\,\pi_{n,m} = \lambda\pi_{n-1,m} + \mu_S\pi_{n+1,m-1} \quad 2 \le m \le (B-1) \tag{13d}$$
$$(\lambda+\mu_S)\,\pi_{n,B} = \lambda\pi_{n-1,B} + \mu_S\pi_{n+1,B-1} \tag{13e}$$

Y en los estados (N, m),

$$\mu_H\pi_{N,-1} = \lambda\pi_{N-1,-1} \tag{14a}$$
$$\mu_V\pi_{N,0} = \lambda\pi_{N-1,0} + \mu_H\pi_{N,-1} \tag{14b}$$
$$\mu_S\pi_{N,1} = \lambda\pi_{N-1,1} + \mu_V\pi_{N,0} \tag{14c}$$
$$\mu_S\pi_{N,m} = \lambda\pi_{N-1,m} \quad 2 \le m \le (B-1) \tag{14d}$$
$$\mu_S\pi_{N,B} = \lambda\pi_{N-1,B} \tag{14e}$$

Además, se tiene la condición de normalización.

$$\pi_{0,0} + \sum_{n=1}^{N}\sum_{m=-1}^{B}\pi_{n,m} = 1 \tag{15}$$

Para calcular las probabilidades $\pi_{n,m}$, se realiza un desarrollo matricial para el que se definen los siguientes vectores de probabilidades de dimensión $(B+2)\times 1$

$$\vec{\pi}_0 = \begin{pmatrix} \pi_{0,0} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \quad \vec{\pi}_n = \begin{pmatrix} \pi_{n,-1} \\ \pi_{n,0} \\ \pi_{n,1} \\ \vdots \\ \pi_{n,B-1} \\ \pi_{n,B} \end{pmatrix} \quad 1 \le n \le N \tag{16}$$

El desarrollo concluye con unas relaciones matriciales entre los vectores de probabilidades de forma que todos ellos quedan en función de la probabilidad $\pi_{0,0}$.

$$\begin{cases} \vec{\pi}_1 = Z_1\vec{\pi}_0 \\ \vec{\pi}_n = Z_n\vec{\pi}_{n-1} \quad 2 \le n \le N \end{cases} \tag{17}$$

A su vez, las matrices Z_n son resultado de operaciones matriciales donde intervienen las tasas λ, μ_S, μ_V, y μ_H.

$$\begin{cases} Z_N = -\lambda A^{-1} \\ Z_n = -\lambda\left(A - \lambda I + BZ_{n+1}\right)^{-1} \quad 2 \le n \le N-1 \\ Z_1 = -\left(A - \lambda I + BZ_2\right)^{-1} C \end{cases} \tag{18}$$

Las matrices A, B y C, que aparecen en (18), son cuadradas, de dimensiones $(B+2)\times(B+2)$, y contienen los siguientes elementos:

$$A = \begin{pmatrix} -\mu_H & 0 & 0 & 0 & \dots & \dots & 0 \\ \mu_H & -\mu_V & 0 & 0 & \dots & \dots & 0 \\ 0 & \mu_V & -\mu_S & 0 & \dots & \dots & \vdots \\ 0 & 0 & 0 & -\mu_S & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -\mu_S \end{pmatrix}$$

$$B = \begin{pmatrix} 0 & 0 & \dots & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 & \mu_S \\ 0 & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \mu_S & 0 & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \mu_S & 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 0 & \mu_S & 0 \end{pmatrix}$$

$$C = \begin{pmatrix} \lambda & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & 0 & 0 \end{pmatrix}$$

Sustituyendo $\pi_{n,m}$ $(1 \le n \le N, -1 \le m \le B)$ en la condición de normalización (15) por las expresiones extraídas de (16)-(18), se obtiene el valor de $\pi_{0,0}$.

$$\pi_{0,0} = \frac{1}{S}; \quad S = 1 + \left[\sum_{n=1}^{N}\vec{e}_1^{\,T}\left(\prod_{i=0}^{n-1}Z_{n-i}\right)\vec{e}_2\right] \tag{19}$$

$\vec{e}_1$ y $\vec{e}_2$ son vectores de dimensiones $(B+2)\times 1$.

$$\begin{aligned} \vec{e}_1^{\,T} &= (1 \quad 1 \quad \dots\dots \quad 1) \\ \vec{e}_2^{\,T} &= (1 \quad 0 \quad \dots\dots \quad 0) \end{aligned} \tag{20}$$

Una vez que se tiene el valor de $\pi_{0,0}$, ya se pueden determinar todas las probabilidades de estado $\pi_{n,m}$, aplicando (16).

C. Parámetros de rendimiento

Con las probabilidades de estado en régimen estacionario, $\pi_{n,m}$, ya es posible calcular los parámetros de rendimiento de interés.

Probabilidad de bloqueo (P_B). Es igual a la suma de las probabilidades de los estados de bloqueo (N, m).

$$P_B = \sum_{m=-1}^{B}\pi_{N,m} \tag{21}$$

Throughput de captura (X_C). Tiene en cuenta la probabilidad de estar ejecutando una *softirq* y la tasa de servicio μ_S.

$$X_C = \mu_S \sum_{n=1}^{N} \sum_{m=1}^{B} \pi_{n,m} \tag{22}$$

Utilización de procesador en captura (U_C). Tiene en cuenta los estados en los que el procesador se dedica a *hardirqs* y *softirqs*.

$$U_C = \sum_{n=1}^{N} \pi_{n,-1} + \sum_{n=1}^{N} \sum_{m=1}^{B} \pi_{n,m} \tag{23}$$

Directamente relacionada con la utilización, se tiene la disponibilidad del procesador para otras tareas distintas de la captura, D.

$$D = \pi_{0,0} + \sum_{n=1}^{N} \pi_{n,0} = 1 - U_C \tag{24}$$

Frecuencia de *hardirq* (f_{hirq}). Este parámetro mide el número de veces por unidad de tiempo que se atiende a la rutina de atención a la *hardirq*. Para calcular dicha frecuencia, se estima T_{ciclo}, tiempo medio del ciclo compuesto por sistema vacío, *hardirq* y conjunto de *vacations* y *softirqs* hasta vaciar de nuevo el sistema.

$$T_{ciclo} = \frac{1}{\lambda} + \frac{1}{\mu_H} + k_s \left(\frac{1}{\mu_V} + T_{sirq} \right) \tag{25}$$

T_{sirq} es el tiempo medio de *softirq* y k_s es el número medio de *softirqs* dentro del ciclo. Dado que se ha asumido que las llegadas siguen un proceso de Poisson:

$$\begin{aligned} \frac{1}{\lambda} &= T_{ciclo} \cdot \pi_{0,0} \\ f_{hirq} = \frac{1}{T_{ciclo}} &= \lambda \cdot \pi_{0,0} \end{aligned} \tag{26}$$

Frecuencia de *softirq* (f_{sirq}). Es equivalente a la frecuencia de captura (f_C) definida en el modelo M1. En M2, la expresión que toma es la siguiente:

$$f_{sirq} = \mu_V \sum_{n=1}^{N} \pi_{n,0} \tag{27}$$

Tiempo medio de *softirq* (T_{sirq}). Es la duración media de una *softirq*. Coincide con el periodo activo del procesador en captura.

$$T_{sirq} = \frac{\sum_{n=1}^{N} \sum_{m=1}^{B} \pi_{n,m}}{\mu_V \sum_{n=1}^{N} \pi_{n,0}} \tag{28}$$

Una vez calculado el tiempo medio de *softirq*, se puede estimar el número medio de paquetes tratados en una *softirq*, $\overline{m}_{sirq}$.

$$\overline{m}_{sirq} = \frac{T_{sirq}}{\frac{1}{\mu_S}} = \mu_S T_{sirq} = \frac{\mu_S \sum_{n=1}^{N} \sum_{m=1}^{B} \pi_{n,m}}{\mu_V \sum_{n=1}^{N} \pi_{n,0}} \tag{29}$$

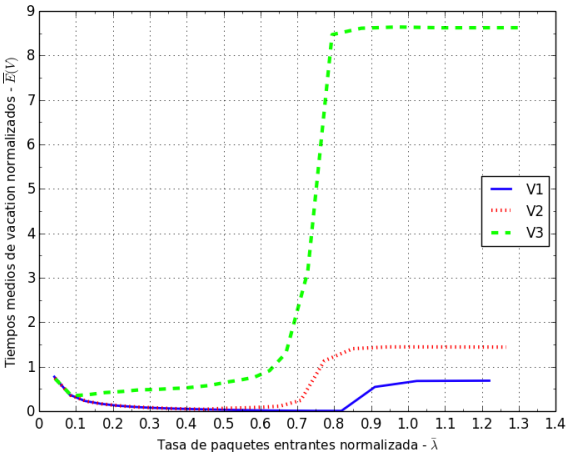


Fig. 4. Escenarios de evaluación V1, V2 y V3

IV. EVALUACIÓN DE MODELOS

En este apartado se muestran resultados de la evaluación de los modelos M1 y M2. Las curvas analíticas se han obtenido a través de la implementación en MATLAB de las ecuaciones derivadas de los modelos. Asimismo, con objeto de realizar comparativas, se presentan valores experimentales obtenidos con una sonda real de captura y análisis de paquetes basada en Linux [16] que opera dentro de una plataforma de medidas de laboratorio [17]. Para que sean comparables ambos tipos de resultados, algunos parámetros de entrada de los modelos (λ, $1/\mu_S$, $1/\mu_V$, $1/\mu_H$) toman sus valores a partir de mediciones realizadas sobre la plataforma experimental.

Se definen tres escenarios de evaluación (V1, V2 y V3) que se caracterizan por tener distinto comportamiento de *vacation* tal y como se muestra en la Fig. 4. Concretamente, se representa, para diferentes tasas normalizadas de entrada de paquetes, $\overline{\lambda} = \lambda/\mu_S$, el tiempo medio de *vacation* normalizado con respecto a la duración máxima de una *softirq* de Linux, $\overline{E}(V) = (1/\mu_V)/(B/\mu_S)$, con $B = 300$ ya que éste es el valor típico de *budget*. En la Fig. 4 se aprecia que los tiempos $\overline{E}(V)$ no permanecen constantes. Esto se debe a que son tiempos extraídos de la sonda real y, en ésta, las operaciones de análisis posteriores a la captura requieren mayores consumos computacionales a medida que aumenta la tasa de entrada de paquetes. Únicamente se ha reproducido el escenario V1 en laboratorio, por lo que, solamente para este caso se realizan comparativas entre los modelos (M1 y M2) y los valores de la sonda.

La Fig. 5 muestra resultados de medidas de rendimiento, tanto de la evaluación de los modelos M1 y M2 como de las medidas de la sonda real de laboratorio (referidas como "Lab"). En primer lugar, la Fig. 5a expone cómo varía el throughput de captura normalizado, $\overline{X}_C = X_C/\mu_S$, con los diferentes escenarios. Se evalúa con $N = 200$ y $B = 300$ (valores coincidentes con la configuración del driver de la tarjeta de red de la sonda). Se observa que el modelo M1 tiene un comportamiento similar para los

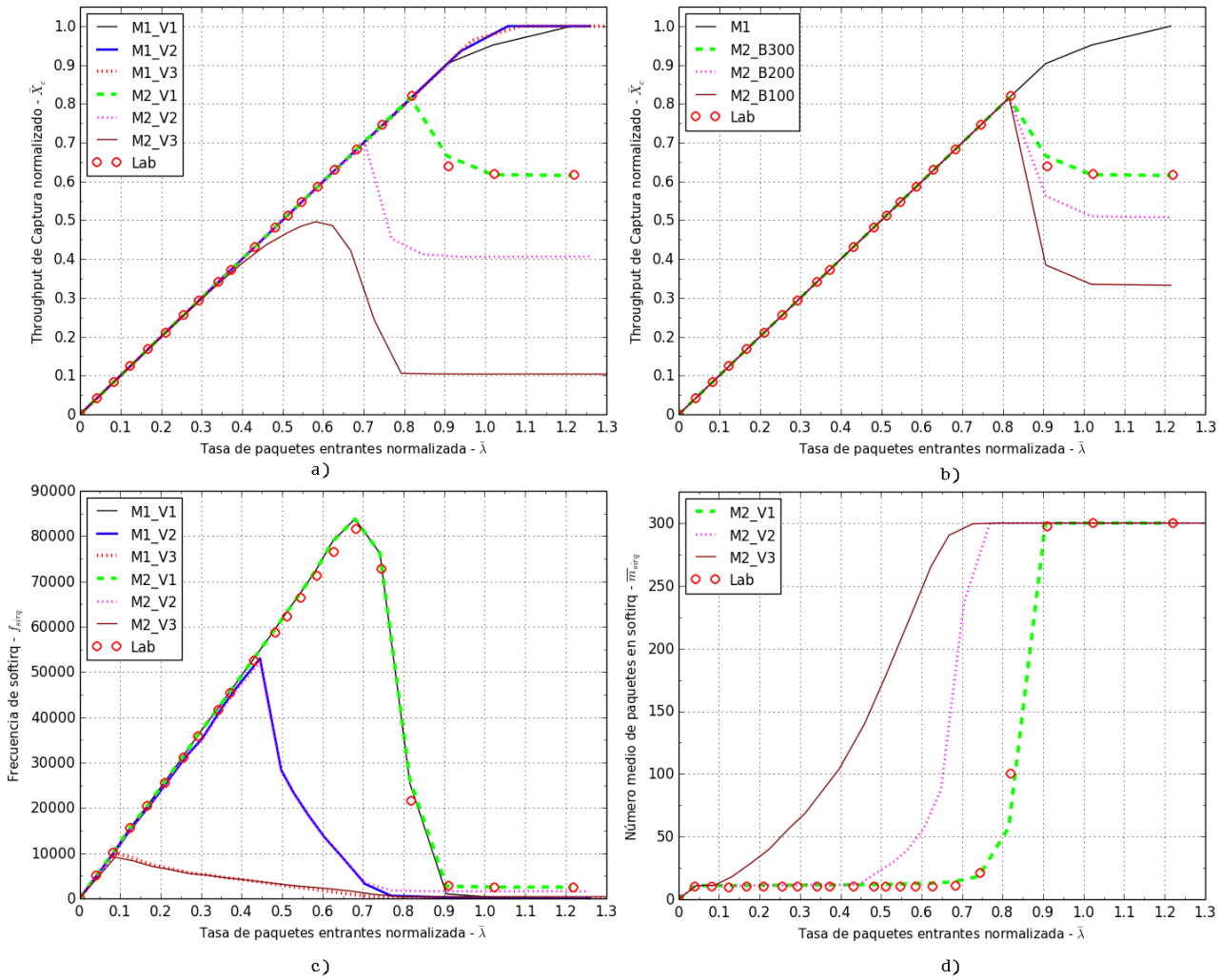


Fig. 5. Medidas de rendimiento de los modelos M1 y M2

escenarios V1, V2, V3; es decir, prácticamente no se ve afectado por los tiempos de *vacation*. Esto se debe a la disciplina exhaustiva que se ha supuesto para el proceso de captura, que permite alargar indefinidamente el tiempo dedicado a la captura y obtener el máximo de throughput en saturación, $\overline{X}_C \approx 1$, a costa de eliminar, en la práctica, los periodos de *vacation*. El modelo M1 empieza a saturarse a partir del valor $\overline{\lambda} \approx 0.9$. Si se compara con los datos de laboratorio, sólo se ajusta para valores por debajo de $\overline{\lambda} \approx 0.8$. Por su parte, el modelo M2, a diferencia de M1, sí se ve afectado por los periodos de inactividad y el throughput de captura disminuye. El caso M2_V1 se ajusta a los valores de laboratorio en su totalidad. Alcanza el máximo para $\overline{X}_C \approx 0.8$. El decrecimiento posterior se debe al aumento de los tiempos de *vacation* y a la finalización de la *softirq* por *budget*. Para las tasas más altas, $\overline{\lambda} > 1$, el throughput se vuelve aproximadamente constante debido a que los tiempos medios de *vacation* se estabilizan y la duración del proceso de captura viene fijada por el límite B. Los casos M2_V2 y M2_V3 sirven para predecir el comportamiento del sistema en otros escenarios. Se ve que la forma es similar a la de M2_V1, pero se alcanzan valores de throughput menores.

A continuación, la Fig. 5b muestra la variación del throughput con respecto al *budget*, B, sobre el escenario V1. Se mantiene $N = 200$. El modelo M1 no tiene definido propiamente el parámetro *budget*, pero puede considerarse equivalente al caso extremo $B \to \infty$. El caso M2_B300 se ajusta a los valores de laboratorio. Con valores de *budget* menores, casos M2_B200 y M2_B100, se observa que el throughput en la zona de saturación ($\overline{\lambda} > 0.8$) disminuye, tanto más cuanto menor sea el valor de B. Esto se debe a que, cuando se agota el *budget*, el tiempo dedicado a la captura es menor cuanto menor es B. Si se evalúa el modelo M2 para valores de $B > 300$, se obtienen throughputs mayores al caso M2_B300, pero siempre por debajo del caso extremo M1.

Por su parte, la Fig. 5c muestra la variación de la frecuencia de *softirq*. Este valor pueda dar idea del número de cambios de contexto que se dan entre los periodos de captura y *vacation*. En los tres escenarios (V1, V2 y V3) y para ambos modelos (M1 y M2), se pueden distinguir tres zonas: una para tasas bajas, en la que a medida que aumenta la tasa de entrada, y con ella la

actividad de captura, el número de *softirq* por segundo crece hasta que llega a un punto máximo. A partir de ahí, con tasas de entrada intermedias, debido al aumento de los tiempos de *vacation*, la frecuencia de *softirq* disminuye. Finalmente, hay una tercera zona para tasas de entrada más altas donde, en el caso del modelo M1, la frecuencia de *softirq* disminuye porque la duración de la *softirq* se alarga indefinidamente, no existen prácticamente *vacations* y $f_{sirq} \to 0$; por contra, en la tercera zona del modelo M2 la frecuencia de *softirq* se mantiene constante, con un valor bajo, ya que la duración media de la *softirq* toma el valor B/μ_S. En la Fig. 5c también se da que el caso M2_V1 se ajusta a la medida de laboratorio.

También se ha analizado los resultados de frecuencia de *hardirq*, f_{hirq} en el modelo M2. Se ha comprobado que, para tasas de entrada en las que no se alcanza el *budget*, $f_{hirq} \approx f_{sirq}$; y que, para tasas en la zona de saturación, se llega a un valor extremo tal que $f_{hirq} \to 0$.

Por último, la Fig. 5d presenta el número medio de paquetes por *softirq*, definido como $\overline{m} = \mu_S T_{sirq}$. Permite visualizar, a partir de qué tasa de entrada, la ejecución de la *softirq* alcanza el valor del *budget* ($B = 300$ en la gráfica). Obviamente, el escenario V3 es el que agota el *budget* con menor tasa de entrada, y el escenario V1 el que lo alcanza con mayor tasa. Una vez más, el caso M2_V1 es el que se ajusta a los valores medidos en la sonda de laboratorio.

Cabe mencionar que también se han evaluado los modelos con tamaños de búfer mayores que 200 (en concreto, $N = 512$), pero los resultados obtenidos han sido similares. También se ha probado qué ocurre si se desprecian las *hardirq* en el modelo M2 (suponiendo $1/\mu_H \to 0$) y la conclusión ha sido que es factible esa suposición, ya que los resultados obtenidos son prácticamente iguales.

V. CONCLUSIONES

Este trabajo plantea la evaluación del rendimiento de un sistema de captura de paquetes a partir de un modelado analítico consistente en un sistema de cola con *vacations*. Este concepto permite modelar el comportamiento del procesador responsable de capturar paquetes y, a su vez, de realizar tareas adicionales en los denominados tiempos de *vacation*.

M1, el primer modelo propuesto, puede considerarse un modelo simple que caracteriza a un sistema de captura genérico que, dada la disciplina de servicio exhaustiva, prioriza el proceso de captura. En condiciones en las que el sistema no está saturado, garantiza la ejecución de tareas adicionales en los denominados tiempos de *vacation*; por el contrario, en condiciones de saturación, el proceso de captura acapara el tiempo del procesador en detrimento de realizar otras tareas adicionales, pero se consiguen throughput de captura aceptables .

M2, el segundo modelo, es más complejo y su disciplina de servicio limitada recoge las particularidades de sistemas de captura como Linux que establecen un límite para la ejecución del proceso de captura. Esto garantiza la ejecución de otras tareas para todos los casos, pero conlleva una pérdida de throughput de captura. En estos casos será interesante valorar el beneficio que supone disponer de tiempos de *vacation* para dichas tareas adicionales, frente a la pérdida potencial de throughput de captura.

Las conclusiones de este trabajo son satisfactorias en lo que respecta al comportamiento de los modelos. No obstante, este trabajo, que combina estudio analítico con medidas experimentales, trae varios aspectos a considerar en un futuro próximo. En primer lugar, se considera interesante estudiar modelos con *vacations* con procesos que no sean de tipo Poisson. En estos casos, si la resolución analítica se vuelve intratable, no se descarta la opción de evaluar el rendimiento mediante técnicas de simulación. Por otro lado, dado que, en un sistema de monitorización de tráfico de red, la etapa de captura puede estar relacionada con tareas posteriores de análisis o de otra índole, puede ser susceptible de estudio el modelado y la evaluación del rendimiento del sistema completo de monitorización de red, teniendo presente la influencia del planificador de tareas.

REFERENCIAS

[1] N. Tian, Z.G. Zhang, "Vacation Queuing Models. Theory and Applications", Springer Science+Business Media, 2006.

[2] H. Takagi, "Queueing Analysis. A Foundation of Performance Evaluation Volume 1: Vacation and Priority Systems (Part 1)", North-Holland, 1991.

[3] S. Lee, K. Levanti, H.S. Kim, "Network monitoring: Present and future", Computer Networks, vol. 65, n. 2, pp. 84-98, 2014.

[4] B. Li, J. Springer, G. Bebis, M. H. Gunes, "A survey of network flow applications", Journal of Network and Computer Applications, vol. 36, n. 2, pp. 567-581, 2013.

[5] S. Gallenmuller, P. Emmerich, F. Wohlfart, D. Raumer, G. Carle, "Comparison of frameworks for high-performance packet IO", Proceedings 2015 ACM/IEEE Symposium on Architectures for Networking and Communications Systems, pp. 29-38, 2015.

[6] ntop project, http://www.ntop.org

[7] S. Han, K. Jang, K. Park, S. Moon "PacketShader: a GPU-accelerated software router", ACM SIGCOMM Computer Communication Review, vol. 41, n. 4, pp. 195-206, 2011.

[8] L. Rizzo, "netmap: A Novel Framework for Fast Packet I/O", USENIX Annual Technical Conference, pp. 101-112, 2012.

[9] N. Bonelli, A. Di Pietro, S. Giordano, G. Procissi, "On multi-gigabit packet capturing with multi-core commodity hardware", Proceedings Passive and Active Measurement, pp. 64-73, 2012.

[10] OpenOnload, http://www.openonload.org/

[11] V. Moreno, P.M. Santiago del Río, J. Ramos, J.L. García-Dorado, I. González, F.J. Gómez, J. Aracil, "Packet Storage at Multi-gigabit Rates Using Off-the-Shelf Systems", 16th IEEE International Conference on High Performance and Communications, 2014.

[12] W. Wu, M. Crawford, M. Bowde, "The performance analysis of Linux networking - packet receiving", Computer Communications, vol. 30, n. 5, pp. 1044-1057, 2007.

[13] K. Salah, K. El-Badawi, R. Boutaba, "Performance modeling and analysis of network firewalls", IEEE Transactions on Network and Service Management, vol. 9, n. 1, pp. 12-21, 2012.

[14] G. Bolch, S. Greiner, H. Meer, K.S. Trivedi, "Queuing Networks and Markov Chains", John Wiley&Sons, 1998.

[15] I. Adan, J. Resing, "Queueing Theory", Eindhoven University of Technology, 2002.

[16] A. Muñoz, A. Ferro, F. Liberal, J. Lopez, "A Kernel-Level Monitor over Multiprocessor Architectures for High-Performance Network Analysis with Commodity Hardware" Proceedings SensorComm 2007 Valencia, 2007.

[17] A. Pineda, L. Zabala, A. Ferro, "Network Architecture to Automatically Test Traffic Monitoring Systems", Mosharaka International Conference on Communications and Signal Processing, 2012.

[18] D.P. Bovet, M. Cesati, "Understanding the Linux Kernel, Third Edition". O'Reilly Media, 2005.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6582

Throughput Analysis and Optimization of Multi-layer FFR-aided OFDMA Networks

Jan García-Morales, Guillem Femenias, Felip Riera-Palou, and John S. Thompson.
Mobile Communications Group – University of the Balearic Islands.
Institute for Digital Communications – The University of Edinburgh.
{jan.garcia,guillem.femenias,felip.riera}@uib.es, john.thompson@ed.ac.uk

***Abstract*—In OFDMA networks, the use of universal frequency reuse plans improves cell capacity but causes very high levels of inter-cell interference (ICI), particularly affecting users located in the cell-edge regions. In order to mitigate ICI while achieving high spectral efficiencies, fractional frequency reuse (FFR) shows a good tradeoff between cell-edge throughput and overall cell spectral efficiency. Recently, multi-layer FFR-aided OFDMA-based designs, splitting the cell into inner, middle and outer layers have been proposed and studied with the aim of increasing the spectrum utilization and improving the user fairness throughout the cell. This paper presents an analytical framework allowing the performance evaluation and optimization of multi-layer FFR designs in OFDMA-based networks. Tractable mathematical expressions of the average cell throughput as well as the layer spectral efficiency have been derived for both proportional fair (PF) and round robin (RR) scheduling policies.**

***Keywords*—OFDMA cellular networks, multi-layer FFR, spectral efficiency, throughput, optimization.**

I. INTRODUCTION

Orthogonal frequency division multiple access (OFDMA) is one of the most prominent air-interfaces in modern cellular standards [1]. Owing to the orthogonality among subcarriers, OFDMA makes the intra-cell interference negligible. However, inter-cell interference (ICI) remains an issue due to the use of *aggressive* high spectral efficiency universal frequency reuse plans where all cells use the same set of frequency subbands (reuse-1). In this setup, ICI critically affects the user mobile stations (MSs) located in the edge of the cells because the serving base station (BS) and the interfering ones are at similar distances. In contrast, the well-known reuse-3 scheme decreases ICI but sacrificing spectral efficiency. With the aim of mitigating ICI experienced by the cell-edge users while still achieving high spectral efficiencies, multiple ICI control (ICIC) strategies have been proposed [2], among which, *static* fractional frequency reuse (FFR) and all its variants show a good tradeoff among cell-edge throughput enhancement, provision of high spectral efficiency and implementation complexity [3].

The FFR scheme divides the cell into two layers, the inner and the outer one (also known as cell-center and cell-edge regions). In FFR-based cellular systems, a low frequency reuse factor is used for the cell-inner MSs (typically reuse-1), less affected by co-channel interference, and a larger frequency reuse factor is selected for the cell-outer MSs (e.g., reuse-3). However, traditional two-layer FFR has some drawbacks: (i) when the inner layer is large, the MSs located in the edge of the inner layer suffer from high levels of ICI, (ii) when the outer layer becomes large, the spectrum utilization becomes low and the spectral efficiency drops. In an attempt to reconcile these two conflicting situations, FFR schemes with more than 2 layers have been recently proposed [4]. The main idea of the multi-layer FFR scheme is to increase the spectrum utilization, enhance the average cell throughput and improve the MS fairness throughout the cell by incorporating middle layers in between the inner and outer ones.

Regardless of the particular ICIC technique in use, spectral efficiency can be significantly enhanced by using channel-aware schedulers that allocate, on a slot-by-slot basis, each subcarrier to a user with favourable channel conditions (i.e., a user experiencing a high signal-to-interference-plus-noise ratio (SINR)), thus exploiting multiuser diversity. Remarkably, the proportional fair (PF) scheduler has been shown to provide a good tradeoff between spectral efficiency and fairness [5]. Then, ICI can be decreased using a PF scheduler in combination with FFR schemes, while at the same time the possibility of a MS with a very bad link suffering from long periods of starvation can be drastically reduced.

The analytical performance evaluation of FFR-aided OFDMA-based cellular networks has been tackled using Poisson Point Processes (PPPs) for modeling the location of the BSs [6], [7]. This approach allows the characterization of the system performance by spatially averaging over all possible network realizations, but precludes from accurately analyzing the performance of a given cell,

a metric of particular importance to network designers that, provided a planned set of BS locations along with traffic load conditions, may be interested in calculating the performance obtained within a specific region in the coverage area of the network. Fan Jin *et al.* [8] studied an FFR-aided twin-tier OFDMA network where stochastic geometry was used to characterize the random distribution of femtocells, and the macrocells were overlaid on top of the femtocells following a regular tessellation. However, the analytical framework was limited to resource allocation schemes based on the round robin scheduling policy. Similar approaches, lacking the consideration of scheduling policies and small scale fading, were also proposed by Assaad in [9] and Najjar *et al.* in [10] to optimize FFR-based parameters in a single-tier network. These limitations were overcome in part by Xu *et al.* in [11] and Garcia *et al.* in [12] (see also [13]), but only taking into account the use of opportunistic maximum SINR (MSINR) schedulers.

In contrast to the above background work, following studies have considered the use of the multi-layer FFR scheme to control the ICI. Xie and Walke [14] proposed a three-layer FFR scheme using reuse-1 and low power for inner layer, reuse-3 and moderate power for middle layer, and reuse-9 and high power for outer layer. A theoretical analysis of a series of reuse partitioning approaches was carried out in this paper using mathematically tractable expressions, but with the only consideration of the pathloss effect and thus precluding any attempt to analyze the system performance under the use of channel-aware schedulers. Ghaffar and Knopp [15] proposed a three-layer scheme that divided the whole spectrum into four subbands. They used reuse-1 for the inner layer and reuse-3/2 for the middle and outer layers. The use of this approach provided a reduction of power consumption at the BSs leading to an improvement of the average spectral efficiency but at the cost of increasing the ICI. A multi-layer soft frequency reuse (SFR) scheme was proposed by Yang in [16] with different power levels for each layer. Using this approach allowed the achievement of a better interference pattern than that obtained using a two-layer SFR, thus improving the overall spectral efficiency. In [15] and [16], the average cell and layer spectral efficiencies were formulated, but the authors did not provide neither closed-form solutions nor mathematically tractable expressions and consequently, only results obtained through Monte-Carlo simulations were presented. Particularly interesting is the work of Wang *et al.* in [4], where a tractable multi-layer FFR model was proposed. Moreover, optimal designs and closed-form expressions of the average spatial capacities of certain typical regions of a cell were derived. One of the main conclusions of this work was that multi-layer schemes can provide better average spatial capacity and fairness than the traditional two-layer scheme. The main limitation of this work, however, was the use of rather unrealistic assumptions such as neglecting the small scale fading effects and, consequently, limiting the proposed analytical framework to resource allocation schemes based on the round robin scheduling policy.

In this paper, a novel approach for a multi-layer FFR-aided downlink OFDMA-based multi-cellular network is introduced, studied and compared. To this end, an analytical framework is presented allowing the performance evaluation of both two-layer FFR and multi-layer FFR using a PF or a RR scheduling policy. The main contributions of this paper can be summarized as follows:

- Based on the statistical channel characterization and a unified cell throughput approach, an analytical framework allowing the evaluation of the impact that any of the FFR layers may produce on the cell throughput is provided.
- Tractable mathematical expressions of the average cell throughput as well as the layer capacity have been derived for both proportional fair (PF) and round robin (RR) scheduling policies.
- The worst MSs, typically located at the edge of each layer, are considered in the analysis aiming at determining the size of the fractional frequency scheme-related spatial and frequency partitions guaranteeing proper QoS and fairness levels throughout the cell coverage.

It is worth stressing at this point that the proposed analytical framework also opens the door to the theoretical spectral efficiency evaluation and optimization of OFDMA-based cellular networks using more sophisticated ICIC techniques such as adaptive frequency reuse or network MIMO, as well as to the assessment of cellular heterogeneous networks.

II. CELLULAR NETWORK MODEL

Let us consider the downlink of an OFDMA-based cellular system where a set of BSs are assumed to be deployed following a conscious planning and thus, are regularly arranged over the whole coverage area. This cellular environment can be safely modeled as a regular tessellation of hexagonally-shaped coverage areas, as shown in Figs. 1 and 2 , with the BSs located at the centre of the hexagons[1]. For the sake of analytical tractability, the central cell, covered by BS 0, which will be referred to as the tagged BS, will be approximated by a circle whose area is the same as the hexagonal one. That is, assuming that the side of the regular hexagon is R_h, the radius of the circular cell is $R = R_h\sqrt{3\sqrt{3}/(2\pi)}$, and the total cell coverage area is $A_r^T = \pi(R^2 - R_0^2)$, where R_0 is the minimum distance of a MS from its serving BS.

The locations of the MSs at a given time instant are assumed to form a stationary PPP of normalized intensity λ (measured in MSs per area unit). A consequence of this assumption is that the probability distribution of the number $M_{\mathcal{S}}$ of MSs falling within any spatial region $\mathcal{S}$ of area $A_r^{\mathcal{S}}$ follows a Poisson distribution, thus implying

$$\mathbb{P}\{M_{\mathcal{S}} = k\} = \frac{(\lambda A_r^{\mathcal{S}})^k e^{-\lambda A_r^{\mathcal{S}}}}{k!}. \quad (1)$$

[1]Omnidirectional antenna BSs are assumed in this paper. In future work this will be extended to consider the use of sectorization.

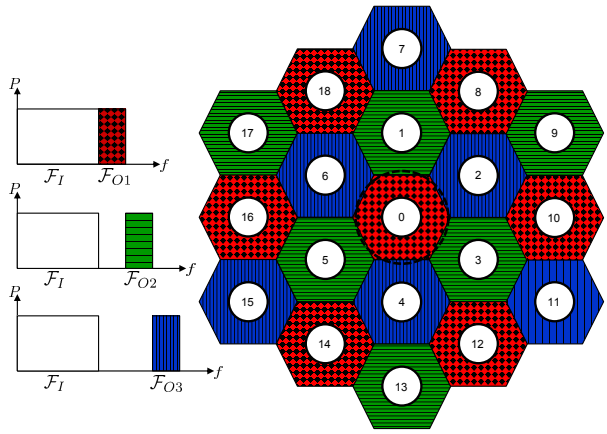


Fig. 1: Schematic representation of the two-layer FFR-aided OFDMA-based cellular network.

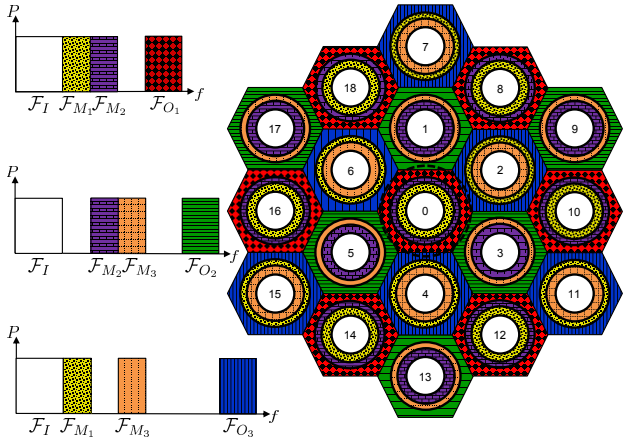


Fig. 2: Schematic representation of a multi-layer FFR-aided OFDMA-based cellular network.

A. Two-layer FFR network layout

In order to control the ICI, MSs are classified according to the received average SINR as either cell-inner MSs, when the received average SINR is above a given threshold, or cell-outer MSs, when it is below the threshold. A two-layer FFR scheme is then applied by allocating non-overlapping resources (subcarriers) to cell-inner and cell-outer MSs, while employing a frequency reuse factor equal to one (reuse-1) for the cell-inner MSs and a higher frequency reuse factor for the cell-outer MSs that, is assumed to be 3 in this paper (reuse-3). For analytical tractability, inner and outer regions (or layers) will be separated by a circumference of radius R_{th} (threshold distance).

The total system bandwidth is exploited by means of a set $\mathcal{F}_T$ of N_T orthogonal subcarriers with a bandwidth Δf small enough to assume that all subcarriers experience frequency flat fading. The set $\mathcal{F}_T$ is split into a set $\mathcal{F}_I$ of subcarriers allocated to the inner layer and a set $\mathcal{F}_O = \mathcal{F}_T \backslash \mathcal{F}_I$ of subcarriers allocated to outer layers. The set $\mathcal{F}_O$ is further split into three equal parts, namely $\mathcal{F}_{O1}$, $\mathcal{F}_{O2}$ and $\mathcal{F}_{O3}$, which are allocated to outer-cell MSs in such a way that adjacent cells will operate on different sets of subcarriers, as shown in Fig. 1. Note that, denoting by N_I and N_O the number of subcarriers allocated to the inner layer and each of the outer layers, respectively, it holds that $N_T = N_I + 3N_O$.

B. Multi-layer FFR network layout

When multi-layer FFR scheme is applied, a middle layer is inserted between the inner and the outer ones. This is necessary because neither the inner layers nor the outer layers should be large as explained in Section I. The reuse factor of inner layers should be small (e.g., reuse-1) to keep a relatively high spectrum utilization. Meanwhile, the reuse factor of outer layers should be large (e.g., reuse-3) in order to avoid high levels of ICI affecting the MSs located far from the BS. Wang *et al.* in [4] divide the middle layer into two sublayers, i.e., middle1 and middle2, with reuse factor 3/2, as shown in Fig. 2, that is a feasible and practical choice for the design of a multi-layer FFR scheme, and also preferred from a performance point of view. The inner and outer layers are designed in the same way as the traditional FFR scheme. Again, for analytical tractability, the inner, middle1, middle2 and outer layers will be separated by circumferences of radii R_{th}, R_{M_1} and R_{M_2}.

In the context of this paper, the ratio between the middle and outer areas are set to 0.2, as proposed by Wang *et al.* in [4], which is shown to be a good choice in order to increase fairness among MSs [17]. Accordingly, when using the multi-layer FFR scheme, we have

$$\frac{R_{M_1}^2 - R_{th}^2}{R^2 - R_{M_2}^2} = \frac{R_{M_2}^2 - R_{M_1}^2}{R^2 - R_{M_2}^2} = \frac{1}{5}. \tag{2}$$

Note that, from (2), R_{M_1} and R_{M_2} can be written in terms of the distance threshold R_{th}.

Furthermore, the set $\mathcal{F}_T$ is split into sets $\mathcal{F}_I$, $\mathcal{F}_M$ and $\mathcal{F}_O$ of subcarriers allocated to the centre, middle and outer layers, respectively. Sets $\mathcal{F}_M$ and $\mathcal{F}_O$ are further split into three equal parts, namely $\mathcal{F}_{M1}$, $\mathcal{F}_{M2}$ and $\mathcal{F}_{M3}$, which are allocated to middle-cell MSs and $\mathcal{F}_{O1}$, $\mathcal{F}_{O2}$ and $\mathcal{F}_{O3}$ which are allocated to outer-cell MSs , respectively (see Fig. 2). We have that $N_T = N_I + 3N_M + 3N_O$, where N_M is the number of subcarriers allocated to each of the middle layers.

III. STATISTICAL CHANNEL CHARACTERIZATION

The downlink channel is subject to path loss and small-scale fading[2]. The path loss characterising the link between the bth BS and the uth MS can be modeled as

$$L_{\text{dB}}\left(d_{b,u}\right) = K + 10\alpha \log_{10}\left(d_{b,u}\right), \tag{3}$$

where K and α are, respectively, a constant and the path loss exponent, and $d_{b,u}$ is the distance (in metres) between the BS b and the MS u.

[2]In line with the studies in [8], [11], for analytical simplicity, only pathloss and small scale fading are considered in this paper. In future work this will be extended to consider also shadowing as well.

The instantaneous SINR experienced by MS u in the cell of interest on the nth subcarrier during the scheduling period t can then be expressed as

$$\gamma_{u,n}(t) = \frac{P_s L(d_{0,u}) |H_{0,u,n}(t)|^2}{N_0 \Delta f + I_{u,n}(t)}, \tag{4}$$

where P_s is the power allocated per subcarrier, $H_{b,u,n}(t) \sim \mathcal{CN}(0,1)$ is the frequency response resulting from the small-scale fading channel linking the bth BS to MS u on the nth subcarrier during scheduling period t, N_0 is the noise power spectral density, and $I_{u,n}(t)$ denotes the interference term that is given by

$$I_{u,n}(t) = \sum_{b \in \Phi_n} P_s L\left(d_{b,u}\right) |H_{b,u,n}(t)|^2, \tag{5}$$

with Φ_n representing the set of interfering BSs, which is subcarrier-dependent as the set of interfering BSs depends on which layer subcarrier n belongs to. In fact, for the two-layer FFR scheme we have

$$\Phi_n = \begin{cases} \{1,2,...,18\}, & n \in \mathcal{F}_I \\ \{8,10,12,14,16,18\}, & n \in \mathcal{F}_{O_1} \end{cases}, \tag{6}$$

and when the multi-layer FFR scheme is used, we have

$$\Phi_n = \begin{cases} \{1,2,...,18\}, & n \in \mathcal{F}_I \\ \{2,4,6,7,8,10,11,12,14,15,16,18\}, & n \in \mathcal{F}_{M_1} \\ \{1,3,5,8,9,10,12,13,14,16,17,18\}, & n \in \mathcal{F}_{M_2} \\ \{8,10,12,14,16,18\}, & n \in \mathcal{F}_{O_1}. \end{cases} \tag{7}$$

Assuming the use of uniform power allocation, the P_s can be obtained as

$$\begin{aligned} \text{Two-layer FFR: } P_s &= \frac{P_T}{(N_I + N_O)}, \\ \text{Multi-layer FFR: } P_s &= \frac{P_T}{(N_I + 2N_M + N_O)}, \end{aligned} \tag{8}$$

where P_T represents the available transmit power at the BS.

As an important notational remark, note that $L\left(d_{b,u}\right)$ can be expressed in terms of the polar coordinates of MS u with respect to BS 0 as $L(d_{0,u}, \theta_{0,u})$ and thus, strictly speaking, $\gamma_{u,n}(t)$ is a function of $d_{0,u}$ and $\theta_{0,u}$. Furthermore, it is shown in [18] that the instantaneous SINR in multicell networks barely depends on the polar angle and thus, from this point onwards, the dependence of $\gamma_{u,n}(t)$ on $\theta_{0,u}$ will be omitted.

Since[3] $h_b \triangleq |H_{b,u,n}|^2$ conforms to an exponential distribution with probability density function (PDF) $f_{h_b}(x) = e^{-x}u(x)$, where $u(x)$ represents the unit step function, its corresponding cumulative distribution function (CDF) can be obtained as $\mathbb{P}\{h_b \leq x\} = (1 - e^{-x})u(x)$. Hence, the CDF of the instantaneous SINR $\gamma_{u,n}$ conditioned on the set of small-scale fading gains $\boldsymbol{h} \triangleq \{h_b\}_{\forall b \neq 0}$ for a given location of MS u, can be derived from (4) as

$$\begin{aligned} F_{\gamma_{u,n}|\boldsymbol{d},\boldsymbol{h}}(x|d,\boldsymbol{h}) &\triangleq \mathbb{P}\{\gamma_{u,n} \leq x | d_{0,u}, \boldsymbol{h}\} \\ &= \mathbb{P}\left\{ h_0 \leq \frac{(N_0 \Delta f + I_{u,n})}{\bar{\gamma}_0} x | d_{0,u}, \boldsymbol{h} \right\} \\ &= 1 - e^{-\frac{x(N_0 \Delta f + I_{u,n})}{\bar{\gamma}_0}}, \qquad x \geq 0, \end{aligned} \tag{9}$$

where $\bar{\gamma}_0 = P_n L(d)$ represents the average received signal. Note that distances in the set $\boldsymbol{d}$ can be written in terms of the distance $d_{0,u} = d$ from the serving BS to MS u.

Now, using (9) and averaging over the PDFs of the i.i.d. random variables $\boldsymbol{h}$, the conditional CDF of the instantaneous SINR $\gamma^A_{u,n}$ experienced by MS u located at distance $d_{0,u} = d$ from the serving BS and in the region A, can be obtained as

$$\begin{aligned} &F_{\gamma^A_{u,n}|d_{0,u}}(x|d) \triangleq \mathbb{P}\{\gamma^A_{u,n} \leq x | d_{0,u}\} \\ &= \int_0^\infty ... \int_0^\infty \left(1 - e^{-\frac{x(N_0 \Delta f + I_{u,n})}{\bar{\gamma}_0}}\right) \prod_{i \in \Phi_n} f_{h_i}(h_i) \mathrm{d}h_i \\ &= 1 - e^{-\frac{x N_0 \Delta f}{\bar{\gamma}_0}} \int_0^\infty ... \int_0^\infty e^{-\frac{x\left(\sum_{i \in \Phi_n} h_i \bar{\gamma}_i\right)}{\bar{\gamma}_0}} \prod_{i \in \Phi_n} e^{-h_i} \mathrm{d}h_i \\ &= 1 - e^{-\frac{x N_0 \Delta f}{\bar{\gamma}_0}} \prod_{i \in \Phi_n} \frac{1}{1 + \frac{x \bar{\gamma}_i}{\bar{\gamma}_0}}, \ x \geq 0, \end{aligned} \tag{10}$$

where A is a token used to represent the cell layers (or regions) I, M_1, M_2, or O_1, $f_{h_i}(h_i)$ is the PDF of the variable $h_i = |H_{i,u,n}|^2$, and $\bar{\gamma}_i = P_n L\left(d_{i,u}\right)$ is the average interfering signal from each interfering BS.

IV. THROUGHPUT ANALYSIS

The average cell throughput (measured in bps) for the downlink of the fractional frequency reuse schemes-aided OFDMA-based cellular network can be expressed as

$$\overline{\eta}^T = \sum_{\forall A} \overline{\eta}^A, \tag{11}$$

where $\overline{\eta}^A$ is the average throughput in cell layer A.

Let us define M_0 as a positive integer random variable representing the number of MSs in the region served by the tagged BS. As MSs are assumed to be uniformly distributed in entire cell region, the probability that an MS is located in cell layer A is

$$P^A_r = \frac{\left(R^A_U\right)^2 - \left(R^A_L\right)^2}{R^2 - R^2_0}, \tag{12}$$

where R^A_L and R^A_U denote the lower and upper radii of the circumferences defining layer A. Using these definitions, the average throughput in cell layer A can be expressed as shown in (13) and (14) on top of the next page, for both the two-layer and the multi-layer FFR schemes, respectively, where $\overline{\eta}^A_n(k_A)$ is the average throughput on the nth subcarrier when there are k_A MSs in cell layer A.

Now, defining M_A as a non-negative integer random variable representing the number of MSs in cell region A,

[3]Note that since the channel is assumed to be stationary, from this point onwards the time dependence (i.e., (t)) of all the variables will be dropped unless otherwise stated.

$$\text{Two-layer FFR scheme: } \overline{\eta}^A = \sum_{k=1}^{\infty} \mathbb{P}\{M_0 = k\} \sum_{k_I=0}^{k} \binom{k}{k_I, k-k_I} \left(P_r^I\right)^{k_I} \left(P_r^{O_1}\right)^{k-k_I} \left[N_A \overline{\eta}_n^A(k_A)\right]. \tag{13}$$

$$\begin{aligned}\text{Multi-layer FFR scheme: } \overline{\eta}^A = \sum_{k=1}^{\infty} \mathbb{P}\{M_0 = k\} \sum_{k_I=0}^{k} \sum_{k_{M_1}=0}^{k-k_I} \sum_{k_{M_2}=0}^{k-k_I-k_{M_1}} \binom{k}{k_I, k_{M_1}, k_{M_2}, k-k_I-k_{M_1}-k_{M_2}} \\ \times \left(P_r^I\right)^{k_I} \left(P_r^{M_1}\right)^{k_{M_1}} \left(P_r^{M_2}\right)^{k_{M_2}} \left(P_r^{O_1}\right)^{k-k_I-k_{M_1}-k_{M_2}} \left[N_A \overline{\eta}_n^A(k_A)\right].\end{aligned} \tag{14}$$

the average throughput on the nth subcarrier allocated to cell region A when $M_A = k$, can be obtained as

$$\begin{aligned}\overline{\eta}_n^A(k) &= \Delta f \; \mathbb{E}_{\gamma_n^A|M_A} \left\{ \log_2\left(1+\gamma_n^A\right) | M_A = k \right\} \\ &= \Delta f \log_2 e \int_0^{\infty} \frac{1 - F_{\gamma_n^A|M_A}(x|k)}{1+x} \, \mathrm{d}x.\end{aligned} \tag{15}$$

In order to obtain tractable mathematical expressions for the CDF $F_{\gamma_n^A|M_A}(x|k)$ has to be calculated and this depends on the specific scheduling policy applied by the resource allocation algorithm. In the following subsections, this CDF will be obtained for the PF and RR scheduling rules.

A. PF scheduling

A PF scheduler, exploiting the knowledge of the instantaneous SINRs experienced by all MSs $q \in \mathcal{M}_A$, allocates the subcarrier $n \in \mathcal{F}_A$ to MS $u \in \mathcal{M}_A$ satisfying

$$u = \arg \max_{q \in \mathcal{M}_A} \{w_q(t) \gamma_{q,n}(t)\}, \tag{16}$$

where $\mathcal{M}_A$ is the set indexing all MSs in cell region A, and $w_q(t) = 1/\mu_q(t)$ is the weighting (prioritisation) coefficient for MS q that, in this case, depends on the short-term average evolution of channel-state information. This can be obtained using a moving average over a window of W scheduling periods as

$$\mu_q(t) = \left(1 - \frac{1}{W}\right) \mu_q(t-1) + \sum_{n \in \mathcal{F}_A} \iota_{q,n}(t) \frac{\gamma_{q,n}(t)}{W}, \tag{17}$$

with $\iota_{q,n}(t)$ denoting the indicator function of the event that MS q is scheduled to transmit on the nth subcarrier during scheduling period t, that is,

$$\iota_{q,n}(t) = \begin{cases} 1, & \text{if MS } q \text{ is scheduled on carrier } n \text{ in } t \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Using this definition, and taking into account that on each subcarrier n in region A, and after averaging over the distance to the BS, the MSs are statistically equivalent in terms of SINR for the PF scheduler [19], the conditional CDF in (15) is given by

$$F_{\gamma_n^A|M_A,\boldsymbol{d}}(x|k,\boldsymbol{d}) = \frac{1}{k} \sum_{u \in \mathcal{M}_A} F^k_{\gamma_{u,n}^A|d_{0,u}}\left(x|d_{0,u}\right). \tag{19}$$

Now, taking into account that on each subcarrier n in region A, and after averaging over the distance to the BS, the MSs are statistically equivalent in terms of SINR, the (unconditional) random variables $\{\gamma_{q,n}(t)\}_{\forall\, q \in \mathcal{M}_A}$ are i.i.d., and the conditional CDF in (15) can be obtained as

$$F^{\text{PF}}_{\gamma_n^A|M_A}(x|k) = \int_{R_L^A}^{R_U^A} F^k_{\gamma_{u,n}^A|d_{0,u}}\left(x|d\right) f_{d_{0,u}}(d) \mathrm{d}d, \tag{20}$$

where $f_{d_{0,u}}(d)$ is the PDF of the random variable $d_{0,u}$ that can be expressed as

$$f_{d_{0,u}}(d) = \frac{2d}{{R_U^A}^2 - {R_L^A}^2}, \; R_L^A \le d \le R_U^A. \tag{21}$$

Using (21), (20) and (15) in (13) or (14) and after some algebraic manipulations, the average throughput in cell layer A, for the PF scheduling rule, can be obtained as shown in (22) on top of the next page.

In order to analyze the capacity achieved by the worst MSs[4] of each layer, we define the edge of region A as a thin angular region with lower radius $R_L^{A,\text{edge}} = R_U^A - \delta$ and upper radius $R_U^{A,\text{edge}} = R_U^A$, where $\delta \le \left(R_U^A - R_L^A\right)/2$.

B. RR scheduling

A RR scheduler allocates subcarriers to MSs in a fair time-sharing approach. Since the SINRs experienced by MSs in region A on each subcarrier n are statistically equivalent, serving $M_A = k$ MSs using a RR scheduling policy is equivalent to serving $M_A = 1$ MS with PF (even when MSs are selected with non uniform probability). Therefore, the conditional CDF in (15) simplifies to

$$F^{\text{RR}}_{\gamma_n^A}(x) = F^{\text{PF}}_{\gamma_n^A|M_A}(x|1). \tag{23}$$

Finally, using (23), (21), and (15) in (13) or (14) and after some algebraic manipulations, the average throughput in the cell layer A , for the RR scheduling rule, can be obtained as shown in (24) on top of the next page.

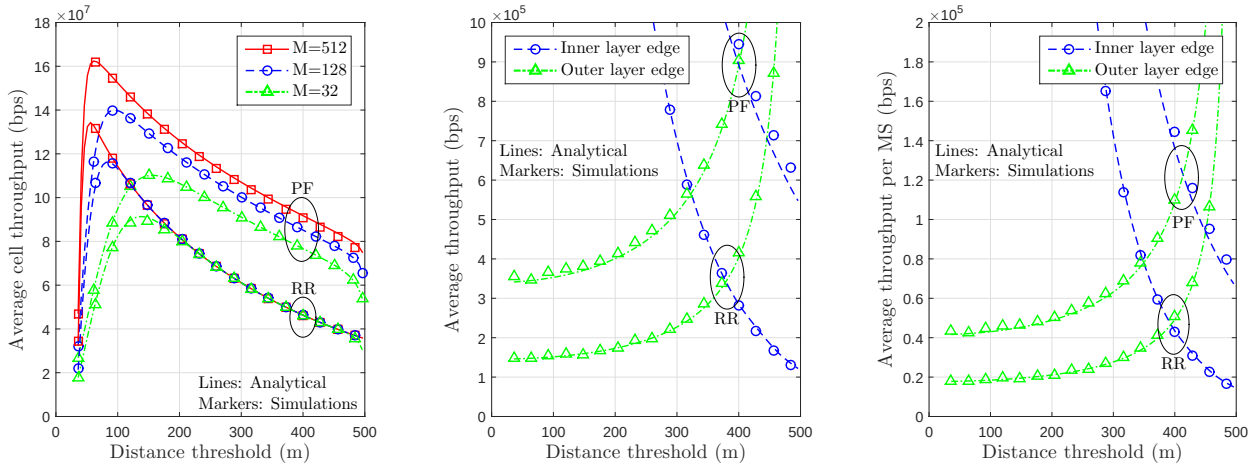
V. PERFORMANCE EVALUATION

In order to validate the proposed analytical framework, a 19-cell network is considered, where the cell of interest is surrounded by two rings of interfering BSs (see Figs. 1 and 2). As stated in previous sections, MSs are distributed over the coverage area using a PPP of normalized intensity λ (measured in MSs per area unit). For the sake of

[4]When the shadowing is not taken into consideration, the worst MSs are located in the edge region of each layer.

$$\text{PF: } \overline{\eta}^A = \frac{2\log_2 eN_A\Delta f}{R_U^{A\,2} - R_L^{A\,2}} \int_0^{\infty} \int_{R_L^A}^{R_U^A} \left(1 - \exp\left[-\pi\lambda(R^2 - R_0^2)P_r^A \left(1 - F_{\gamma_{u,n}^A|d_{0,u}}(x|d)\right)\right]\right) \frac{d}{1+x} \mathrm{d}d\ \mathrm{d}x. \tag{22}$$

$$\text{RR: } \overline{\eta}^A = \frac{2\log_2 eN_A\Delta f}{R_U^{A\,2} - R_L^{A\,2}} \left(1 - \exp\left[-\pi\lambda(R^2 - R_0^2)P_r^A\right]\right) \int_0^{\infty} \int_{R_L^A}^{R_U^A} \left(1 - F_{\gamma_{u,n}^A|d_{0,u}}(x|d)\right) \frac{d}{1+x} \mathrm{d}d\ \mathrm{d}x. \tag{24}$$



(a) Total cell throughput (b) Layer's edge throughput, $M = 512$ (c) Layer's edge throughput per MS

Fig. 3: Average cell throughput, layer's edge throughput and layer's edge throughput per MS, for both RR and PF scheduling policies (two-layer FFR).

Table I: Network parameters

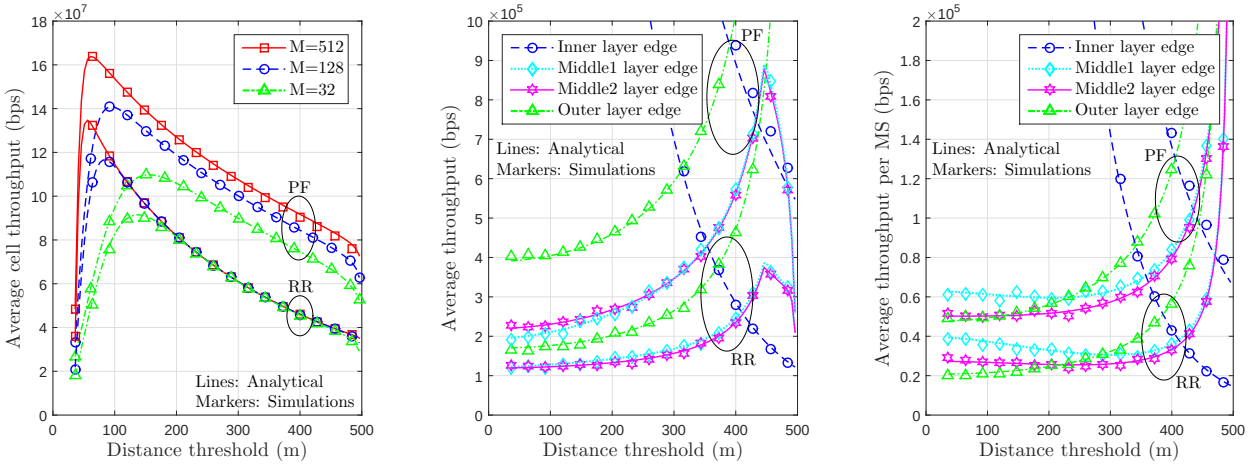
System parameter	Value
Cell radius	500 m
Minimum distance between BS and MSs	35 m
Distance δ defining the layer edge	4 m
Transmit power at the BS	46 dBm
Antenna gain at the BS	14 dBi
Noise power spectral density	-174 dBm/Hz
Receiver noise figure	7 dB
Total bandwidth	20 MHz
Subcarrier spacing	15 kHz
Occupied subcarriers (including DC)	1201
Number of inner subcarriers	624
Number of middle subcarriers	32
Path loss model (dB)	$15.3 + 37.6\log_{10}(d)$
Monte Carlo trials	1,000

presentation clarity, results in this section will be shown as a function of the average number of MSs per cell ($M \triangleq \pi\lambda(R^2 - R_0^2)$) The main system parameters used to generate both the analytical and simulation results are based on [20] and are summarized in Table I.

Illustrating the system behaviour, results in Figure 3 are provided applying the two-layer FFR scheme and using both PF and RR scheduling policies. For the sake of clarity, lines are used to represent the analytical results and markers correspond to Monte-Carlo simulations. It is worth noting the very good agreement between the simulated and analytical results, thus validating the novel mathematical framework.

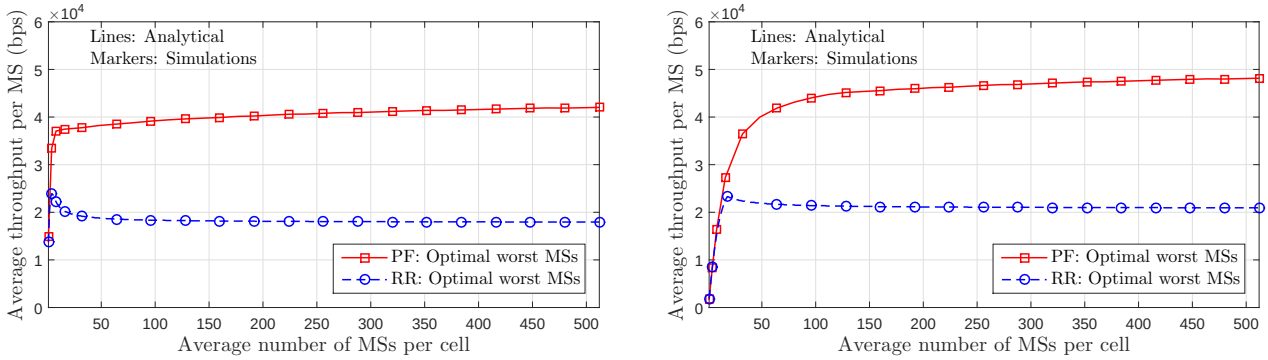
Focusing now on performance aspects, Fig. 3a presents the average throughput, considering the whole coverage cell, as a function of the distance threshold R_{th}. As expected, PF outperforms RR because PF is a channel-aware scheduler exploiting the multiuser diversity. The maximum average cell throughput increases with the average number of MSs per cell. This is basically due to two distinct effects. The first one, only exploited by the PF scheduling rule, is caused by the degree of multiuser diversity provided by the increase of M. The second effect, affecting all the schedulers but more noticeable when using the RR scheduler, is because increasing the average number of MSs per cell raises the probabilities of having at least one inner MS and one outer MS, hence reducing the probability of waste of resources.

Figure 3b presents the average throughput of the worst MSs typically located at each layer's edge, as a function of the distance threshold R_{th}. Note that, as the inner layer becomes larger the throughput of the inner layer's edge decreases because the MSs located in the edge of the inner layer suffer from high levels of ICI, whereas the throughput of outer layer increases. This phenomenon is consistent with the results shown in Fig. 3c. In order to analyze the average capacity achieved by an arbitrary MS located in the edge of certain layer, this figure presets the layer's edge throughput per MS as a function of the distance threshold R_{th}. It is interesting to note that the worst MSs are not necessarily located in the outer layer.



(a) Total cell throughput (b) Layer's edge throughput, $M = 512$ (c) Layer's edge throughput per MS

Fig. 4: Average cell throughput, layer's edge throughput and layer's edge throughput per MS, for both RR and PF scheduling policies (multi-layer FFR).



(a) Two-layer FFR (b) Multi-layer FFR

Fig. 5: Optimal layer's edge throughput, for both RR and PF scheduling policies, and under both two-layer and multi-layer schemes).

Figure 4 shows the performance of the multi-layer scheme, the same trend is observed when comparing both scheduling policies (see Fig. 3). Checking the maximum average cell throughput in Fig. 4a, notice that the multi-layer FFR has only slightly improved the whole cell performance compared with the two-layer scheme, due to the increment of the spectrum utilization when the multi-layer scheme is used. Regardless of the scheduling policy in use, the middle1 layer performance is always better than the middle2 layer performance due to the fact that both of them hold the same reuse factor while the middle1 layer is nearer to the corresponding BS (see Figs. 4b and 4c). It is also interesting to note in Fig. 4c that, for a low value of R_{th}, the worst MSs are located in the edge of the outer layer, however, as R_{th} increases, the MSs located in the edge of the middle2 or the inner layer become the worst.

The optimization outcomes are shown in Fig. 5 when using both PF and RR scheduling policies, and under both two-layer and multi-layer schemes. Regardless of the scheme in use, for each maximum value of average cell throughput, the corresponding optimal value of the worst MSs' capacity increases little with M when using PF (e.g. for $M = 32$ under the two-layer FFR this value is equal to 37.76 Kbps, whereas for $M = 512$ this value is equal to 42.02 Kbps). In contrast, the corresponding optimal value of the worst MSs' capacity decreases little with M when using RR (e.g. for $M = 32$ under the two-layer FFR this value is equal to 19.14 Kbps, whereas for $M = 512$ this value is equal to 17.95 Kbps). The main advantage of the multi-layer scheme is that, without any sacrifice in spectral efficiency, it is able to provide higher levels of fairness between the MSs located accros the coverage area of the network, a QoS metric of paramount importance in beyond-4G cellular networks. In particular, for the multi-layer FFR, it can be observed that the average throughput for the worst MSs when $M = 512$ is equal to 48.16 Kbps when using PF and equal to 20.94 Kbps when using RR, corresponding to a 14.6% and 16.7% improvement, respectively, compared to the benchmark two-layer FFR scheme.

VI. CONCLUSION

This paper has presented and validated a novel analytical framework to evaluate the performance of the multi-layer FFR scheme in a downlink OFDMA-based cellular network. Mathematically tractable solutions have been derived for popular scheduling rules, namely, PF and RR. For the specific case of PF scheduling, the cell throughput improvement is further accentuated by a greater exploitation of the multiuser diversity. Remarkably, as the average number of MSs per cell increases, the maximum average cell throughput also increases whereas the optimal value of the worst MSs' capacity remains virtually constant. Results show that the multi-layer FFR scheme does not decrease the cell spectral efficiency or the average throughput of the whole cell while significantly increases the worst MSs' capacity throughout the cell at the cost of sacrificing the throughput of MSs located close to the BS. In other words, the multi-layer FFR scheme leads to an overall cell capacity virtually identical to that of two-layer FFR, but it is able to archive a higher degree of fairness. Further work will concentrate on the use of more sophisticated ICIC techniques (e.g., soft/adaptive frequency reuse schemes, higher order sectorization, network MIMO).

ACKNOWLEDGMENTS

Work supported by the Agencia Estatal de Investigación and Fondo Europeo de Desarrollo Regional (AEI/FEDER, UE) under project ELISA (subproject TEC2014-59255-C3-2-R), Ministerio de Economía y Competitividad (MINECO), Spain, and the Conselleria d'Educació, Cultura i Universitats (Govern de les Illes Balears) under grant FPI/1538/2013 (co-financed by the European Social Fund). The research leading to these results has also received funding from "la Caixa" Banking Foundation.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 2nd ed. Elsevier Science, 2013.

[2] A. S. Hamza, S. S. Khalifa, H. S. Hamza, and K. Elsayed, "A survey on inter-cell interference coordination techniques in OFDMA-based cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1642–1670, 2013.

[3] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in LTE-Advanced HetNets," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 113–122, 2013.

[4] L. Wang, F. Fang, N. Nikaein, and L. Cottatellucci, "An analytical framework for multilayer partial frequency reuse scheme design in mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7593–7605, 2016.

[5] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *The Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[6] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4294–4305, December 2011.

[7] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 996–1019, 2013.

[8] F. Jin, R. Zhang, and L. Hanzo, "Fractional frequency reuse aided twin-layer femtocell networks: Analysis, design and optimization," *IEEE Transactions on Communications*, vol. 61, no. 5, pp. 2074–2085, 2013.

[9] M. Assaad, "Optimal fractional frequency reuse (FFR) in multicellular OFDMA system," in *IEEE 68th Vehicular Technology Conference (VTC-Fall)*, 2008, pp. 1–5.

[10] A. Najjar, N. Hamdi, and A. Bouallegue, "Efficient frequency reuse scheme for multi-cell OFDMA systems," in *IEEE Symposium on Computers and Communications (ISCC)*, 2009, pp. 261–265.

[11] Z. Xu, G. Y. Li, C. Yang, and X. Zhu, "Throughput and optimal threshold for FFR schemes in OFDMA cellular networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2776–2785, 2012.

[12] J. García-Morales, G. Femenias, and F. Riera-Palou, "Analytical performance evaluation of OFDMA-based heterogeneous cellular networks using FFR," in *IEEE 81st Vehicular Technology Conference (VTC-Spring)*, 2015.

[13] G. Femenias and F. Riera-Palou, "Corrections to, and comments on, "Throughput and Optimal Threshold for FFR Schemes in OFDMA Cellular Networks"," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2926 – 2928, 2015.

[14] Z. Xie and B. Walke, "Performance analysis of reuse partitioning techniques in ofdma based cellular radio networks," in *Telecommunications (ICT), 2010 IEEE 17th International Conference on*. IEEE, 2010, pp. 272–279.

[15] R. Ghaffar and R. Knopp, "Fractional frequency reuse and interference suppression for ofdma networks," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*. IEEE, 2010, pp. 273–277.

[16] X. Yang, "A multilevel soft frequency reuse technique for wireless communication systems," *IEEE Communications Letters*, vol. 18, no. 11, pp. 1983–1986, 2014.

[17] L. Wang, F. Fang, K. Min, N. Nikaein, and L. Cottatellucci, "Toward multi-layer partial frequency reuse in future mobile communication systems," in *Communications in China (ICCC), 2014 IEEE/CIC International Conference on*. IEEE, 2014, pp. 647–652.

[18] M. Maqbool, P. Godlewski, M. Coupechoux, and J.-M. Kélif, "Analytical performance evaluation of various frequency reuse and scheduling schemes in cellular OFDMA networks," *Performance Evaluation*, vol. 67, no. 4, pp. 318–337, 2010.

[19] J. García-Morales, G. Femenias, and F. Riera-Palou, "On the analysis of channel-aware schedulers in OFDMA-based networks using FFR," in *11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2015, pp. 786–793.

[20] G. TR36.921, "Home enode b (HeNB) radio frequency (RF) requirements analysis (release 9)," v9.0.0. Mar. 2010.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6564

Mejora de la Calidad en Redes WLAN Coordinadas a través de SDWN

Julián Fernández Navajas, Luis Sequeira Villarreal, José Ruiz Mas, José Mª Saldaña Medina
Grupo CeNITEQ– Instituto de Investigación en Ingeniería de Aragón (I3A)
Dpt. IEC. EINA, Universidad de Zaragoza
Edif. Ada Byron, 50018, Zaragoza
{navajas, sequeira, jruiz, jsaldana}@unizar.es

***Resumen*- En el presente trabajo se propone una arquitectura capaz de mejorar la calidad de las comunicaciones multimedia en redes WLAN coordinadas, mediante la utilización de SDWN (Software Defined Wireless Networks). Para esto se requiere que soluciones SDN (software Defined Network) sean adaptadas para poder identificar los parámetros propios de las redes Wi-Fi, tales como interferencias, movilidad, selección de canales, etc. y sean capaces de abordar los problemas derivados de los requerimientos de calidad de las estaciones que comparten la misma red inalámbrica. En conclusión, este trabajo que tendrá formato de "work in progress" busca qué pruebas plantear en un entorno controlado de WLAN con SDWN para así mejorar la calidad de las comunicaciones multimedia.**

***Palabras Clave*- WLAN, SDN, SWDN, Wi-Fi**

I. INTRODUCCIÓN

En los últimos años estamos asistiendo a un incremento en el uso de las comunicaciones inalámbricas. Esto es debido a las facilidades de utilización por parte del usuario final, tanto en los dispositivos móviles como en las infraestructuras de red. Como contrapunto, estas tecnologías son más complejas de implementar, sobre todo si queremos desplegarlas en entornos amplios y de forma coordinada, como sucede en escenarios tales como aeropuertos, áreas urbanas, centros comerciales, etc.

En particular, la tecnología inalámbrica que mayor popularidad está adquiriendo es IEEE 802.11 (conocida popularmente como Wi-Fi). Ello explica que diferentes fabricantes hayan desarrollado soluciones que permitan coordinar múltiples puntos de acceso Wi-Fi para dar cobertura a una zona. Sin embargo, estas soluciones son cerradas y resultan poco flexibles frente a los nuevos retos de escalabilidad y de calidad de experiencia en las comunicaciones, que se plantean en los escenarios anteriormente citados y que se encuentran en constante evolución.

Una solución inmediata para Wi-Fi es cambiar a la banda de 5 GHz, que ofrece mejores prestaciones por estar menos saturada. Esto puede solucionar el problema por un corto periodo de tiempo pero, a la postre, volvería a aparecer. Por eso se deben buscar otras alternativas que proporcionen funcionalidades coordinadas en el balanceo de carga, la planificación de canales, etc.

En la actualidad existen trabajos que desarrollan diferentes propuestas para unificar todas estas posibles actuaciones de coordinación. Así en [1, 2] se propone como solución coordinada la utilización de SDN (Software Defined Networks) en estos entornos inalámbricos. Para esto se requiere que dichas infraestructuras SDN sean adaptadas, para poder así identificar los parámetros propios de las redes Wi-Fi, tales como interferencias, movilidad, selección de canales, etc., y sean capaces de abordar los problemas derivados de las necesidades de calidad de las estaciones (STA a partir de ahora) que comparten la misma red inalámbrica.

En Wi-Fi las STA disponen de sus propios algoritmos para seleccionar, de manera local y en base a la potencia de señal recibida, el punto de acceso (AP a partir de ahora) al que asociarse, de entre todos aquellos AP que pertenezcan a la red configurada. Además, cuando una STA detecta otro AP de la misma red, con mayor potencia de señal recibida, sus algoritmos pueden desencadenar un cambio de asociación (*handoff*) de un AP a otro. Una primera consecuencia es que las STA pueden llegar a asociarse a un AP que está saturado, en cuanto a tráfico soportado. Un segundo efecto negativo es que las STA pueden cambiar, sin previo aviso, de un AP a otro, lo que en el mejor de los

casos puede tardar varios cientos de milisegundos, con la consiguiente pérdida de paquetes, o incluso puede suponer la perdida de la comunicación a nivel IP. Todo ello supone un deterioro en la calidad, sobre todo para las comunicaciones en tiempo real.

Una posible solución al problema es inhibir este comportamiento y sustituirlo por un *seamless handoff* controlado por la red en el que las STA crean estar siempre conectadas al mismo AP. Con este fin, se propone una nueva arquitectura, basada en SDN, en la que, incorporando soluciones coordinadas como APs virtuales (LVAP, *Light Virtual Access Point*) a los que se conectan las STA, se mejore la calidad de las comunicaciones en la red inalámbrica.

En resumen, el objetivo del presente trabajo es la evaluación de la calidad de las comunicaciones multimedia existentes en una red Wi-Fi multicanal al introducir una arquitectura SDWN con funcionalidades inteligentes.

II. Trabajos relacionados

Una primera propuesta para solucionar los problemas ya comentados está descrita en [3]. La idea es que los APs reales utilicen diferentes LVAP, uno por cada STA presente en la red. De esta manera cada STA ve un único AP (LVAP), incluso si va cambiando de AP real a causa de la propia movilidad. Como consecuencia, las STA evitan la reasociación pero no el asociarse a un AP saturado. Una limitación de esta propuesta es que todos los AP deben trabajar en el mismo canal.

La anterior solución no aborda en profundidad el problema de cómo se gestiona el paso de un AP a otro. Para ello se plantea un segundo trabajo [4] en el que se define un protocolo distribuido entre los AP para informarse de la presencia de las diferentes STA. De nuevo está basado en el uso de LVAP y utiliza un único canal Wi-Fi. También hay que indicar que este método no entra en profundidad en el problema de que las STA se conecten a un AP saturado.

El siguiente paso, presentado en [5], consiste en definir un protocolo centralizado (coordinación) para la gestión de los AP y STA presentes, basado en SDN. Aparece la figura de un controlador que gestiona cada uno de los LVAP asociados a cada una de las STA. Así, el controlador puede decidir que una STA no se conecte a un AP que se encuentre saturado, mediante un algoritmo de balanceo de carga. Esto supone tener en cuenta la calidad de las comunicaciones para el correcto funcionamiento del sistema. Este método sigue trabajando en un único canal y presenta problemas de escalabilidad cuando crece el número de STA (o lo que es lo mismo, de LVAP) en la red, porque por cada LVAP, cada AP real debe enviar periódicamente un *beacon unicast* a cada STA.

III. Arquitectura SDWN propuesta

La arquitectura propuesta está basada en SDN, pero adaptada al entorno *wireless* que nos ocupa. Como ya se ha comentado, el objetivo es controlar los mecanismos por los cuales las STA, que tienen configurada una misma red a la que conectarse, se asocian a un determinado AP y se desasocian de éste para asociarse a otro que consideran con mejores prestaciones. Para ello vamos a enumerar una serie de condiciones que debemos cumplir en nuestro diseño:

-La arquitectura propuesta no debe requerir ningún cambio en las STA.

-Los usuarios se desplazan a una velocidad propia de quien va caminando y no contemplamos velocidades superiores.

-El sistema de control debe ser centralizado y tener clara la separación entre plano de control y plano de datos. De esta manera queremos que nuestra arquitectura esté desarrollada previendo una futura interacción con otras arquitecturas de movilidad como 3G o 4G.

-Los servicios escogidos a los que debemos proporcionar calidad son multimedia en tiempo real.

-El principal parámetro a tener en cuenta para la realización de la asociación y desasociación de los STA en los AP debe ser la calidad percibida por los usuarios de los servicios.

-Debemos trabajar en un entorno multicanal y tener presentes los problemas de escalabilidad.

En la Figura 1 se muestra la arquitectura SDN propuesta, donde están separados los planos de control y de datos. En ella se puede observar cómo un conjunto de aplicaciones de red inteligentes *(smart funcionalities)* interactúan *(northbound protocols)* con un plano de control *(controller)* para gobernar *(southbound protocols)* el funcionamiento y la transferencia de información en el plano de datos.

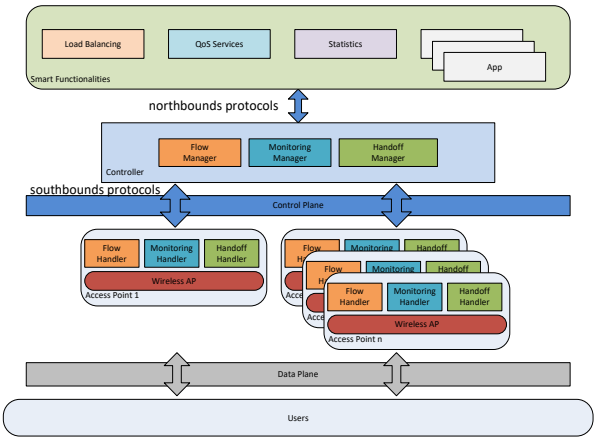


Fig. 1. Arquitectura SDWN propuesta.

Las aplicaciones de red nos permiten dotar a nuestra arquitectura de la inteligencia necesaria para llevar a cabo funcionalidades tales como *Load Balancing*, *QoS Services* o simplemente *Statistics*, entre otras, gracias a los procesos de gestión *(manager)* existentes en el controlador. Estos gestores, en acciones tanto conjuntas como separadas, rigen el comportamiento de procesos residentes en los elementos de red controlados:

-*Flow Handler*: proceso que permite identificar, ubicar y controlar los flujos de datos asociados a los usuarios. En el caso de Wi-Fi permite la gestión de elementos de

red como Virtual Switch (OpenFlow) y de LVAP y, por tanto, la identificación y (re)ubicación de STA.
-*Monitoring Handler*: proceso continuo de recogida de información relevante *(scanning)* asociada a los equipos y usuarios existentes en la red. En el caso de Wi-Fi sería información relevante de la red y servicios, como tasa, ancho de banda, tamaño paquetes, potencias, interferencias, etc.
-*Handoff handler*: proceso que genera las acciones necesarias de red para la (re)ubicación real de equipos de usuario. En el caso de Wi-Fi posibilita el cambio de AP y canal a una STA sin que necesite funciones adicionales que corten la comunicación a nivel IP o superiores (*Channel Switch Announcement*)

Una acción conjunta de los gestores sería, por ejemplo, el caso de un *seamless handover* debido a una redistribución de la carga de tráfico *(Load Balancing)*. Así, el gestor de *handoff* se pone en marcha tras concluir con el gestor de monitorización la necesidad de un traspaso de una STA hacia un determinado AP que reúne las condiciones adecuadas para mantener la calidad del servicio. Será el gestor de flujos el que propicie el traspaso de unos flujos de datos de un AP a otro AP haciendo los cambios oportunos en los elementos de control implicados (LVAP, *virtual switches* a través de OpenFlow, etc.).

IV. Implementación

La implementación llevada a cabo incorpora los elementos presentados en la anterior sección. Para demostrar la viabilidad de la propuesta, junto a las aplicaciones de red, la arquitectura desarrollada consta de un único controlador que puede comunicarse con varios AP que constituyen la red Wi-Fi, configurada con un único SSID al que se conectan las STA. En el plano de control se establece la comunicación entre el controlador y los AP mediante una red cableada (Ethernet conmutada). En el plano de datos se establecen las comunicaciones propias de los servicios presentes en las STA, y está formada por un segmento cableado que conecta los AP con la red de datos (Ethernet conmutada, *router* y salida a Internet) y los segmentos inalámbricos (diferentes canales Wi-Fi).

El proceso *Flow Handler* utiliza, como ya ha sido indicado, southbound protocols, entre ellos OpenFlow, para la gestión de elementos de red como *virtual switches* y LVAP y, por tanto, la identificación y (re)ubicación de STA. El proceso *Monitoring Handler*, además de la recogida continua de información vía *Radiotap*, añade a cada AP un interfaz Wi-Fi adicional que trabaja en un canal independientemente al canal de datos. Así, el proceso no interfiere en la calidad de las comunicaciones Wi-Fi, que es la más sensible. Asimismo, hemos de recordar que en la parte cableada existe una Ethernet conmutada que no tiene por qué dar problemas para la calidad de las comunicaciones.

El proceso de *handoff Handler* incluye un cambio de canal, especificado ya en la norma IEEE 802.11n, llamado *Channel Switch Announcement* (CSA). Dicho procedimiento está basado en el envío de *beacon* con el elemento CSA, y está pensado para que un AP anuncie a todas las STA conectadas que se va a cambiar de canal. Al final del proceso, tanto el AP como todas las STA cambian al canal destino anunciado en los *beacon* CSA. Este mecanismo no fuerza una desconexión IP. Este procedimiento lo vamos a utilizar para cada LVAP, que se ocupa de una única STA, por lo que los beacon deberán ser *unicast*. Además, en la implementación propuesta, el AP que comenzó el proceso no cambia de canal, sólo la STA a la que se le envía, consiguiendo así el *handoff* particularizado.

El tiempo entre *beacon* es un parámetro crítico en el CSA, dado que de éste depende el tiempo de interrupción de la comunicación IP. Si el tiempo es pequeño, disminuye el tiempo de interrupción de las comunicaciones IP y también sus efectos negativos sobre la calidad. Sin embargo, esto implica una mayor ocupación temporal del canal, pudiendo provocar congestión en el caso de que haya muchas STA. La solución propuesta es introducir un tiempo entre *beacon* variable, para que mientras no haya necesidad de CSA, se mantenga un valor de cientos de milisegundos. Sin embargo, llegado el momento del cambio de canal, disminuye a valores de decenas de milisegundos, enviando una ráfaga de una duración corta, pero lo suficiente como para que la STA haya realizado la operación con éxito.

Todo este procedimiento se ha puesto en marcha junto con la aplicación de red correspondiente para evaluar un *seamless handover*, tal como se puede observar en la Figura 2 y consultar en [8]. Los aspectos fundamentales para su correcto funcionamiento en la arquitectura desarrollada son:
-Definir un mapa de canales utilizados por los diferentes AP, de tal forma que cada AP sepa cuáles son los canales utilizados por los AP cercanos y poder así monitorizar las STA cercanas y con posibilidad de ser asociadas a nuestro AP.
-Disponer en cada AP de un interfaz Wi-Fi adicional en el que hacer las medidas propias en los diferentes canales sin interferir en el suyo propio.
-Definir adecuadamente las métricas y los algoritmos que nos permitirán, en base a la información de monitorización recibida, realizar las reasignaciones correspondientes que permitan el *seamless handover* que asegure los niveles de calidad deseados.
-Utilizar un intervalo temporal adecuado con el que se transmiten los *beacon* para conseguir un *seamless handover* óptimo. Cuanta mayor sea la frecuencia de la ráfaga de *beacon*, mejores resultados obtendremos en cuanto a rapidez. Pero si es demasiado elevada, es posible que inundemos la red con dicha información de control, lo que haría el sistema poco escalable. Por este motivo será importante analizar los efectos que tiene dicho parámetro en la calidad obtenida.

V. Escenario de pruebas

El scenario de pruebas (Figura 2) incluye dos AP (TP-Link AC1750 con OpenWrt 15.05) configurados en los canales 4 y 9 de Wi-Fi. También consta de una STA a la que forzaremos a desasociarse y reasociarse a

los AP. Por último disponemos de un controlador, un generador y un receptor de tráfico y un *sniffer*.

Este escenario, así configurado, evita los problemas de las capturas en el medio radio, puesto que el generador de tráfico no coincide con la STA (que sería lo esperable) sino que lo conectamos vía Ethernet a éste. De este modo el *sniffer* estará conectado al generador de tráfico por un lado y al receptor de tráfico por otro, lo que nos permite conseguir medidas de pérdidas y retardos mucho más precisas. Esto es válido porque se acepta que no hay pérdidas en la red Ethernet en comparación con la red Wi-Fi, y la capacidad de proceso del STA es tal que no añade retardo, al tratarse de un PC de propósito general. El retardo y las pérdidas están, por lo tanto, generados sólo por los dispositivos Wi-Fi y la red inalámbrica.

Además, a la STA podemos conectarle diferentes dispositivos Wi-Fi y así estudiar las diferencias de comportamiento sin necesidad de cambiar el escenario. Los dispositivos Wi-Fi escogidos para las pruebas son: Linksys WUSB54GC, WiPi WLAN USB b/g/n y TP-LINK TL-WN722N).

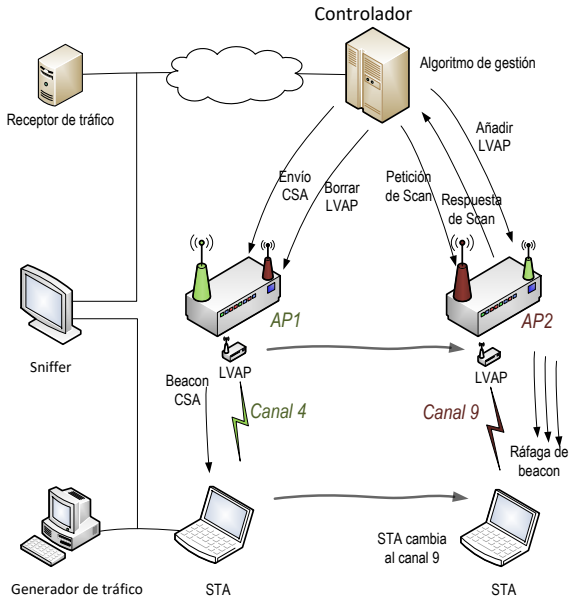


Fig. 2. Escenario de pruebas.

VI. Evaluación

El objetivo del trabajo no es presentar unos resultados concluyentes sino plantear la necesidad de nuevas pruebas que permitan analizar los efectos del *handoff* en la QoE (*Quality of Experince*) de servicios multimedia en tiempo real. Como ejemplo se presentan unas pruebas extraídas de [8] utilizando *Quake 3* (un juego *First Person Shooter sobre UDP)*, con restricciones de latencia, pero no de pérdidas. Para las pruebas se ha utilizado el generador de tráfico D-ITG [9] que dispone de los modelos de tráfico del juego. Se han realizado varias pruebas con diferentes valores de tiempo entre *beacon*: 5, 10, 20, 30, 40 y 50 ms. El controlador ha sido configurado para forzar el 20 *handoff* del STA cada 30 segundos.

Para obtener valores de QoE, se ha utilizado el G-Model [10] que nos proporciona un MOS. En la Tabla I, se presentan los resultados para diferentes dispositivos Wi-Fi. Su comportamientos puede ser muy diferentes a pesar de estar usando el mismo sistema de *handoff*.

Tabla I

Tiempo entre beacon	G-Model		
	Linksys	WiPi	TP-Link
5	0.27	4.19	4.06
10	1.10	4.23	4.16
20	1.15	4.17	4.25
30	1.21	4.28	4.25
40	1.37	4.26	4.26
50	2.72	4.28	4.26

VII. Conclusiones

En este trabajo se ha presentado un escenario de pruebas para evaluar la calidad de las comunicaciones multimedia en nuestra propuesta de redes WLAN coordinadas. Son tres las variables que podemos modificar: la aplicación multimedia, los dispositivos Wi-Fi y la arquitectura de WLAN coordinadas. Aunque no son resultados definitivos, cabe destacar que no existe una solución única y que los parámetros a escoger para nuestro sistema pueden ser acertados para ciertas configuraciones del entorno pero no para otras.

Agradecimientos

Este trabajo ha sido financiado parcialmente por la Comisión Europea a través del Proyecto Wi5 (H2020 G.A. no 644262), por la DGA y el fondo social europeo a través de CeNITEQ, y por el Gobierno de España a través del proyecto TISFIBE (TIN2015-64770-R).

Referencias

[1] R. Riggio, T. Rasheed, and M. K. Marina, “Poster: Programming software-defined wireless networks,” in Proc. MobiCom, 2014, pp. 413–416.

[2] R. Riggio, T. M. Rasheed, and R. Narayanan, “Virtual network functions orchestration in enterprise WLANs,” in Proc. IM, May 2015, pp. 1220–1225.

[3] Y. Grunenberger, F. Rousseau, “Virtual Access Points for Transparent Mobility in Wireless LANs”.: WCNC, 2010, p 1-6

[4] M. E. Berezin, F. Rousseau, A. Duda, “Multichannel Virtual Access Points for Seamless Handoffs” in IEEE 802.11 Wireless Networks., in: VTC Spring, IEEE, p. 15.

[5] J. Schulz-Zander, L. Suresh, N. Sarrar, A. Feldmann, T. Hhn, R. Merz, “Programmatic Orchestration of WiFi Networks”, in: 2014 USENIX Annual425 Technical Conference (USENIX ATC 14), USENIX Association, Philadelphia, PA, p. 347358

[6] “Vision: Augmenting WiFi Offloading with An Open-source Collaborative Platform”, in: Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services, MCS ’15, ACM, New York, NY, USA, 430 2015, p. 4448.

[7] M. E. Berezin, F. Rousseau, A. Duda, “Multichannel Virtual Access Points for Seamless Handoffs in IEEE 802.11 Wireless Networks”., in: VTC Spring, IEEE, 2011, p. 15.

[8] L. Sequeira, J. L. de la Cruz, J. Ruiz-Mas, J. Saldana, J. Fernandez-Navajas, J. L. Almodovar, “Building a SDN Enterprise WLAN Based On Virtual APs,” IEEE Communications Letters, Vol. 21, No. 2, pp 374-377, Feb. 2017.

[9] A. Botta, A. Dainotti, A. Pescap, “A tool for the generation of realistic network workload for emerging networking scenarios”, Comp. Network. vol. 56, no. 15, pp. 3531-3547, 2012.

[10] R. M. Pereira, L. M. R. Tarouco, “Adaptive Multiplexing Based on E-model for Reducing Network Overhead in Voice over IP Security Ensuring Conversation Quality”, in: 2009 Fourth International Conference on Digital Telecommunications, IEEE, 2009, p. 5358

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6572

Multivariate statistical technique over QoS variables to analyze video quality metrics on IEEE 802.11ac networks

Miguel García-Pineda, Santiago Felici-Castell, Jaume Segura-García
Dpt. Informàtica, Universitat de València
Burjassot, España
Email: migarpi@uv.es, felici@uv.es, jsegura@uv.es

***Resumen*—We present the results from a measurement-based performance evaluation of wireless networks based on IEEE 802.11ac standard in an indoor environment, with the aim to analyze their performance under high definition streaming video applications. We focus our study on analyzing the highest performance of these standards using off-the-shelf equipment as well as the behavior of Quality of Service variables and how they affect to the video quality. Thus, we have analyzed and measured these variables and have applied a multivariate statistical technique, called Factor Analysis, and finally discuss their behavior.**

***Palabras Clave*—multivariate statistical technique, factor analysis, video quality metrics, quality of service, quality of experience**

I. INTRODUCTION

Big companies like Google, Cisco Systems, Apple and Microsoft predict that by 2020, more than 90% of Internet traffic will be multimedia content (images, 3D images, High Definition (HD) video and audio, etc.) [1]. With these applications, streaming around home as well as mobility has become an issue.

The IEEE 802.11 Working Group has approved several standards to deliver gigabit rates in Wireless Local Area Networks (WLAN), in particular with the IEEE 802.11ac standard, working only in the 5 GHz band. Theoretically, IEEE 802.11ac has expected multi-station WLAN throughput of 1 gigabit per second (Gbps) and single link throughput of 500 megabits per second (Mbps).

Nevertheless, the actual network performance results in real experimental environments are quite different and require a further analysis [2] [3]. Then, it is very important to perform a comprehensive analysis of IEEE 802.11ac network using relatively new off-the-shelf commercial products to validate these specifications, under the scenarios for which they were designed, in particular for video streaming with Full and Ultra HD (or 4K) videos under a subjective video quality point of view or Quality of Experience (QoE), measured in terms of Mean Opinion Score (MOS).

Thus, in order to detect critical issues in this process, we measure and analyze different variables in the streaming process related to the physical layer as well as Quality of Service (QoS) parameters. As many factors are affecting QoE, that could hide relevant information, we apply a multivariate statistical technique, called Factor Analysis (FA) [4] to reduce the whole set of measurable variables and to find out which ones have influence on the subjective video quality. This technique is also used in big data science. This could ease the design of new objective video quality metrics to estimate or predict the MOS, denoted by $\widehat{MOS}$. It must be noticed that this last step is out of the scope of this paper, because it would require more space, both to include the design as well as the performance evaluation comparing with relevant metrics.

About IEEE 802.11ac standard, it introduces enhancements to the IEEE 802.11n, which is based on MIMO (Multiple Input Multiple Output). It contains many advanced features designed to improve the user experience, including wider radio frequency bandwidth (up to 160 MHz), more MIMO spatial streams (up to 8), Multi-User MIMO (MU-MIMO) and high-density modulation (up to 256-QAM). The standard was developed between 2011 and 2013, while devices compliant with it were released by 2015. This 5 GHz band has several advantages over 2.4 GHz networks, for example they have non-overlapping channels (unlike 2.4 GHz channels) and more channels are available for higher throughputs. Nevertheless, 5 GHz signals suffer from greater attenuation that can be mitigated using beamforming and MIMO techniques.

The goal of this paper is to analyze the performance of IEEE 802.11ac standard under real indoor deployments from a video quality metric point of view, in particular when using Full and Ultra HD (or 4K) resolutions, while we analyze of the behavior of the measured variables

throughout the streaming system in the WLAN to detect which ones have more influence on the video quality.

II. RELATED WORK AND DISCUSSION

First, we present the recent literature related to the evaluation of IEEE 802.11ac using real scenarios and equipments.

M. Dianu et al. [2] study the performance of IEEE 802.11ac in an indoor environment using UDP traffic and WPA-2 encryption. They conclude that for short distances, IEEE 802.11ac offers significantly better performance compared to IEEE 802.11n with data rates exceeding 700 Mbps for a 3x3 MIMO configuration. But these improvements are very sensitive to the channel conditions with the achieved data rates rapidly declining as the distance between the transmitter and the receiver increases. Y. Zeng et al. [5] present an early performance characterization of IEEE 802.11ac standard, analyzing the impact of utilizing wider channel widths on energy efficiency and interference, that are basically the main basis for IEEE 802.11ac. The authors show that 80 MHz channel width yields substantial throughput improvement, but the improvements come at the cost of higher power consumption. The authors conclude that increasing the number of spatial streams is more energy efficient compared to increasing the channel width in achieving the same percentage increase in throughput. However, it is worth mentioning that the number of streams depends basically on the adapter, where most of them only support 1 or 2 spatial streams. In addition, the authors confirm that IEEE 802.11ac link suffers severe unfairness issues when it coexists with legacy IEEE 802.11 as confirmed in [2]. M. Abu-Tair et al. [3] analyze the performance and the energy efficiency of IEEE 802.11n and IEEE 802.11ac, and by comparing the results at 5 GHz, the authors find that IEEE 802.11ac achieves only 8% more throughput than IEEE 802.11n. Finally, M. Li et al. [6] propose a QoE-aware scheduling scheme for video streaming over IEEE 802.11n/ac networks with high density of users, tunning packet delay and channel transmission rate and enhancing the video quality, measured in terms of objective Peak Signal-to-Noise Ratio (PSNR).

About video quality metrics, they are classified as Full Reference (FR) and Non Reference (NR) [7]. FR approaches require access both the original (or reference) and the received video, while NR does not require the original video. In particular, NR metrics are more interesting from the practical point of view. Using similar steps as we show in this paper, in [8] the authors design NR video quality metrics based on bitstream, assisted by FR metrics, for real-time network monitoring. Notice that other NR standarized metrics are O.23 (ITU-T P.1201 [9]) and G.1070 [10].

In order to design new video quality metrics we can find in the literature different techniques. Basically, they are classified in regression techniques [11] or machine learning (artificial neural networks techniques). On one hand, when we work with regression techniques, we work with all variables regardless of their inter dependences. Then with this approach, when using complex functions to fit the regression, the number of variables hamper to find out the coefficients. Thus as we suggest in this paper, FA is an interesting alternative by reducing the dimensionality of the variables, by grouping them. On the other hand, machine learning techniques and neural networks require large number of datasets, are extremely time consuming and computationally intensive [12].

Fig. 1: Network deployment. Detail of the Video Streaming Server, the IEEE 802.11ac/n Access Point and Wireless clients.

It is worth mentioning that in [13] the authors use FA to model the data traffic generated by video streaming applications when using High Definition resolutions. Their goal is to define a model to support a better understanding of video stream workload characteristics and their impact on network traffic, to help in the network scheduling and resource allocation fields. Using the same technique, in [14], it shows a comprehensive simulation analysis of LTE Discontinuous Reception (DRX), allowing the Base Station (BS) to schedule User Equipments (UE) for periodic wake/sleep cycles to save energy. The authors employ FA applied to variables related directly to the DRX configuration, to determine their impact on several applications, in particular for streaming video.

We conclude from the previous works related to IEEE 802.11ac performance evaluation, although they provide valuable information with different thorough studies, none of them have analyzed the performance in a real scenario of video streaming from a subjective point of view, that is basically the main reason for this type of standards with gigabit throughputs and by analyzing the measured variables we can help improving the design of these networks or devise new video quality metrics.

III. EXPERIMENT DESIGN AND TEST-BED DESCRIPTION

In order to carry out the performance analysis of the WLAN we have used the topology described in Fig. 1, where the streaming video server is connected via Gigabit Ethernet to the Access Point (AP) and the wireless clients with USB Wi-Fi adapters are located at 5 m distance from the AP as in a usual domestic environment. For the measurements, we used a student lab (Lab 1.1.6 at

Computer Science Dpt. in our university) that is 20 x 15 square meters, with two lateral walls of glass and the other two of wood/polyester.

In our testbed, we have used off-the-shelf equipments with default configurations. The hardware description of the equipment is as follows:

1) *Access Point*: model Linksys WRT 1900 AC dual band (2.4 & 5GHz) Gigabit Wi-Fi router [15] (firmware 1.1.8.161917) with network standards IEEE 802.11a/b/g/n/ac with 4 adjustable R-SMA antennas and maximum data rates 600 and 1300 Mbps for 2.4 and 5 GHz respectively. In particular the user manual of the AP says that it implements a draft version of the IEEE 802.11ac
2) *Wireless Adapters*: four USB Linksys AC1200 wireless adapters WUSB6300 [16] (firmware 1027.5.105.2015) with 2 antennas and maximum rates 300 and 867 Mbps for 2.4 and 5 GHz respectively
3) *Clients and server*: one server connected to the access point through Gigabit Ethernet and four clients equipped with aforementioned Wi-Fi adapter. All these computers are equipped with Intel Core i7 processor, 16 GiB RAM with USB port 3.0. Clients are static on the lab's table and at the same distance from the AP.

With this configuration of antennas between the access point and receivers, MIMO is working with 4x2:2 (maximum number of transmit antennas x maximum number of receive antennas : maximum number of data spatial streams). IEEE 802.11ac only operates in 5 GHz and we will use a channel bandwidth of 80MHz. In addition we do not use frame aggregation, allowing a fair distribution of the bandwidth as well as reducing the delay and the jitter.

The measurements have been conducted in the student labs of our university and we defined two different scenarios. It should be noted that we have different APs in the neighborhood although in different channels, as in a real environment.

In the first scenario, we measure the congestion throughput generating traffic between the server and several clients at 5 m from the AP. This will show us the base line of available bandwidth. In this case, each Wi-Fi client runs several simultaneous streaming connections (flows). To generate synthetic traffic we used LAN Traffic v2 (Enhanced software) [17], using UDP protocol with maximum packet sizes in IPv4 without fragmentation, 1460 and 1472 bytes respectively as in a normal video streaming scenario. To reach the saturation point of the network in order to determine the maximum throughput values, we generated packets continuously, without delay between sent packets.

In the second scenario, we analyze the performance of the WLAN with different simultaneous flows of video streaming using different wireless clients. As we expect high throughputs in these Gigabit WLAN, we used the "Big Buck Bunny" video sequence available at [18] with a duration of 1 minute with two different resolutions:

Fig. 2: A frame example from Big Buck Bunny video sequence

Full HD (HD 1920x1080) and 4K (Ultra-HD UHD, 3840x2160) with 30 and 60 frame per second (fps). These resolutions now on will be denoted as A, B, C and D for HD@30, HD@60, 4K@30 and 4K@60 respectively. Fig. 2 shows a frame of this movie. We use the codec H.264/Advanced Video Coding (AVC) with profiles High@L4.1 for HD 30 fps, High@L4.2 for HD 60 fps and High@L5.1 for UHD-4K (both 30 and 60 fps), with a variable and adaptive GOP size. The video streaming is done using MPEG-TS (Moving Picture Experts Group-Transport Stream) as container over RTP (Real Time Protocol) and UDP. The maximum bit rates for the different streams are 16.7, 19.7, 35.1 and 37.8 Mbps respectively for A, B, C and D.

We consider the maximum number of video streams simultaneously allowed, while the subjective video quality of the streams, measured in terms of Mean Opinion Score (MOS), does not fall by more than 50% of the initial quality. That is, it will happen when start to appear permanently errors such as blocking, blurriness, freezing, etc. To asses MOS, we have used traditional subjective MOS through surveys following the recommendations given by ITU-R (BT.500-13 [19], P.910 [20]). The evaluators provide one rating for the overall video quality using a discrete five-level scale ranging from Bad (1) to Excellent (5).

The study was conducted over two sessions, each lasting less than half an hour with video sequences of 10 s approximately, as recommended in [19] in order to minimize evaluator fatigue. The evaluators' pool consisted of 35 under-graduate students of different ages (on average 21 years old) from our university. They are students (male and female) with a male majority, of the last course of Multimedia Engineering and they have sufficient knowledge about multimedia streaming. It should be noted that although no vision test was performed, a verbal confirmation of soundness of (corrected) vision from the subject was taken to be sufficient. For the surveys, we used the same computers described above with 22" Samsung monitor.We filtered out the measurements given by evaluators whose scores were out of a range, given by the mean and $\pm$ two times the standard deviation. The subjective MOS has been calculated to meet a confidence interval of 95%.

For video streaming, we use the open-source FFmpeg tool [21] that allows both the configuration of the video streaming server and the client. In this case we use a buffer jitter of 500 ms. In addition, we have used Wireshark [22], an open-source packet analyzer to capture packets with IEEE 802.11 radio information sent and measure the signal strength (in terms of Receive Signal Strength Indication (RSSI)) as well as Signal Quality (SQ) given by the adapter, Throughput (Th.), Lost Packets (LP), packet Length (Len), Delay (D) and Jitter (J). In addition to these variable, FFmpeg tool provides in their reports a muxing OverHead (OH) included also a an input variable to be monitored.

IV. STATISTICAL MODELING: FACTOR ANALYSIS

Factor Analysis (FA) is a multivariate statistical method used to identify the factors underlying the variables by means of grouping related variables under the same factor, where we make the assumption that an underlying causal model exists. It is a dimension reduction technique by reducing the large number of variables into few factors without sacrificing much, the power of explained variability by the variables. In other words, we can describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors [4], denoted by F_i where i is the factor identifier. Variables are grouped into different factors on the basis of their interrelation. These factors partly explain the behavior of global variance of the variables. Thus, factors are unmeasured variables, defined as a linear combination of a reduced set of measurable variables, as follows:

$$F_i = constant_i + \sum_{j=1} \alpha_j \cdot variable_j; \qquad (1)$$

where α_j are coefficients of the linear regressions for the factors.

Variables may belong to more than one factor, but by using the factor rotation technique, these factors may be made mutually exclusive. These steps allow us to analyze the system as a linear application, where the weights of the different components are given by their eigenvalues. Thus, we set a criteria to group these variables under the same factor, when they show eigenvalues greater than one. Thus, in FA the extraction of the principal components is done by analyzing the correlation matrix of the different variables. Once we have a reduced set of components, we apply a rotation of these factors to obtain an orthogonal set by means of the *Varimax* algorithm and the *Kaiser* normalization [23].

It must be stressed that because FA is based on the correlation between the different variables, for each measurable variable we could even consider the first four standardized moments of the measured variables: *mean*, *standard deviation*, *skewness* and *kurtosis* in as similar way as shown in [13]. All of these statistics are also considered as new variables within FA. In this paper, for simplicity we will consider only the first two standardized moments.

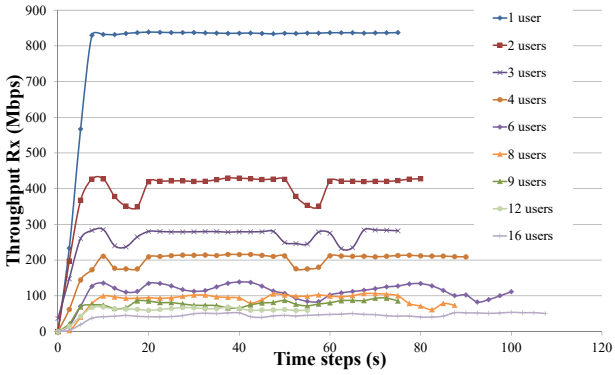


Fig. 3: Congestion Throughput in Mbps with different number of users with UDP traffic, for IEEE 802.11ac, with 80 MHz channel bandwidth

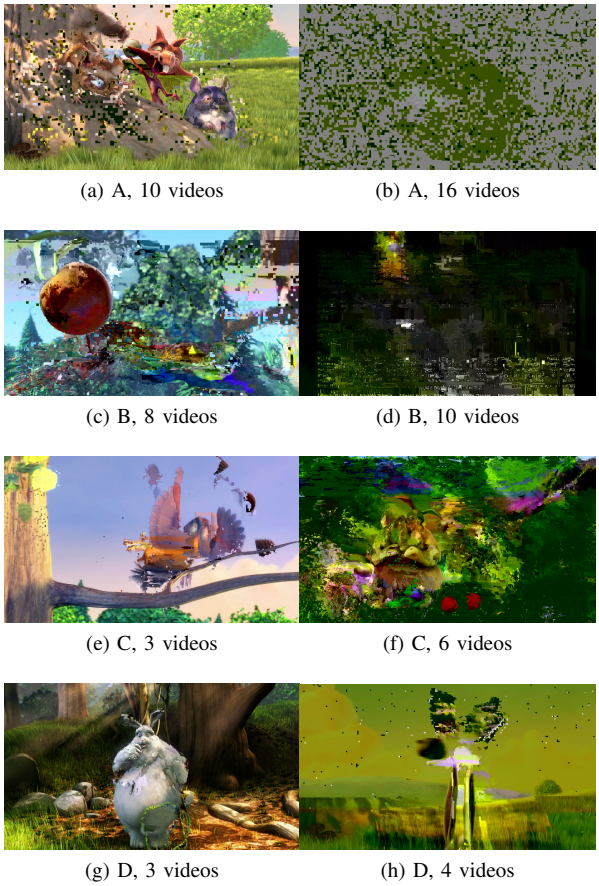
Using these factors, we could devise new video quality metrics such as $\widehat{MOS} = f(F_i)$, where $f()$ defines a regression function over these factors. This step is out of the scope of this paper.

V. MONITORED VARIABLES

With the first scenario, Fig. 3 shows the temporal variation of the throughput in the time line till it reaches the saturation point with UDP traffic. It is noteworthy that when the number of users increases, the deviation of the throughput also increases. As we can see, the available bandwidth is 840 Mbps approximately, that it is shared between the different clients.

In the second scenario and taking into account the subjective video quality, we determine the maximum number of video streams just when MOS values fall by more than 50% compared with the original MOS. Under these conditions, the maximum number of flows allowed are 20, 20, 6 and 5 for A, B, C and D resolutions respectively. Fig. 4 shows different errors and artifacts produced in the simultaneous video streams. These artifacts produce significant subjective errors, such as blocking, blurriness, freezing, etc. It is worth mentioning that the errors in HD streams are significantly more noticeable than in UHD (4K) streams.

Now, we will analyze the behavior of these measured variables. In Table I, we describe these variables and their abbreviations. Fig. 5 shows variables related to the physical layer, that is the RSSI at the Wi-Fi adapter when receiving the video streams at different resolutions (A, B, C and D) and number of simultaneous clients, as well as average Signal Quality (%) given by the adapter. In order to show the RSSI fluctuations of the signal, we plot the standard deviation for each resolution on RSSI. We show the mean RSSI between different resolutions against clients as well.



(a) A, 10 videos (b) A, 16 videos

(c) B, 8 videos (d) B, 10 videos

(e) C, 3 videos (f) C, 6 videos

(g) D, 3 videos (h) D, 4 videos

Fig. 4: Received frames with several errors for different resolutions (denoted as A, B, C and D), fps and number of simultaneous streaming videos

Tabla I: Measured variables and abbreviations.

Variable [unit]	abbreviation
Res [A..B]	Resolution
aRSSI [dBm]	avg. RSSI
sRSSI [dBm]	std. RSSI
rRRSI [dBm]	range RSSI
aSQ [%]	avg. SQ
OH [%]	overhead
LP [%]	Lost Packets
mD [ms]	mean Delay
sD [ms]	std. Delay
mJ [ms]	mean Jitter
sJ [ms]	std. Jitter
mT [Mbps]	mean Throughput
sT [Mbps]	std. Throughput
mLen [byte]	mean Length Packet
sLen [byte]	std. Length Packet

Fig. 6 shows all the measured variables related to QoS, such as Jitter (ms), Delay (ms), Th. (Mbps) and Packet Length (bytes) both with mean and standard deviation, as well as Lost Packets (%), for different video resolutions (A, B, C and D) against clients. At a first glance, it can be seen from this figure, that values remain constant except for Jitter. This is due to many things. Both clients and servers are on the same LAN thus the time uncertainty is reduced as well as the MAC algorithms split or share the available bandwidth in a fair way because of the high throughput. However, these arguments do not go with Jitter, that clearly increases as the number of clients increase, basically as a result of the traffic as well as the internal queue management in the AP.

In addition, we measure MOS values as shown in Fig. 7. It is important to highlight that we see a video quality improvement when increasing fps. Besides, subjective errors due to blocking, blurriness, freezing, etc., that highly impact the subjective video quality, appear randomly independently of the resolution.

As a preliminary step, in Fig. 8 we show the correlation coefficients for the different variables against MOS. By analyzing the significance level of this correlation coefficients, we see that $aRSSI, sRSSI, rRRSI, aSQ, mD$ and mJ are not correlated with MOS.

VI. STATISTICAL ANALYSIS AND RESULTS

We have used *SPSS.r22* software package for statistical analysis and Matlab R2015a. Following the steps explained in Section IV, to perform a FA, first we have to analyze the correlation matrix, shown in Table II, on the basis of which variables are grouped into factors. Based on this information, we determine the number factors and map the variables into them, taking into account the variance explained of the data. In this case, we find that with 4 factors we explain 89.46% of the total variance. Each factor explains 49.93, 19.13, 11.48 and 8.93% respectively. The results of this mapping process are shown in Table III, where each variable is assigned (after performing a Varimax rotation), based on the Pearson's coefficient (R^2), to each factor. At this point from our results and for this scenario, on one hand it worth mentioning that some factors are less relevant than others, in particular F_3 and F_4 that only explain 11.48 and 8.93% respectively of the total variance. These factors include variables related to the physical layer (RSSI and SQ). On the other hand, on F_1 we see the main critical variables.

Once we have found out the factors composition, using linear regression we define the expressions that determine them, as shown in Table IV along with their R^2 coefficients. Notice that $OH, sLen, mJ$ from F_1 and mT from F_2 have been excluded when modeling these factors due to their low contribution.

Now, with these factors (a reduced data set) we could perform different regression techniques in order to devise a NR new video quality and would assist the design of new NR video quality metrics, but this is out of the scope of this paper. Besides, it must be stressed that the presented data is biased version because we did not include all data. We skipped data related to bad MOS measurements because we focused only on the performance evaluation of IEEE 802.11ac.

VII. CONCLUSION

In this paper we have studied the measurement-based performance of IEEE 802.11ac standard, both analyzing the maximum saturation throughput and the maximum number of high definition streaming video flows under a subjective video quality point of view. From these results,

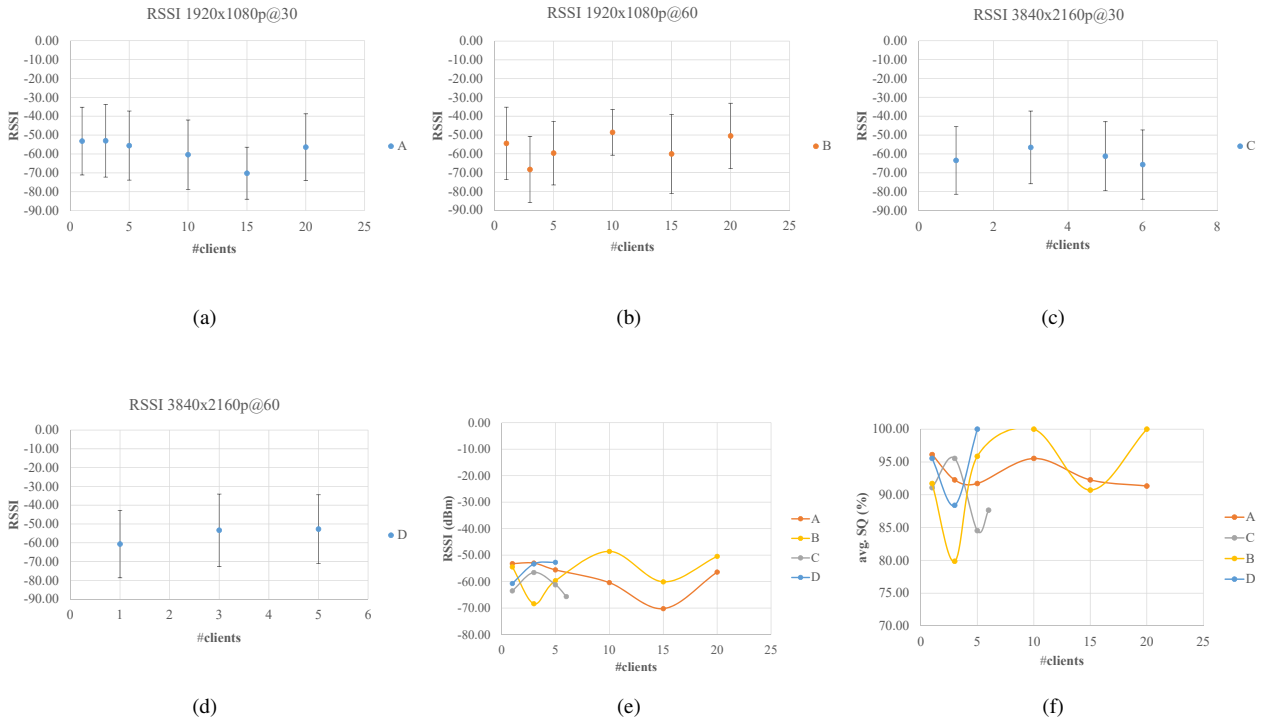


Fig. 5: RSSI (dBm) with detail of standard deviation and average Signal Quality (%) for the different video resolutions (denoted as A, B, C and D) against clients

Tabla II: Correlation matrix for the different measured variables.

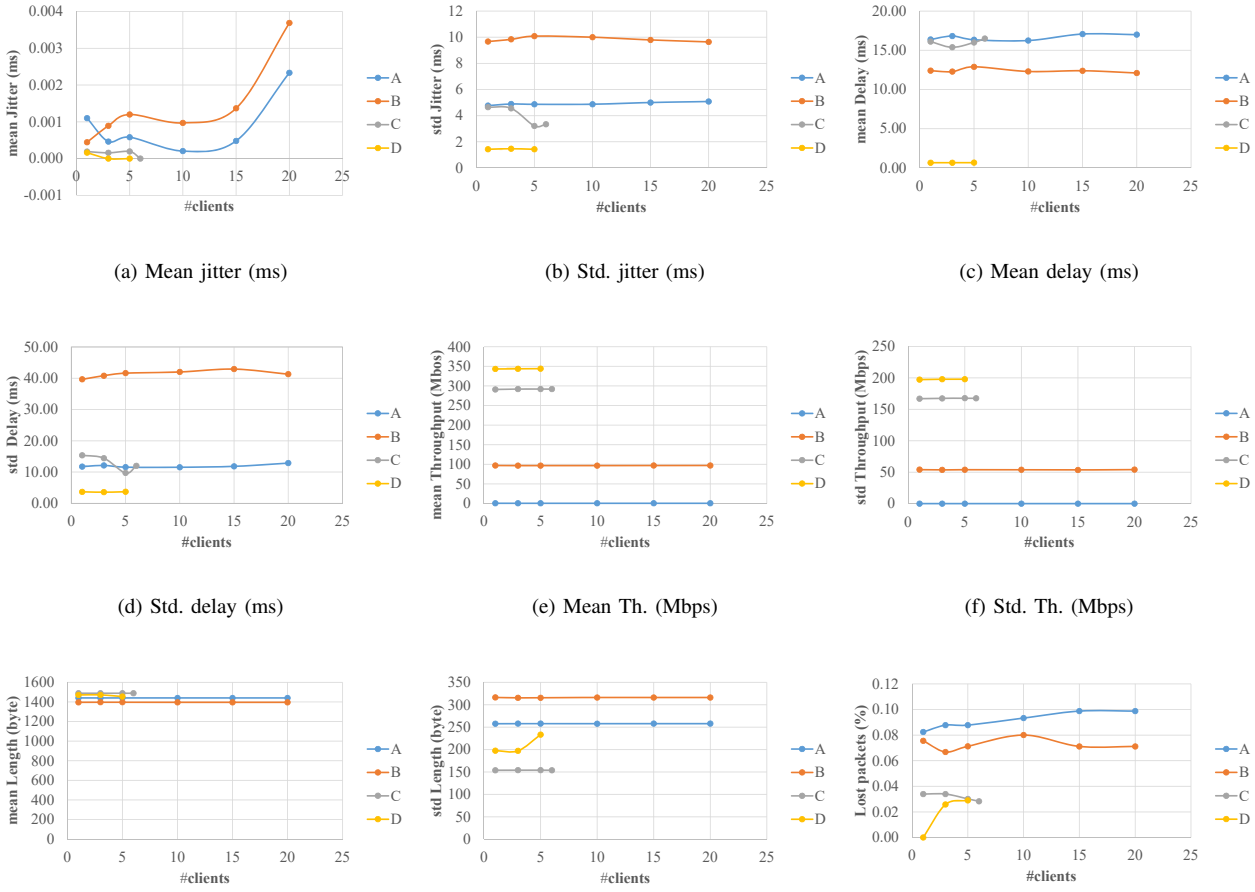
	Res	*aRSSI*	*sRSSI*	*rRRSI*	*aSQ*	*OH*	*LP*	*mD*	*sD*	*mJ*	*sJ*	*mT*	*sT*	*mLen*	*sLen*
Res	1	0.01	0.19	-0.18	-0.03	-0.52	-0.95	-0.75	-0.27	-0.41	-0.46	0.98	0.98	0.49	-0.55
aRSSI		1	-0.07	0.01	0.62	0.20	0.10	-0.25	0.09	0.31	0.07	-0.06	-0.06	-0.23	0.25
sRSSI			1	0.72	-0.46	-0.33	-0.29	0.10	-0.17	-0.16	-0.21	0.26	0.26	0.34	-0.37
rRRSI				1	-0.31	-0.24	0.07	0.43	-0.16	-0.22	-0.10	-0.08	-0.08	0.28	-0.29
aSQ					1	0.14	0.12	-0.20	0.03	0.27	0.04	-0.09	-0.09	-0.18	0.22
OH						1	0.65	0.11	0.86	0.60	0.90	-0.65	-0.66	-0.99	0.98
LP							1	0.61	0.36	0.46	0.54	-0.97	-0.97	-0.63	0.69
mD								1	0.17	0.19	0.33	-0.61	-0.61	-0.06	0.04
sD									1	0.54	0.98	-0.35	-0.36	-0.85	0.76
mJ										1	0.57	-0.46	-0.47	-0.59	0.57
sJ											1	-0.53	-0.53	-0.88	0.81
mT												1	1	0.63	-0.69
sT													1	0.63	-0.70
mLen														1	-0.99
sLen															1

we see that this standard fails to match the expectations created. It must be pointed out that these results heavily depend on the implementation of the IEEE 802.11ac standard of the commercial access points. These comments agree with [2], [3] because the network performance results in real environments are quite disparate from those we should expect from the standard. Nevertheless, although commercial products are not still 100% IEEE 802.11ac compliant, they actually can support several simultaneous UHD (4K) video streams at home (in our case, till 3 with an excellent quality).

Finally, we see how using FA we define a reduced data set, useful in particular when managing many variables and we want to devise a new NR video metric. This technique is used in big data science. In addition, we have seen that the measured variables related to the physical layer, are not relevant in the design of new NR video quality metrics at least if we reach a certain threshold of RSSI.

ACKNOWLEDGMENT

This work was supported by the Universitat de València under the projects UV-INV-PRECOMP14-207134, UV-INVAE15-339582, by the Generalitat Valenciana under the



(a) Mean jitter (ms) (b) Std. jitter (ms) (c) Mean delay (ms)

(d) Std. delay (ms) (e) Mean Th. (Mbps) (f) Std. Th. (Mbps)

(g) Mean p.length (byte) (h) Std. p.length (byte) (i) Lost packets (%)

Fig. 6: Relevant measured variables jitter (ms), delay (ms), throughput (Th.) (Mbps), packet length (bytes) with mean and standard deviation as well as lost packets (%), for different video resolutions against clients

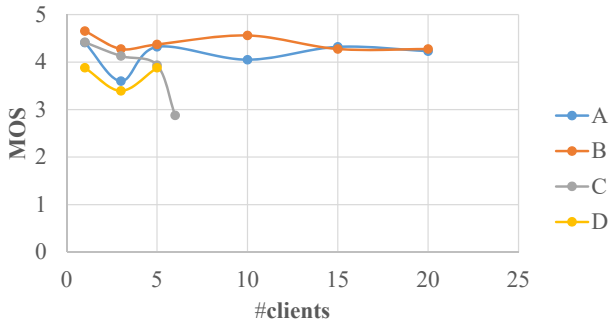


Fig. 7: MOS values for the different video resolutions (denoted as A, B, C and D) against clients

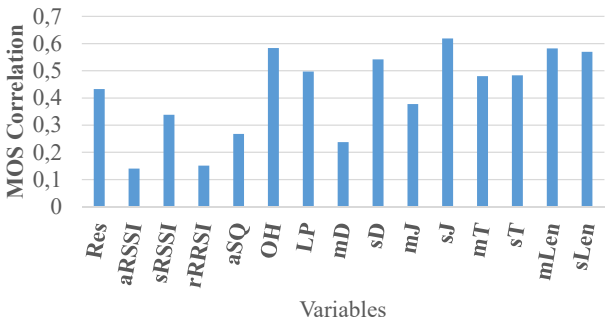


Fig. 8: Correlation coefficients of the measured variables with MOS

project GV-2016-002 and by the Ministry of Economy under the project BIA2016-76957-C3-1-R.

REFERENCIAS

[1] Cisco, "Visual Networking Index Global Mobile Data Traffic Forecast, 20172020. White paper." Cisco Systems, Corp., Tech. Rep., 02 2017. [Online]. Available: http://www.cisco.com/c/en/us/

Tabla III: Rotated component matrix with their loading factor

Variable	F_1	F_2	F_3	F_4
sD	**0.96**	-0.05	0.00	-0.06
sJ	**0.93**	-0.26	0.00	-0.06
OH	**0.92**	-0.31	-0.18	0.09
mLen	**-0.91**	0.28	0.21	-0.13
sLen	**0.84**	-0.36	-0.25	0.18
mJ	**0.59**	-0.27	-0.05	0.33
Res	-0.24	**0.97**	-0.01	-0.01
mT	-0.36	**0.92**	0.08	-0.07
sT	-0.37	**0.91**	0.08	-0.07
LP	0.36	**-0.90**	-0.09	0.10
mD	0.00	**-0.78**	0.30	-0.26
sRSSI	-0.12	0.18	**0.91**	-0.13
rRRSI	-0.19	-0.24	**0.89**	-0.03
aRSSI	0.15	0.06	0.11	**0.93**
aSQ	0.01	-0.03	-0.37	**0.82**

Tabla IV: Expressions of the factors

F_1=	$18.713 + 0.084 \cdot sD - 0.013 \cdot mLen - 0.236 \cdot sJ; (R^2 = 0.974)$
F_2=	$-5.304 + 3.282 \cdot Res - 0.031 \cdot sT + 0.09 \cdot mD - 7.917 \cdot LP; (R^2 = 0.993)$
F_3=	$-10.057 + 0.252 \cdot sRSSI + 0.114 \cdot rRRSI; (R^2 = 0.939)$
F_4=	$-0.424 + 0.114 \cdot aRSSI + 0.076 \cdot aSQ; (R^2 = 0.956)$

solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] M.-D. Dianu, J. Riihijarvi, and M. Petrova, "Measurement-based study of the performance of IEEE 802.11ac in an indoor environment," in *Communications (ICC), 2014 IEEE International Conference on*, June 2014, pp. 5771–5776.

[3] M. Abu-Tair and S. Bhatti, "Upgrading 802.11 deployments: A Critical Examination of Performance," in *Advanced Information Networking and Applications (AINA), 2015 IEEE 29th International Conference on*, March 2015, pp. 844–851.

[4] R. L. Gorsuch, *Factor analysis*. Hillsdale: Lawrence Erlbaum Associates, 1983.

[5] Y. Zeng, P. Pathak, and P. Mohapatra, "A first look at 802.11ac in action: Energy efficiency and interference characterization," in *Networking Conference, 2014 IFIP*, June 2014, pp. 1–9.

[6] M. Li, P. H. Tan, S. Sun, and Y. H. Chew, "Qoe-aware scheduling for video streaming in 802.11n/ac-based high user density networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[7] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 165–182, June 2011.

[8] I. Sedano, K. Brunnström, M. Kihl, and A. Aurelius, "Full-reference video quality metric assisted the development of no-reference bitstream video quality metrics for real-time network monitoring," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–15, 2014. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2014-4

[9] ITU-T, "Parametric non-intrusive assessment of audiovisual media streaming quality," *P.1201 Amendment 2*, vol. Series P: Terminals and subjective and objective assessment methods, 12 2013. [Online]. Available: http://www.itu.int/rec/T-REC-P.1201-201312-I!Amd2

[10] ——, "Opinion Model for Video-Telephony Applications." *G.1070, 2012a*, vol. Series G: Transmission systems and media, digital systems and networks, 7 2012. [Online]. Available: https://www.itu.int/rec/T-REC-G.1070-201207-I/en

[11] M. Alreshoodi and J. Woods, "Survey on qoe\qos correlation models for multimedia services," *International Journal of Distributed and Parallel Systems (IJDPS)*, vol. 4, 2013. [Online]. Available: http://arxiv.org/abs/1306.0221

[12] S. Aroussi and A. Mellouk, "Survey on machine learning-based qoe-qos correlation models," in *2014 International Conference on Computing, Management and Telecommunications (ComManTel)*, April 2014, pp. 200–204.

[13] A. K. A. Tamimi, R. Jain, and C. So-In, "Statistical analysis and modeling of high definition video traces," in *ICME*. IEEE, 2010, pp. 596–601.

[14] G. Stea and A. Virdis, "A comprehensive simulation analysis of LTE Discontinuous Reception (DRX)," *Computer Networks*, vol. 73, no. 0, pp. 22 – 40, 2014.

[15] "Linksys WRT 1900 AC dual band," 2015, accessed: 1/10/2015. [Online]. Available: http://www.linksys.com/us/p/P-WRT1900AC

[16] "Linksys USB AC1200 wireless adapters WUSB6300," 2015, accessed: 1/10/2015. [Online]. Available: http://www.linksys.com/us/p/P-WUSB6300/

[17] "LAN Traffic v2 (Enhanced software)," 2015, accessed: 8/2/2015. [Online]. Available: http://www.zti-communications.com/documentation/LanTrafficV2Enhanced_UserGuide.pdf

[18] "Big Buck Bunny, Sunflower version." 2013, accessed: 15/9/2015. [Online]. Available: http://bbb3d.renderfarming.net/download.html

[19] ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures," *BT.500-13*, 1 2013. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf

[20] ——, "Subjective Video Quality Assessment Methods for Multimedia Applications," *P.910*, 9 1999.

[21] FFmpeg, "Tools," 2016. [Online]. Available: https://www.ffmpeg.org/

[22] A. Orebaugh, G. Ramirez, J. Burke, and L. Pesce, *Wireshark & Ethereal Network Protocol Analyzer Toolkit (Jay Beale's Open Source Security)*. Syngress Publishing, 2006.

[23] H. T. Kaiser, "The Application of Electronic Computers to factor analysis," *Educational and Psychological Measurement*, vol. 20, p. 141151, 1960.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6573

QoE en el contexto de *Internet of Everything*

Maria-Dolores Cano

Dpto. Tecnologías de la Información y las Comunicaciones

[1]Universidad Politécnica de Cartagena

Campus Muralla del Mar, Edif. Antigones, s/n 30202 Cartagena (España)

mdolores.cano@upct.es

***Resumen*- La investigación y el desarrollo hacia la *Internet of Everything* (IoE) inteligente es un empeño ambicioso y altamente interdisciplinar que debe abordarse en los diferentes niveles de su arquitectura. Un efecto colateral de esta separación se produce en la medición de la calidad de los servicios proporcionados y en cómo evaluar sus prestaciones. La calidad de experiencia de usuario (*Quality of user Experience*, QoE) se ha convertido en un marco de referencia de evaluación de prestaciones capaz de acoger los nuevos requisitos y demandas de este nuevo paradigma. En este artículo, proporcionamos una visión actual de la QoE en la IoE. Para ello, realizamos un estudio de las propuestas de medición de QoE, de las métricas que se sugieren para su modelado y sus relaciones. Como resultado, se identifica la necesidad de un nuevo enfoque para la evaluación de prestaciones de los servicios y aplicaciones de la IoE capaz de capturar su idiosincrasia, es decir, las nuevas métricas que definen calidad, nivel de conocimiento, nivel de inteligencia, consumo, etc., así como la carencia de un modelado entre los diferentes componentes de la QoE en IoE.**

***Palabras Clave*- IoE, IoT, QoE, QoD, QoI.**

I. Introducción

Nos encontramos en una situación única en la que nuestro universo físico está adquiriendo una nueva existencia digital, donde cualquier ser/objeto va a estar conectado y capacitado para comunicarse y colaborar. En esta nueva realidad, con millones de dispositivos conectándose a la nube para intercambiar, procesar, y almacenar información, paradigma de *Internet of Everything* (IoE) [1], la arquitectura de red debe adaptar de forma cognitiva tanto sus principales parámetros de capacidad/caudal y latencia, como la calidad de sus servicios, de la forma más ágil, inteligente y eficiente posible. Es decir, como si se tratase de un servicio público básico (*utility service*) en función de las demandas de sus productores/consumidores de información.

En este marco general, el proyecto AIM (*Augmenting Intelligence, cognitive capabilities, efficiency and value of coMmunication technologies for the IoE*) pretende avanzar en las nuevas tecnologías de red para que permitan crear valor y cubrir efectivamente las necesidades futuras de comunicación y colaboración de una sociedad evolucionada. Durante la última década, se ha realizado un esfuerzo de investigación importante en el ámbito de *Internet of Things* (IoT), en particular en el avance de las tecnologías habilitadoras [2] (aquéllas que aseguran su correcta operación y permiten el despliegue de aplicaciones elementales y servicios sectoriales). Sin embargo, el trabajo relacionado con la adaptabilidad y las capacidades cognitivas de IoE donde "todas las cosas" deberían mejorar su disposición para aprender, comprender y actuar en el mundo físico, el digital y el social no ha hecho más que comenzar. Desde las TIC, entendemos que el aumento de la inteligencia en IoE debe interpretarse como la capacidad de los sistemas que la componen de aprender de su experiencia y en base a ésta adaptar/mejorar sus prestaciones [3], permitiendo la asignación inteligente de recursos, la operación automática de red y la provisión inteligente de servicios.

La investigación y el desarrollo hacia la IoE inteligente es un empeño ambicioso y altamente interdisciplinar que debe abordarse en los diferentes niveles de su arquitectura. Se han publicado numerosos trabajos describiendo arquitecturas IoE multicapa [4-6].

A grandes rasgos, existen tres niveles que interactúan: (i) la red de comunicación (y especialmente su tecnología de subred de acceso), (ii) la nube (cloud), y (iii) la capa de aplicación. En este contexto, el principal objetivo del proyecto AIM es contribuir a la extensión, inteligencia, adaptabilidad cognitiva, eficiencia y aporte de valor del paradigma de IoE, explorando soluciones TIC factibles en cada uno de los niveles citados. De entre los diferentes bloques de trabajo que componen este proyecto, nos centraremos en este artículo en la capa de aplicación de IoE y en el papel que la calidad de experiencia de usuario (*Quality of user Experience*, QoE) puede tener en este entorno.

El resto del artículo se organiza como se indica a continuación. La sección II describe la capa de aplicación en IoE y cómo esta se convierte en un elemento que proporciona valor añadido y donde la QoE, en términos de cómo alcanzarla y en cómo medirla, se convierte en un reto a resolver. La sección III explora los modelos de evaluación de la QoE para IoE presentados en la literatura especializada. Finalmente, el artículo finalizada en la sección IV, donde se discute la situación actual en base a la información revisada y se propone el trabajo futuro a realizar.

II. La capa de aplicación en IoE

Nos centramos en este apartado en describir la creación de valor añadido en la capa de aplicación de IoE, más allá de las expectativas estándar, aportando ventajas competitivas. El aumento de la inteligencia, de la eficiencia y del número de dispositivos conectados en IoE proporcionará al sector productivo no sólo mejores prestaciones sino también la posibilidad de explotar la información fruto de los propios procesos, por lo que la creación de valor será un elemento más de la cadena de suministro [7]. Aquí surgen nuevas cuestiones: ¿cómo diseñar y llevar a cabo la necesaria provisión de recursos para implementar un servicio/aplicación IoE? ¿Cómo evaluar las prestaciones de estos servicios/aplicaciones para satisfacer niveles apropiados de calidad de servicio (*Quality of Service*, QoS) o de experiencia (*Quality of user Experience*, QoE)? ¿Qué implica para una compañía/usuario final una aplicación cognitiva en IoE? ¿Cómo se puede medir la inteligencia alcanzable por el despliegue de tales servicios/aplicaciones? Etc. Las respuestas pasan por evaluar las prestaciones de IoE, lo que no es tarea sencilla. IoE proporciona un entorno de creación de servicios dinámico para nuevas aplicaciones y medios y herramientas para monitorizar los procesos de negocio o los parámetros de red, pero a la vez, está sujeta a restricciones comunes como la energética, la adaptación en tiempo real a la adquisición de datos o la tolerancia a fallos. Si bien los modelos de evaluación de prestaciones para redes tradicionales han sido ampliamente estudiados, para IoE están todavía en sus inicios [8].

Algunos autores coinciden en la necesidad de tener en cuenta nuevos factores en la evaluación de prestaciones de IoE, y en la definición de nuevas métricas para capturar el comportamiento tanto de capas individuales como de la arquitectura completa. Existen experiencias preliminares del Grupo de investigación de Ingeniería Telemática de la Universidad Politécnica de Cartagena al respecto [9]. Como evolución natural de trabajos de investigación previos, proponemos seguir las indicaciones de uno de los pocos artículos con esta visión sobre el nuevo paradigma cognitivo IoE [10]. En él, se dividen las métricas en dos dimensiones, coste y beneficio. La dimensión beneficio se refiere a métricas tales como calidad de los datos, de la información y de experiencia. La primera pretende evaluar la calidad de los datos adquiridos a nivel de dispositivo y hace referencia por ejemplo a su exactitud, veracidad, integridad o actualidad. La calidad de la información está íntimamente ligada a la información a nivel de toma de decisiones, una vez la comunican y procesan el dispositivo, las redes y la nube. Por último, la QoE se contempla como "*the overall acceptability of an application or service, as perceived subjectively by the end user*", según se define por la International Telecommunications Union (ITU) [11]. Esta métrica se puede extender más allá de la clásica provisión de QoS, teniendo en cuenta por ejemplo, el acceso a la IoE de dispositivos móviles o estáticos mediante una conexión apropiada, la capacidad de la comunicación para garantizar la ejecución de una aplicación, la disponibilidad de suficientes recursos de computación a nivel del *cloud* o el efecto del servicio que percibe el usuario final. Por otro lado, la dimensión coste permite estimar la eficiencia de la utilización de los recursos. La eficiencia de recursos incluye figuras de mérito relacionadas con la eficiencia en la utilización de múltiples dispositivos o individual, la carga computacional en que se incurre en escenarios multi-aplicación, el consumo energético en los distintos niveles de la arquitectura, la eficiencia de los recursos de almacenamiento, etc. Los trabajos de investigación actuales están todavía lejos de una visión de IoE cognitiva. La definición, la caracterización y la evaluación de estas métricas de evaluación de prestaciones contribuirán a conocer mejor los sistemas IoE a gran escala y a cerrar la brecha actual entre teoría y práctica para dimensionar y planificar a nivel de sistema los servicios y aplicaciones.

III. Identificación de métricas

Como se ha indicado anteriormente, Wu propone en [10] una separación de los componentes que determinan las prestaciones en términos de calidad para la IoE en tres capas: calidad de datos (*Quality of Data*, QoD), calidad de la información (*Quality of Información*, QoI) y calidad de experiencia de usuario (*Quality of user Experience*, QoE) como se muestra en la Tabla I. Esta misma clasificación de prestaciones en tres niveles, entiéndase QoD, QoI y QoE, también se emplea en [12]. En este trabajo, los autores intentan fusionar los conceptos de *Social Internet of Things*

(SIoT) y *Cognitive Internet of Things* (CIoT) dentro del paradigma de IoE, presentando un ejemplo de implementación con *smart software agents*.

Para la QoD, Wu [10] se basa en el trabajo anterior realizado por Bellavista *et al.* [13], donde se propone medir la calidad de los datos recolectados, ya sea a través de la captura mediante sensores u otros componentes/servicios, y su impacto en el contexto. Respecto a QoI, su definición la extrae de [14], donde se introduce un modelo paramétrico de medida dado por (1), donde las componentes están normalizadas. Claramente podemos observar que hay métricas duplicadas en los niveles QoD y QoI, por ejemplo la precisión. Según [10], la diferencia radica en la unidad lógica que se está midiendo, siendo en QoD el dato en sí mientras que en QoI es la información, con un nivel más alto de procesado. No obstante, este (posible) solape puede ser contradictorio como discutiremos más adelante.

Por último, se incluye en la taxonomía la capa QoE. En ella se diferencian cuatro subniveles (véase la Tabla I). El subnivel comunicación entendemos que se correspondería con las métricas clásicas de QoS, esto es, pérdidas de paquetes, retardo, *jitter* y ancho de banda. También podrían añadirse aquí otras métricas empleadas en modelos paramétricos de QoE obtenidos a partir de mediciones objetivas de un servicio, como por el ejemplo el tiempo de *buffering*, el *codec* empleado en una transmisión de video o la tasa de codificación de bits. Si además juntamos este subnivel de comunicación con el subnivel acceso, entendemos que nos encontramos aquí con los modelos de medición de QoE propuestos en la literatura especializada, como pueden ser el modelo E (*E-model*, [15]) para Voz sobre IP (VoIP), que específicamente incluye un parámetro denominado Access y que hace referencia al tipo de acceso al servicio que tiene el usuario, o los modelos ITU-T P.1201 [16] e ITU-T P.1202 [17] para la medición de QoE en servicios de transmisión de video. El subnivel aplicación intentaría capturar aquellos parámetros subjetivos propios del consumo de un servicio, y cuyo impacto en la QoE ha quedado demostrado en trabajos previos, como son la interfaz gráfica o la facilidad de uso [18] [19].

$$QoI = Q \cdot P \cdot R \cdot A \cdot D \cdot T \cdot V \qquad (1)$$

Un aspecto diferenciador de la propuesta de evaluación de prestaciones de calidad de [10] es la introducción de un elemento "negativo", el coste, al que hemos denominado *Quality Cost* (QC), a semejanza de los modelos de análisis coste-beneficio muy empleados en teoría de decisión en disciplinas como la economía o la biología, entre otras. Aunque sí que existen algunos trabajos en la literatura especializada que abordan la compensación entre QoS y eficiencia energética [20-22], no se había planteado desde una perspectiva más amplia para QoE y con otros factores reductores de la calidad. Si miramos en detalle la Tabla I, el efecto de la eficiencia computacional y del almacenamiento de información se verá sin duda minimizado por la integración y colaboración entre la IoE y el *cloud*. No obstante, el uso o la delegación de estas tareas al *cloud*

Tabla I
DESCRIPCIÓN DE LA PROPUESTA DE MÉTRICAS DE EVALUACIÓN DE PRESTACIONES INTRODUCIDA EN [10]

Capa o Nivel	Métrica	Descripción	Beneficio	Coste
Quality of Data (QoD)	*Data accuracy*	Precisión	X	
	Data truthfulness	Grado de fiabilidad	X	
	Data completeness	Ratio de datos adquiridos sobre requeridos	X	
	Data up-to-dateness	Validez en la toma de decisión	X	
Quality of Information (QoI)	*Quantity (Q)*	Cuánta información válida se ha obtenido para la toma de decisión	X	
	Precission (P)	Proporción de información relevante respecto a toda la información capturada por sensores/redes/servicios	X	
	Recall (R)	Proporción de información relevante respecto a toda la información capturada sin sensores/redes/servicios	X	
	Accuracy (A)	Grado de precisión según los requisitos del tomador de decisiones	X	
	Detail (D)	Grado de completación de cara al tomador de decisiones	X	
	Timeliness (T)	Relación entre la información obtenida y el momento de uso para la toma de decisión (inversamente proporcional al retardo)	X	
	Validity (V)	Veracidad de la información obtenida	X	
Quality of user Experience (QoE)	*Access*	Conexión	X	
	Communication	Mediciones de la calidad de la comunicación, QoS	X	
	Resources	Disponibilidad de recursos, por ejemplo para servicios con altas demandas de computación	X	
	Application	Variables que afectan directamente al usuario (p.e., facilidad de uso)	X	
Quality Cost (QC)	*Device utilization efficiency*	Grado eficiencia en la utilización de los recursos		X
	Computation efficiency	Grado de eficiencia computacional (recursos de cómputo en la CIoT)		X
	Energy efficiency	Grado de eficiencia energética de los dispositivos en toda su perspectiva: cómputo, acceso a red, comunicación, etc.		X
	Storage efficiency	Grado de habilidad para el almacenamiento físico (y su gestión) de la información		X

podría tener un impacto en otras métricas, como por ejemplo un aumento en el gasto energético derivado de una mayor necesidad de comunicación con la red (en particular con la nube) o por ejemplo una mayor probabilidad de error en el envío de los datos (impactando entonces en la QoS).

Otra propuesta de evaluación de calidad en IoE se presentó en [23]. Los autores proponen dividir la QoE en diferentes dominios, asociados a los diferentes niveles de una arquitectura IoT/IoE, y medir la QoE como la combinación de la calidad en cada dominio individual. Los niveles empleados son cinco: dispositivos físicos, red, combinación (en referencia a la virtualización y la integración con la nube), aplicación y contexto. La Tabla II resume la descripción de cada uno de ellos. Los autores evalúan de forma práctica su modelo con un despliegue experimental que también incluye mediciones subjetivas de calidad para comparación. Aunque obtienen un modelo simple de medición de QoE, éste es muy limitado en cuanto a variables (métricas) empleadas. Como principal resultado destacan la baja influencia de la calidad de video sobre la QoE final, aunque no queda claro en qué capa de su propuesta se ubicaría (*physical layer* o *network layer*). Desde un punto de vista más social, algunos investigadores [24] proponen asociar directamente la QoE en IoE a Net Promoter Score (NPS) [25]. NPS es un indicador de la lealtad de los clientes a un producto basado en una única pregunta que responden los clientes: "con qué probabilidad recomendarías este producto o servicio a un familiar o amigo?". Esta propuesta, aunque interesante, presenta bastantes controversias, la principal su excesiva simplicidad.

En [26], los autores sugieren un modelo para la medición de la QoE en IoT también basado en niveles, en este caso tres: usuario, servicio y contexto (véase la Tabla III). La explicación de cada una de estas tres componentes es muy limitada y aunque los autores llevan a cabo un análisis posterior, para éste sólo emplean como métricas el tiempo de respuesta, la tasa de transmisión, la estabilidad y la precisión, limitando su aplicabilidad.

Por último, hacemos referencia a un interesante trabajo realizado en [27]. En este caso, el modelado de QoE se hace respecto a cuatro capas: dispositivos, red, *computing*, e interfaz de usuario (véase Tabla IV). Como novedad adicional, se definen dos tipos de métricas: métricas físicas y métricas metafísicas. Para cada capa se definen un conjunto (limitado) de métricas físicas, como por ejemplo resolución de cámaras (capa *device*), ancho de banda (capa *network*), prestaciones del servidor en la nube (capa *computing*) o presentación de la aplicación (capa *user interface*). Respecto a las métricas metafísicas, éstas intentan capturar la característica cognitiva de la IoT, y para ello los autores proponen relacionarlas con el logro de un objetivo. En concreto, como propuesta inicial, las asocian a obtener información del contexto o a disponer de un servicio cómodo. Una vez definidas e identificadas las métricas metafísicas (p.e. precisión, cantidad, *timeliness*, etc.) los autores modelan la relación entre métricas (físicas y metafísicas) y a continuación, modelan la relación entre las métricas metafísicas y la QoE de la aplicación. No obstante, los autores dejan justamente el modelado como trabajo futuro a desarrollar.

Tabla II
MÉTRICAS IDENTIFICADAS PARA LA EVALUACIÓN DE QOE EN IOT/IOE SEGÚN [23]

Dominio o nivel	Descripción	Asociación con modelo [10] de Tabla I
Physical devices	Medición de la calidad que proporcionan los elementos físicos de la IoE	QoD, QoE (QoS)
Network layer	Prestaciones de la infraestructura de red sobre la que opera la IoE	QoE (QoS)
Combination layer	Medición de la interacción con el *cloud*. Objetos virtuales y su combinación interaccionando con objetos físicos	QoI, QoE
Application layer	Calidad en términos de control, interactividad, presentación y usabilidad	QoE
Context layer	Efecto del contexto de uso, p.e., tipo de dispositivo, tipo de consumidor, precio, ubicación geográfica,…	QoE

Tabla III
MÉTRICAS IDENTIFICADAS PARA LA EVALUACIÓN DE QOE EN IOT/IOE SEGÚN [26]

Dominio o nivel	Descripción	Asociación con modelo [10] de Tabla I
User	Métricas: *background, experience, expectation, physiological and mental state*	QoE
Service	Métricas: *Transmission layer* (p.e., métricas de red), *application layer* (p.e., conversión, resolución de video, etc.), *service layer* (prioridades)	QoE, QoI
Context	*Physical context* (condiciones *hardware* and *software*) y *external context* (entorno social, cultural, etc.)	QoE

Tabla IV
MÉTRICAS IDENTIFICADAS PARA LA EVALUACIÓN DE QOE EN IOT/IOE SEGÚN [27]

Capa o nivel	Descripción	Asociación con modelo [10] de Tabla I
Device	Calidad de la generación de datos brutos	QoD
Network	Calidad en la transferencia de datos	QoE (QoS)
Computing	Calidad en el procesado y análisis de los datos (almacenamiento, abstracción, indexación, prestaciones CPU, precisión,…)	QoI, QoE
User interface	Calidad en la interacción aplicación-usuario	QoE

IV. DISCUSIÓN Y CONCLUSIONES

Como se indica en todos los trabajos referenciados, aún supone un reto el estudio y análisis de la QoE a nivel de sistema, de forma generalizada. Tras describir las diferentes visiones holísticas propuestas en la literatura, surgen algunas inquietudes. En primer lugar creemos que existe una clara duplicación de métricas. Será necesario determinar si realmente es necesario incluirlas en diferentes niveles o si su impacto puede absorberse en un único nivel. En segundo lugar, es necesario un estudio detallado de qué métricas han de tenerse en cuenta. Ya sean métricas físicas con una relación directa en la QoE, como se indicaba en [27], o una combinación de éstas donde se podría emplear una metodología similar a los árboles de decisión, aunando las sugerencias realizadas en [23] y [27]. Por último, la tendencia de dar respuesta a la QoE como la unión del nivel de calidad en varios niveles o capas parece acertada. No obstante, la relación (el modelado) entre los distintos niveles no ha sido definido en ninguno de los trabajos considerados. Además, creemos necesario profundizar en el estudio de la inclusión de un componente "negativo" en la formulación final de la QoE, como podría ser el *Quality Cost*.

Consideramos por tanto necesario un nuevo enfoque para la evaluación de prestaciones de los servicios y aplicaciones de la IoE capaz de capturar su idiosincrasia, es decir, las nuevas métricas que definen calidad, nivel de conocimiento, nivel de inteligencia, consumo, etc. En consecuencia, uno de los objetivos del proyecto AIM será proporcionar una guía rigurosa para la evaluación de prestaciones, la planificación y el dimensionado de los servicios y aplicaciones en IoE. Para ello, planteamos dos fases. Una primera fase en la que definir, caracterizar y cuantificar nuevas métricas que en base a una visión holística satisfagan QoE en IoE y una segunda fase de búsqueda de diseño de herramientas de optimización capaz de capturar la relación entre los distintos niveles en los que se categorice la QoE.

AGRADECIMIENTOS

This research was supported by the AEI/FEDER, UE project grant TEC2016-76465-C2-1-R (AIM).

REFERENCIAS

[1] I. Bojanova, G. Hurlburt, J. Voas, "Imaginering an Internet of Anything", Computer, Vol. 47 (6), pp. 72-77, 2014.
[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications", IEEE Communications Surveys & Tutorials, Vol. 17 (4), pp. 2347-2376, 2015.
[3] A. Sheth, "Internet of Things to Smart IoT through Semantic, Cognitive, and Perceptual Computing", IEEE Intelligent Systems, Vol. 31 (2), pp. 108-112, 2016.
[4] J. Lloret, J. Tomas, A. Canovas, L. Parra "An Integrated IoT Architecture for Smart Metering", IEEE Communications Magazine, Vol. 54 (2), pp.50-57, 2016.
[5] S. M. A. Oteafy, H. S. Hassanein, "Resilient IoT Architectures Over Dynamic Sensor Networks With Adaptive Components", IEEE Internet of Things Journal, vol. 4 (2), ,pp. 474-483, 2017.
[6] N. Kaur, S. K. Sood, "An Energy-Efficient Architecture for the Internet of Things (IoT)", IEEE Systems Journal, Vol. PP (99), pp. 1-10, 2015.
[7] M. Raynor, M. Cotteleer, "The more things change: Value creation, value caption, and the Internet of Things", Deloitte Univ. Press, 2015. [Online] http://goo.gl/XJK09n/
[8] A. Floris et al., "Quality of Experience in the Multimedia Internet of Things: definition and practical use-cases", in Proc. ICC15, pp. 1747-1752, 2015.
[9] R. Sanchez-Iborra, M.-D. Cano, "JOKER: A Novel Opportunistic Routing Protocol", IEEE J. of Selected Areas in Communications, 2016. DOI: 10.1109/JSAC.2016.2545439
[10] Q. Wu et al., "Cognitive Internet of Things: A New Paradigm Beyond Connection", IEEE Internet of Things Journal 1(2), 2014.
[11] "ITU-T: New definitions for inclusion in Recommendation ITU-T P.10/G.100", ITU-T Recommendation P.10/G.100 Amend. 2, 2008.
[12] P. Kasnesis, C. Z. Patriakekios, D. Kogias, L. Toumanidis, I.A. Venieris, "Cognitive friendship and goal management for the social IoT", Computers and Electrical Engineering, Vol. 58, pp. 412-428, 2017.
[13] P. Bellavista, "A survey of context data distribution for mobile ubiquitous systems", ACM Computing Surveys (CSUR), vol. 44 (4), pp. 1-49, 2012.
[14] J. Mitola, "Cognitive radio architecture evolution", Proceedings of IEEE, vol. 97 (4), pp. 626-641, 2009.
[15] "The E-model, a computational model for use in transmission planning", ITU-T Rec. G.107, 2005.
[16] "ITU-T: Parametric non-intrusive assessment of audiovisual media streaming quality", ITU-T Recomm. P.1201, 2012.
[17] "ITU-T: Parametric non-intrusive bitstream assessment of video media streaming quality", ITU-T Recomm. P.1202, 2012.
[18] Maria-Dolores Cano, F. Cerdan, S. Almagro, "Statistical Analysis of a Subjective QoE Assessment for VVoIP Applications", ETRI Journal Vol. 32 (6), pp. 2010.
[19] Dong-Hee Shin, "Conceptualizing and measuring quality of experience of the Internet of Things: Exploring how quality is perceived by users", Information & Management, in press, pp. 1-14, 2017.
[20] Zaheeruddin, D.K. Lobiyal, Sunita Prasad, "Ant based Pareto optimal solution for QoS aware energy efficient multicast in wireless networks", Applied Soft computing, Vol. 55, pp. 72-81, 2017.
[21] Qiaoni Han, Bo Yang, Cailian Chen, Xinping Guan, "Energy-aware and QoS-aware load balancing for HetNets powered by renewable energy", Computer Networks, Vol. 94 (15), pp. 250-262, 2016.
[22] Chi Harold Liu, Jun Fan, Joel W. Branch, Kin K. Leung, "Toward QoI and Energy-Efficiency in Internet-of-Things Sensory Environments", IEEE Transactions on Emerging Topics in Computing, Vol. 2 (4), pp. 473-487, 2014.

[23] A. Floris, L. Atzori, "Quality of Experience in the Multimedia Internet of Things: definition and practical use-cases", in Proc. International Conference on Communicaitons (ICC) - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI), pp. 1747-1752, 2015.

[24] M. Aazam, M. St-Hilaire, Chung-Horng Lung, I. Lambadaris "MeFoRE: QoE based resource estimation at Fog to enhance QoS in IoT", in Proc. 23rd International Conference on Telecommunications, pp. 1-5, 2016.

[25] F. F. Reichheld, "The One Number You Need to Grow", Harvard Business Rewiew, 2003.

[26] L. Li, M. Rong, G. Zhang, "An Internet of Things QoE evaluation method based on multiple linear regression analysis", in Proc. 10th International Conference on Computer Science & Education (ICCSE), pp. 925-928, 2015.

[27] Y. Ikeda, S. Kouno, A. Shiozu, K. Noritake "A framework of scalable QoE modeling for application explosion in the Internet of Things", in Proc. IEEE 3rd World fórum on Internet of Things (WF-IoT), pp. 425-429, 2016.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6569

Aplicación Web para comunicación multimedia en tiempo real y en movilidad

Miguel Gil, Elsa Macías, Alvaro Suárez
Departamento de Ingeniería Telemática
Instituto de Ciencias y Tecnologías Cibernéticas

Universidad de Las Palmas de Gran Canaria

Dirección Postal: Edificio de Electrónica y Telecomunicación – Campus universitario de Tafira – 35017 Las Palmas de G.C.

miguelgilbr@gmail.com, alvaro.suarez@ulpgc.es, elsa.macias@ulpgc.es

***Resumen*- En este documento se presenta una aplicación Web para comunicación multimedia en tiempo real y compatible con dispositivos móviles, combinando varias tecnologías que son tendencia actualmente en el mundo del desarrollo: HTML5, WebRTC y Node.js. A partir de las tecnologías anteriores se ha desarrollado una plataforma que desde un navegador compatible permite: realizar videoconferencias en grupo, enviar y recibir archivos, conversaciones en texto privadas entre usuarios, conversaciones de texto en grupo y grabación de las emisiones multimedia de cualquiera de los usuarios de la sala. A nivel visual, se ha diseñado una interfaz simple de usar y completamente responsiva, es decir, compatible con todos los tamaños de pantallas, incluidas las de cualquier teléfono móvil. Se ha hecho comparativas con aplicaciones actuales comerciales y se observa un rendimiento similar o mejor que algunas de ellas. En cambio, el consumo de energía es muy considerable (al igual que en el resto de aplicaciones analizadas).**

***Palabras Clave*- HTML5, WebRTC, Node.js**

I. Introducción

En los últimos años, los servicios de videoconferencia a través de Internet han evolucionado notablemente, pasando de ser una tecnología cara y que requería de buenas conexiones de datos, a ser servicios habituales, con un consumo moderado de datos y al alcance de todo el mundo.

A día de hoy es posible realizar conferencias simultáneas entre varias personas de una manera simple y gratuita con aplicaciones como *Skype* [1] o *Google Hangouts* [2], aplicaciones ampliamente usadas para realizar este tipo de comunicaciones. El servicio de videoconferencia se suele integrar con otros servicios básicos, por ejemplo, la aplicación de mensajería instantánea más usada (*WhatsApp* [3]) o la red social con más usuarios (*Facebook* [4]).

En el marco de la Tele-enseñanza, los sistemas de videoconferencia todavía no se usan masivamente, a pesar de que en determinados escenarios son imprescindibles. Un ejemplo de esto es el aprendizaje de idiomas, en el que una parte consiste en hablar y escuchar, lo que implica mantener una conversación fluida en persona. Esto solo es posible usando sistemas de videoconferencia. En la práctica se suelen usar, no masivamente: a) sistemas de pago (de calidad negociable con el operador de Telecomunicación), como es el caso de *Polycom* [5], b) sistemas de pago de menor calidad como las de *ISL Group* [6], c) herramientas corporativas de elevadísima calidad como *WebEx* [7], entre otras muchas corporativas, y d) herramientas con versiones gratuitas que dan un servicio básico (como, por ejemplo, *TeamViewer* [8]). Sin embargo, las opciones para usar sistemas de videoconferencia en Tele-Enseñanza, están cambiando gracias a que a día de hoy es más sencillo acceder a herramientas y funcionalidades gratuitas, que se programan haciendo uso de la tecnología *Web Real Time Communications* (*WebRTC*) [9] [10] e *Hiper Text Markup Language versión 5* (*HTML5*) [11] [12] [13], debido a que en teoría es extremadamente fácil programarlas.

En este artículo presentamos el desarrollo integral de un servicio de videoconferencia que hace uso de WebRTC y HTML5. En la práctica se demuestra que la programación de este tipo de servicios eficientes, seguros y responsivos no es una tarea tan sencilla como indica la teoría. Este servicio se puede usar en las plataformas móviles más usadas en la actualidad (Android e iOS), además hace uso bidireccional de la comunicación entre usuarios, maneja flujos de comunicación de texto (*chat*), permite grabar las sesiones de video y enviar archivos. Se ha esmerado el diseño e implementación para que el uso del servicio sea lo más simple posible, que maximice la *Quality of Experience* (*QoE*) y aproveche al máximo los mecanismos de *Quality of Service* (*QoS*) disponibles en la red (en especial en *Wireless Fidelity* (*WiFi*), tecnología para la que hemos hecho pruebas de funcionamiento mostrando un comportamiento similar a servicios como *Skype* y *Hangout*.

II. Arquitectura del Software

El servicio propuesto se compone de dos sistemas principales:

A. Servidor

Desarrollado usando la tecnología *Node.js* [14] [15], ya que, es muy adecuada para escribir de manera simple y rápida servicios Web. Se encarga de las siguientes tareas: almacenar y hacer accesible a los usuarios la aplicación Web, guardar los datos de cada uno de los usuarios que se conecte al sistema, controlar en tiempo real la entrada y salida de usuario e informar al resto de usuarios, encargarse de la señalización a la hora de crear canales de comunicación *Peer 2 Peer* (*P2P*), ya sea para las transmisiones multimedia, o para los canales de datos, gestionar el estado de las conexiones entre los clientes, de manera que el sistema asegure las conexiones entre todos los usuarios, toda la lógica del chat de grupo.

B. Cliente Web responsivo

Desarrollado con HTML5, combinado con algunas bibliotecas *JavaScript* y *Cascade Style Sheets versión 3* (*CSS3*) [16]. Es el código que ejecuta el usuario en su navegador y se encarga de realizar las siguientes tareas: punto de entrada del usuario al sistema, gestión de la conexión con el servidor central mediante WebSockets, acceso a los medios de captura multimedia usando WebRTC, conexión entre los usuarios conectados al sistema, acceso al sistema de archivos del usuario para poder enviarlos, control de ventanas del Chat.

A continuación (Fig. 1), se describen los módulos funcionales de la solución de forma independiente al sistema que esté implicado en su funcionamiento. Esto implica que existen módulos que dependen solo del servidor, de la aplicación Web, o en el caso más común, de ambos.

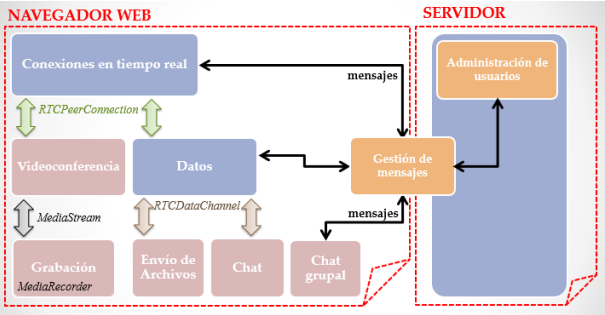


Fig. 1. Módulos funcionales de la aplicación.

El módulo *Conexiones en Tiempo Real*, se encarga de la configuración y gestión de los objetos *RTCPeerConnection*, que son los encargados de gestionar las conexiones P2P entre los usuarios. Una vez que ya tenemos establecido un canal P2P entre dos usuarios, es muy sencillo establecer un canal de comunicaciones extra usando el módulo *Datos*, que se encarga de generar los canales de datos *RTCDataChannel* usando la función *createDataChannel* de los objetos *RTCPeerConnection* generados con el primer módulo.

Ambos módulos requieren de un proceso de sincronización y señalización entre los usuarios implicados en la comunicación, de los que se encarga el módulo *Gestión de mensajes*, para el intercambio de información entre usuarios a través del servidor y el módulo *Administración de usuarios* que se encarga de almacenar y actualizar en tiempo real los datos y el estado de los usuarios del sistema.

Para finalizar, se observan los módulos de aplicación, que se corresponden con funcionalidades concretas del sistema. El módulo *Videoconferencia* se encarga de la transmisión y recepción de las transmisiones audiovisuales de los usuarios, mientras que el módulo *Grabación* graba las transmisiones recibidas, usando el objeto *MediaRecorder*. Los módulos *Envío de Archivos* y *Chat* se apoyan en el módulo *Datos* para funcionar y por último el módulo *Chat grupal* funciona gracias al servidor y a los mismos módulos encargados de la señalización y sincronización y que se comentaron en el párrafo anterior.

III. Rendimiento y Usabilidad

En este apartado presentamos en primer lugar la ventana principal del servicio desarrollado (para que se observe el diseño minimalista presentado) y en segundo lugar unas ideas mínimas sobre su rendimiento.

A. Interfaz gráfica

En la Fig. 2 se muestra la interfaz de usuario de la aplicación que consta de 3 partes fundamentales:

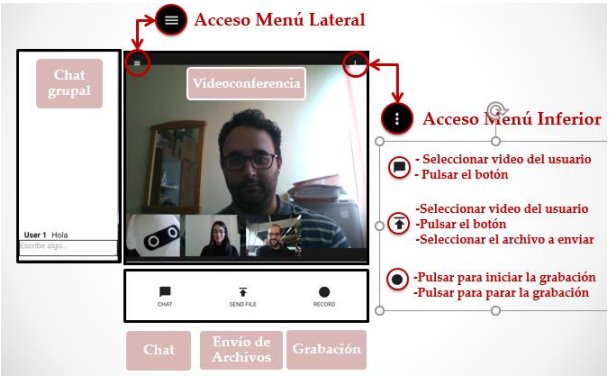


Fig. 2. Interfaz Gráfica de la aplicación

- La ventana principal en la que se observa, en primer plano, la videoconferencia principal y en ventanas más pequeñas sobrepuestas a la principal el resto de videoconferencias. Al pinchar en la videoconferencia de una ventana pequeña, esta pasa a ser la principal.

- El menú lateral, fácilmente desplegable usando el botón superior izquierdo y que se encarga de mostrar la interfaz del chat en grupo.

- El menú inferior, que se encarga del acceso directo al resto de aplicaciones del sistema: el chat privado, el envío de archivos y la grabación. Todas estas funcionalidades afectarán al usuario de la videoconferencia contenida en la ventana principal.

Comentar adicionalmente que la interfaz desarrollada es compatible con todo tipo de tamaños de pantalla (computadores de sobremesa, tabletas y teléfonos móviles), y está optimizada para su correcta visualización en dispositivos móviles.

B. Rendimiento

Para medir el rendimiento del servicio, en primer lugar, se ha analizado el consumo de *Random Access Memory* (*RAM*) del servidor ya que, en teoría, una de las mejores prestaciones del Node.js es su reducido uso de RAM. Para realizar las pruebas, se ha instalado el servidor Node.js en varios sistemas operativos distintos y se ha medido el consumo de RAM.

Los resultados de la prueba se muestran en la Fig. 2 y se han realizado en un computador *Mac* (*Apple*) con el sistema operativo *OS X El Capitan* versión 10.11.6, una configuración de 12 GB de RAM DDR3 a 1600 MHz y un procesador Intel Core i7 de 2,3 GHz. El consumo de memoria RAM es muy bajo, unos 25 MB de RAM.

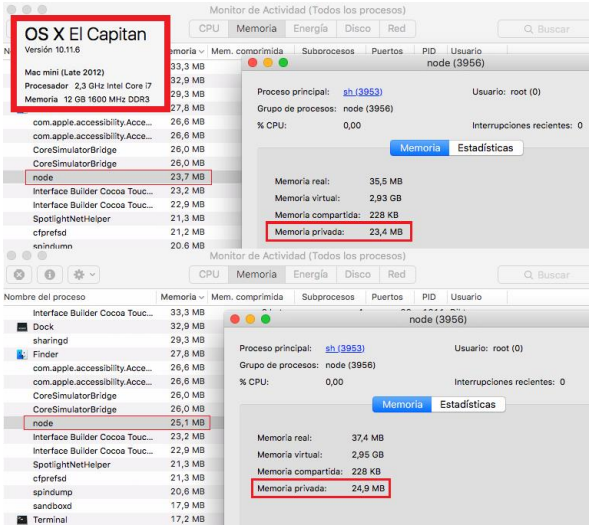


Fig. 3. Consumo de RAM de la aplicación

A continuación, se ha analizado el consumo de datos de una videoconferencia de 1 minuto de duración. Se han analizado los paquetes de la conversación, capturándolos usando el software analizador de redes *Wireshark*. En la Fig. 3 aparece el consumo de datos en bruto de todo el proceso y se distinguen 4 fases.

La primera de ellas se corresponde con el acceso al sistema y la descarga de toda la aplicación Web, cuyo procedimiento desemboca en la pantalla de acceso al sistema. La segunda fase es la entrada al sistema y la conexión con el Servidor. En este momento en el sistema no hay nadie y se está a la espera de que entre alguien con el que conectar. El tráfico en este punto sigue siendo *Transmission Control Protocol* (*TCP*).

La tercera y cuarta fase ocurren casi simultáneamente: cuando un nuevo usuario entra al sistema, el sistema lo comunica. Posteriormente comienza la negociación entre los usuarios, como paso previo a iniciar la conexión multimedia efectiva. Todo este procedimiento ocurre bajo TCP. La cuarta fase es el intercambio de datos en tiempo real y se efectúa en su mayoría con tráfico *User Datagram Protocol* (*UDP*). Una vez estabilizada la conexión, el ancho de banda usado oscila entre los l70 y los 220 paquetes por segundo y un ancho de banda equivalente en torno a los 100 KBps de media.

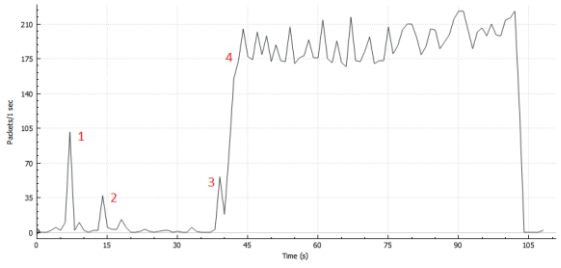


Fig. 4. Consumo de datos de la aplicación

Por último, se ha comparado el consumo de datos con los de Skype y Google Hangout. Los resultados aparecen en la tabla I, en la que se observa que la aplicación tiene un consumo moderado y sobretodo, es muy rápida a la hora de realizar una conexión efectiva.

Tabla I
COMPARATIVA

	Datos	Velocidad Tx	Tiempo conexión
Aplicación	5,19 MB	83 KB/s	**2,62 s**
Skype	**2,79 MB**	**42 KB/s**	6.5 s
Hangouts	14 MB	197 KB/s	11 s

Se hicieron otros tipos de pruebas (recogidas en [17]), en especial se hicieron pruebas de análisis de la QoE y de respuesta de la QoS en WiFi. Después de esos estudios llegamos a la conclusión que se debe dejar al programador (la herramienta que define el servicio es de código abierta) el poder variar el valor de una variable que limita el número máximo de usuarios que pueden hacer videoconferencia simultáneamente. Nosotros encontramos que el mejor número de usuarios para la mayoría de las instalaciones WiFi es 4.

IV. CONCLUSIONES

En este artículo hemos presentado un servicio de videoconferencia muy simple de usar y adecuado especialmente para tele-enseñanza, por cuanto permite flujos de video, chat archivos y grabación de sesiones. A partir de los resultados experimentales, podemos asegurar que la aplicación desarrollada tiene un rendimiento destacable y es fácil de usar.

Como trabajo futuro debemos desarrollar mecanismos de consumo inteligente de sesiones de videoconferencia adaptando inteligentemente el número de usuarios según la QoS de la Red y por otro lado optimizar el consumo de energía en los teléfonos móviles. También comparar con otras herramientas como Facebook Messenger.

AGRADECIMIENTOS

This work has been funded by the Spanish Ministry of Economy and Competitiveness/FEDER under project TEC2015-67387- C4-4-R.

REFERENCIAS

[1] Página oficial de Skype: https://www.skype.com/es/

[2] Página oficial de Google Hangouts: https://hangouts.google.com/?hl=es

[3] Página oficial de WhatsApp: https://www.whatsapp.com/?l=es

[4] Página oficial de Facebook: https://www.facebook.com/

[5] Página oficial de Polycom: http://www.polycom.es/

[6] Página oficial de ISLOnline: http://www.islonline.com/?hl=es

[7] Página oficial de WebEx: https://www.webex.es/

[8] Página oficial de TeamViewer: https://www.teamviewer.com/es/

[9] Salvatore Loreto, Simon Pietro Romano, Real-Time Communication with WebRTC. Peer-To-Peer in the browser. Editorial O'Reilly. Mayo 2014. ISBN 978-1-78216-630-6

[10] Rob Manson, Getting Started with WebRTC. Explore WebRTC for real-time peer-to-peer communication. Editorial Packt Publishing. Septiembre 2013. ISBN 978-1-449-37187-6

[11] Peter Lubbers, Brian Albers, Frank Salim, Pro HTML5 Programming. Powerful APIs for Richer Internet Application Development. Use HTML5 to create cutting-edge Web applications. Editorial Apress. 2010. ISBN 978-1-4302-2791-5

[12] Dale Cruse, Lee Jordan, HTML5 Multimedia Development Cookbook. Recipes for practical, real-world HTML5 multimedia-driven development. Editorial Packt Publishing. Mayo 2011. ISBN 978-1-849691-04-8

[13] Silvia Pfeiffer, The Definitive Guide to HTML5 Video. Everything you need to know about the new HTML5 video element. Editorial Apress. 2010. ISBN 978-1-4302-3091-2

[14] Tom Hughes Croucher, Mike Wilson, Node. Up and Running. Editorial O'Reilly. Mayo 2012. ISBN 978-1-449-39858-3

[15] Pedro Teixeira, Professional Node.js. Building Javascript-Based Scalable Software. Editorial John Wiley & Sons, Inc. 2013. ISBN 978-1-118-18546-9

[16] Página oficial de la librería NativeDroid2: http://nativedroid.godesign.ch/material/

[17] Miguel Gil, Aplicación Web para comunicación multimedia en tiempo real y en movilidad, Proyecto final de Carrera, Escuela de Electrónica y Telecomunicación, Universidad de Las Palmas de Gran Canaria, Directores: Elsa Macías y Alvaro Suárez, 2017.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6587

Quality of Service (QoS) oriented management system in 5G cloud enabled RAN

Rubén Solozabal, José Oscar Fajardo, Bego Blanco, Fidel Liberal
University of the Basque Country (UPV/EHU)

{ruben.solozabal, joseoscar.fajardo, begona.blanco, fidel.liberal}@ehu.eus

***Abstract*—This paper analyze different techniques to implement Quality of service (QoS) on multi-tenant 5G networks. Describes the architecture of the next generation mobile network based on cloud-enabled small cell deployments and also proposes an hybrid-cloud solution coexisting with centralized cloud RAN (C-RAN), in order to achieve a gradual implementation of the technology. In this context, the work here presented deals with the challenges of preserving the quality of experience in a multi-tenant cloud enable RAN bearing in mind the Key Performance Indicator (KPI) agreed in the Service Level Agreement (SLA). To achieve this goal, QoS should be managed at different levels of the architecture. Feedback should be given between learning modules in order to analyze the results and infer enhanced decision rules which may conclude in an architecture replacement.**

***Keywords*—Network Function Virtualization, Network Service instantiation, cloud-enabled small cells, QoS/QoE C-RAN, Mobile Edge Comuting**

I. INTRODUCTION

The flexible Radio Access Network (RAN) proposed in 5G will leverage Software Defined Network (SDN), Network Function Virtualization (NFV) and Mobile Edge Computing (MEC) principles for a simplified network deployment and management, enhancing CAPital EXpenditures (CAPEX) / Operating expense (OPEX) efficiency. Intelligent 5G centralized RAN systems will concentrate processing resources together in shared data centers not only in order to reduce deployment costs, but also to provide low latency connections between different RAN processing units. Traditional deployments of specialized devices with 'hard-wired' functionalities will be replaced by general-purpose reconfigurable computing assets. Making use of cloud computing, SDN and NFV, the centralized RAN will become a Cloud RAN (C-RAN) [1]. The softwarization of the network functions will enable the automation of network service provisioning and management, thus, the adaptation to growing and heterogeneous market requirements at a lower cost.

To facilitate the adaptation of the current architecture to the proposed, SESAME project[2] analyzes the development of multi-tenant cloud enabled RAN (C-RAN) through the evolution of the architecture of traditional commercial Small Cells (SC) to Cloud Enabled Small Cells (CESC). A CESC is a multi-operator SC that integrates a virtualized execution platform to support the executions of novel applications in the network edge using NFV and SDN technology. The proposed solution extends the Small Cell as a Service (SCaaS) model, which provisions of shared radio access capacity to mobile network operators in localized areas. Efficient management of resources, rapid introduction of newer network functions and services, easy of upgrade and maintenance and CAPEX/OPEX reduction are only few examples of various benefits that the proposed solution provides.

Despite the potential technical benefits, viability of the solution strongly depends on several factors such as the guarantee of the Service Level Agreements (SLAs). A SLA which captures the particular Key Performance Indicators (KPI) of a delivery –scope, quality, and responsibilities– can play a significant role towards business success. In this paper several techniques are proposed at different levels of the next generation mobile architecture in order to improve the overall quality of experience.

This paper is organized in six sections. First, Section II deals with cloud architectures evolution. Then, Section III analyzes the service provisioning models over the proposed architecture and Section IV discusses the definition of network services. Next, Section V proposes hybrid cloud approach for the evolution to 5G C-RAN and Section VI introduces the implementation of a QoS into the service life-cycle management. Finally, Section VII summarizes the main conclusions.

II. MULTI-TENANT CLOUD ENABLED RAN

Traditionally, actual installation of physical infrastructure is needed to provide coverage in one Point of Presence (PoP). Such an ownership increases operator's CAPEX and significantly hampers business agility, particularly when considering the high degree of cell densification. In a multi-tenant scenario, an infrastructure provider can grant

access to third parties such as network operators, service providers or Over-The-Top (OTT) players. Sharing the physical infrastructure increases service dynamicity and reduces the overall cost and energy consumption compared to the case where parallel systems are installed in one PoP to support connectivity for different parties.

Beyond the pure centralization of an eNB functions, one of the emerging technologies to cope with more personalized and user-centric service provisioning is the novel MEC. This may be exploited to deploy proximity-enabled services with close-to-zero latency characteristics. Regardless of the adopted architecture for C-RAN, MEC-driven service instances must be deployed over the cloud resources available at the RAN side.

In this centralized solution, the upper RAN functions are located in powerful data centres that are ideally connected to the RRHs through high-speed and low-latency fronthauls. Yet, high fronthaul delays may degrade the performance of certain novel edge services that require close-to-zero latencies as prescribed by 5G objectives. Alternatively, nowadays CESCs architecture may become better suited for deploying mobile edge services. In that case, some processing and storage resources are placed close to the RRH, and thus, the fronthaul delay is significantly reduced. Deploying huge data centres implies a series of requirements in terms of space, energy, etc. Hence, this second option envisages the deployment of a series of HW resources with limited capacity and requirements and in a distributed configuration.

Figure 1 shows an architecture to consolidate multi-tenancy in the mobile communications infrastructures based on a substantial evolution of the SC towards cloud-enabled devices, as proposed in the SESAME project.

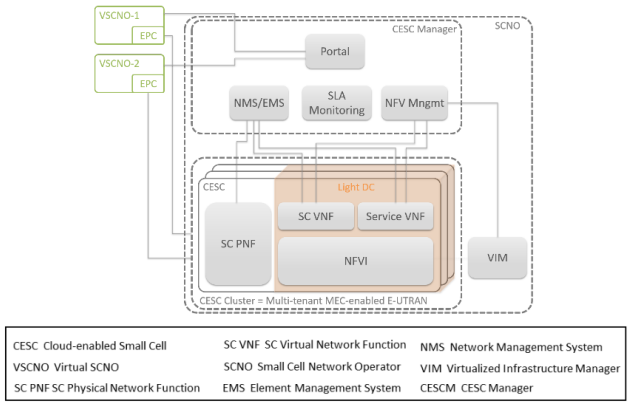


Fig. 1. Multi-Tenant cloud enabled RAN using SCs

The key element of this architecture is the CESC, owned by a Small Cell Network Operator (SCNO), which consists of a micro server integrated with the small cell to support both radio connectivity and edge services. It foresees the split of the small cell into Physical Network Functions (PNF) and Virtual Network Functions (VNF)[3], enabling a multi-tenant environment in support of a Multi-Operator Core Network (MOCN)[4]. This will allow Virtual Small Cell Network Operators (VSCNO) not only to support connectivity but also to provide added value mobile edge services in a PoP.

Resources on a single micro server (i.e. RAM, CPU, storage, HWA) might not be enough to support the mobile edge computing services of all tenants. CESC clustering enables the creation of a micro scale virtualised execution infrastructure in the form of a distributed data centre, denominated Light Data Centre (Light DC), enhancing the virtualisation capabilities and processing power at the edge. The hardware architecture of the Light DC envisages that each micro server will be able to communicate with all others via a dedicated LAN/WLAN guaranteeing the latency and bandwidth requirements needed for sharing resources. It provides also the backhaul connections to the operators Evolved Packet Core (EPC).

In the context of multi-tenant cloud enabled RAN, a Network Service (NS) is understood as a chain of PNFs and VNFs that jointly supports data transmission between a User Equipment (UE) of an operator and the operator's EPC, with the possibility to involve one or several service VNFs in the data path. It clearly highlights that, beyond the conventional orchestration and management of the cloud resources in a virtualised environment, the proposed solution entails a series of specific challenges such as the dynamic composition of the Light DC resources based on the status of CESC cluster(s), coordination of specific type of resources (radio-related resources, service-related HWA, etc.) and isolation of dedicated network slices to each tenant.

All management tasks, e.g. resource allocation and service lifecycle management over the distributed infrastructure, are carried out by a centralized unit called CESC Manager(CESCM). A single instance of CESCM is able to operate over several CESC clusters, each constituting a Light DC, through the use of a dedicated Virtual Infrastructure Manager (VIM) per cluster. CESCM is the main management component in the architecture, covering the orchestration, management and configuration of NSs. The CESCM has a high-level knowledge of the virtual and physical resources available on the C-RAN environment, including the radio access functionalities. CESCM is composed of the following modules:

- The NFVO is the entity in charge of NS lifecycle management (creation, termination, monitoring, scaling etc.) via coordination between ETSI MANO elements, such as VIM, EMS and VNFM. NSs are defined in the form of NS descriptors (NSD), which contain VNF descriptors (VNFD) – defining required IT resources needs to be dedicated for a VNF as well as its specific functionality – and connectivity between VNFs.
- The VNFM is the entity in charge of the lifecycle management of the VNFs, from deployment to termination, keeping track of their status to adjust their configuration if needed.
- The EMS is the entity in charge of the key functionalities as fault, configuration, accounting, performance and security (FCAPS). It manages the traffic between

the different network elements, coordinating configuration of multiple devices. The EMS associated to radio functions also includes autonomous self-x functionalities to reconfigure the mobile network.

- The SLA component enhances service reliability providing monitoring mechanisms to evaluate the performance of NSs in the radio and cloud environments. It communicates with the NFVO, notifying faults in the system for it to perform the appropriate actions that assure the QoS guarantees of each service in a multi-tenant environment.

In order to communicate CESCM and CESC cluster, the VIM as described in the ETSI[5], is the responsible of managing the virtualized infrastructure, that includes the catalog of the allocated resources, forwarding graphs and chaining rules among VNFs and repository of resources, to provide optimized features. VIM is the software entity that monitors and manages the Network Functions Virtualization Infrastructure (NFVI) (i.e., Light DC) and performs the lifecycle management of the virtual units that will host the VNFs. This centralized administration of virtual resources across multiple localized infrastructure, so that instances can be administrated in a coordinated way, provides the flexibility and scalability needed to optimize and maximize the use of such resources. The VIM is enhanced with SDN component for the networking aspect. The controller takes into account the physically distributed NFVI and the stringent requirements in RAN performance metrics. Moreover, it uses SDN for propagating the VNF chaining requests to the NFVI in order to properly manage the networking resources within the Light DC.

III. SERVICE PROVISIONING MODELS

From the business perspective, three major role players are identified. Function provider (FP) is the VNF developer which sells/develops VNFs. Service Provider (SP), is the one who composes NS –i.e. chain of VNFs, PNFs– with the available VNFs and offers them to the customer. Customer is the one who purchases NSs. In multi-tenant cloud enabled RAN, there are two main possible ways to form a joint radio-cloud model, as illustrated in Fig. 2.

- Mobile Edge Computing as a Service (MECaaS): This model has been inspired mainly from the MNO-MVNO business relationship. Briefly, in this model, MVNO relies completely on the infrastructure and other services provided by the MNO. VSCNO asks for high level KPIs on the SLA. Here, VSCNO only has an overall vision of the system and SCNO has to provide enough support, i.e. both in terms of hardware and number/composition VNF chains (i.e. NS), to meet the agreed KPIs. Performance reports are provided to VSCNO on time intervals (even real time). In simple words, with this model, VSCNO does not chain VNFs to form a mobile edge service, and a high level KPI view is enough for it to request a service without going to details.
- CESC Infrastructure as a Service (CESCIaaS): In this model, VSCNO on SLA asks for connectivity in a certain coverage area according to the elements and for aggregated cloud resources on the Light DC, e.g. a certain amount of GB of storage, of RAM, etc. This model corresponds with the famous Infrastructure as a Service (IaaS) paradigm, which is one of the three fundamental service models of cloud computing[6]. With this model in place VSCNO can compose VNF chains on demand. As a consequence, any VNF instantiation (depending on the used hardware resources) consumes a portion of available VSCNO's aggregated resources. Therefore, the deployment of VNF chains is conditioned to the amount of requested resources by the VSCNO.

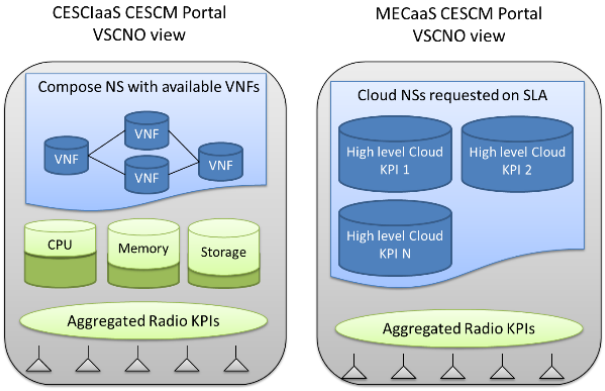


Fig. 2. VSCNO's CSCM portal view on multi-tenant cloud enabled RAN

Figure 2 shows the dashboard view of VSCNO based on the two mentioned joint cloud-radio service provisioning models.

IV. NS DEFINITION

Fig. 3 shows a conceptual view of the ecosystem, which hosts three different VSCNOs over the shared NFVI, i.e. Light DC. The network service is described as a collection of VNFs (including radio related and service related instances) required to deploy a complete 5G mobile service for the end users of the VSCNO. To form the multi-tenant scenario depicted in Fig. 3, NFVO needs to process the functional chaining of VNF requested by each VSCNO and to trigger its instantiation to the VIM that manages the NFVI. Therefore, a NS can be characterized through a series of radio-level and service-level KPIs that are captured in the SLA between the SCNO owning the NFVI and the interested VSCNO. A NS is defined through its associated Network Service Descriptor (NSD).

The main building blocks of a NSD are the VSCNO Network Connectivity Topology (NCT) and the VNF Forwarding Graph (VNF-FG) descriptors. The NCT determines the complete list of VNFs and their Connection Points (CPs), as well as the possible interconnections through a series of Virtual Links (VLs). In this sense, the VSCNO NCT can be seen as the virtual network slice assigned to that VSCNO in the CESC Cluster. VIM will map the logical request to the actual hardware by

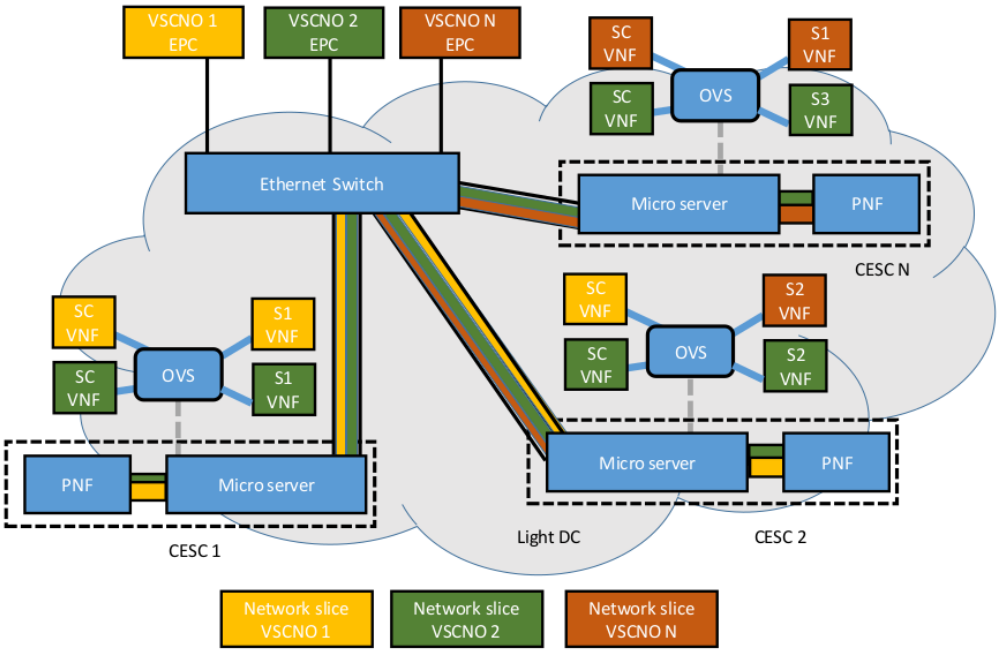


Fig. 3. Edge network services

instantiation of VNFs, which may run in different micro servers. The distributed nature of the edge cloud introduces a novel challenge on the hardware resource allocation. At the end of the placement process, the created NS for the VSCNO can be observed as a separate underlying virtual LAN instantiated and ready to use. However, the actual data flows between VNFs need to be enforced (VNFs can be seen as the origin and destination of data flows). The SDN controller (integrated into the VIM) implements the forwarding rules necessary to move packets according to the VNF-FG.

V. EVOLVING TO 5G C-RAN - HYBRID CLOUD APPROACH

Hybrid approaches have been proposed to deal with the transitions from 4G specific hardware based architectures to software based 5G platforms. Both centralized and distributed clouds will coexist during this evolution to a completely softwarized central C-RAN. Multiple clouds can work together under orders form the same orchestrator to develop a hybrid cloud. In this model, VNFs can be spread between centralized and CESC distributed clouds. Depending on the functional split, central cloud can take part in a different layer of the softwarized upper layer protocols. Depending on the case, it can process just the VNFs form the upper protocol layers or the whole softwarized stack as initially proposed[7]. A Hybrid NFV manager is proposed to orchestrate both clouds in an unified manner.

As lower layers on the protocol stack are virtualised, centralized clusters will take on major relevance as the RAN solution.

VI. QOS OVER MULTI-TENANT C-RAN

Ensuring the QoS per tenant base, assuring the SLA, is another important aspect in the service lifecycle management. QoS assurance demands establishing a feedback loop consist in three main steps:

- Monitoring: a phase in which performance metrics are collected from the radio/cloud/software elements (e.g. SC physical network function, VMs, etc.) and handed over to the next step (decision-making). Depending on the NS deployment and nature the metrics to be collected may vary each time.
- Decision-making: a phase in which performance metrics collected in the previous step are processed. Depending on the situation and available resources, a decision will be taken to ensure the level of QoS (with the help of a dedicated algorithm). Besides available resources, in a multi-tenant scenario, the decision-making process needs to take into account the status of other tenants.
- Reaction: upon making a decision, the management/orchestration system needs to coordinate the interaction with the other lower level modules such as Element Management System (EMS), VNFM and VIM to react appropriately.

Implementing this QoS loop means adding the radio dimension to the standard cloud orchestration system defined by the European Telecommunications Standards Institute (ETSI), and/or to shift the traditional radio network management mentality towards a cloud-oriented mind-set. For instance, leveraging on the radio traffic profile over a certain period (e.g. a day, a week, etc.), it would be possible determining when and where cloud services need to be scaled up/down in a Point of Presence (PoP).

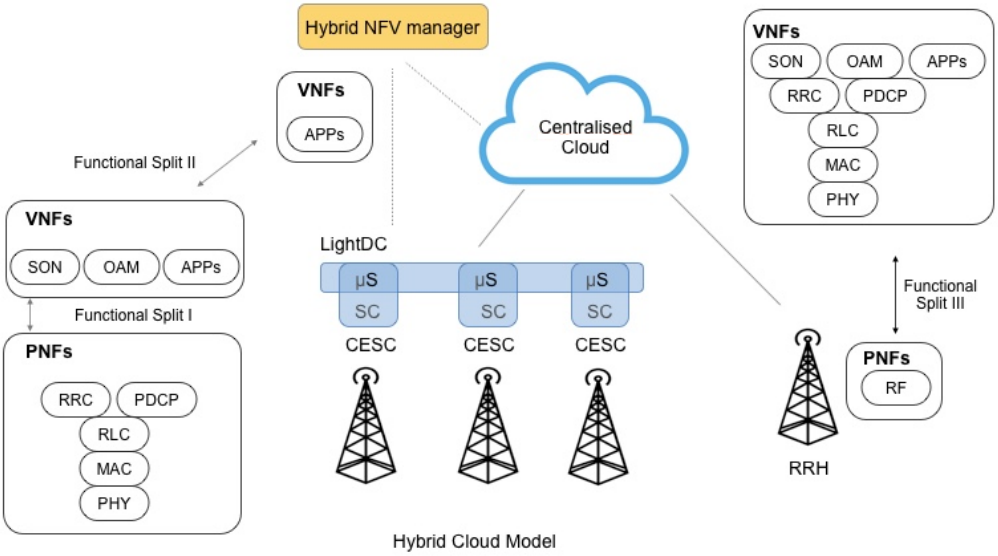


Fig. 4. Hybrid C-RAN

Decision-maker could be placed in several parts along the control plane (VIM, VFNs and NFVO) and the NMS in the management plane[8]. When a change on the architecture is process, a feedback should be given to the learning module in order to analyze the results and infer enhanced decision rules.

Starting at the lower level on the proposed architecture, SDN controller should provide QoS. Networking between VNFs in a chain is an important task in the cloud architecture. Intercepting traffic and forwarding to the correct NS is a challenge resolved by the SDN controller. By taking advantages of standard protocols such as OpenFlow, SDN controller can prioritize traffic according to QoS levels.

At a higher lever, mapping mechanism deals with the allocation of the softwarized components of a NS into the resources of the CESC cluster. In other words, where instantiate the VNF chain that composes a NS. Placement algorithm checks the available virtual resources in the NFVI catalog and the instantiation requirements. Bearing that in mind allocates the resources in the optimal location (see Fig. 5).

Placement algorithm resides inside the VIM. The objective is to place the VNFs chain in order to minimize the end-to-end delay. This placement algorithm could dynamically be recomputed in order to find actively, and not only when the NS is instantiated, the optimal placement in order to achieve the best QoE. Reordering VNFs while executing is possible thanks to live-migrations techniques inside the cloud enable SCs. To achieve a fast migration some requirements have to taken when designing the cloud architecture[9].

The algorithm execution time is critical. Some algorithms deploy new instances over the cloud resources available at the RAN side. In other words, placement of service VNFs that conform the aforementioned edge services are allocated in the remaining infrastructure. In order to not recompute the whole optimal location of all VNFs. There is a commitment between the handover introduced in the system to live-migrate some services and the benefits of QoE obtained.

In the CESCM, VNFM is in charge of the lifecycle management of the VNFs. VMFM is able to apply policies for NS-level rescaling and reconfiguration to achieve high resource utilization. General purpose clouds introduce automatic rescaling algorithms. Generally a minimum and maximum number of intestacies is given to the VMFM in order to select the optimum QoE. In some cases, VNF must be rescaled so it generates a new template indicating the VIM the new architecture. Letting him to compute the best placement for the needed instances.

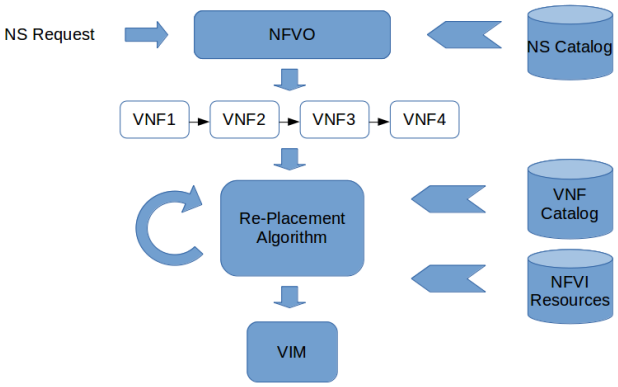


Fig. 5. Replacement algorithm

NMS-EMS are responsible for the management of the network slices that server the NSs. In the proposed solution[10] (Fig. 6) the Metric Aggregator (MA), is responsible for combining and filtering the collected monitored parameters and associates them with the running services over the platform. MA continuously processes

the collected monitoring values for the QoS or SLA evaluation.

The Decision Support System(DSS), as shown in Fig. 6, main responsibility is to detect the level of severity on the QoS evaluation process done in MA and decide whether a reactive or a proactive action is needed. Basically, such a decision will be made based on the high level SLA agreements made with VNOs.

According with the decision some downstream reconfiguration must be done: i- NFVO should reconfigure the flow of data in a NS (i.e. changing the SDN rules), ii- VIM could migrate the NS within the PoP, from one PoP to another, iii- and VNFM scale up/down the whole NS (i.e. instantiation of a parallel service or terminating a running one).

NMS manages the interaction between NSs instantiated by different VSCNOs. And it must be done in a prioritize manner. Different pricing schemes could be applied to prioritize clients according to the QoS offered.

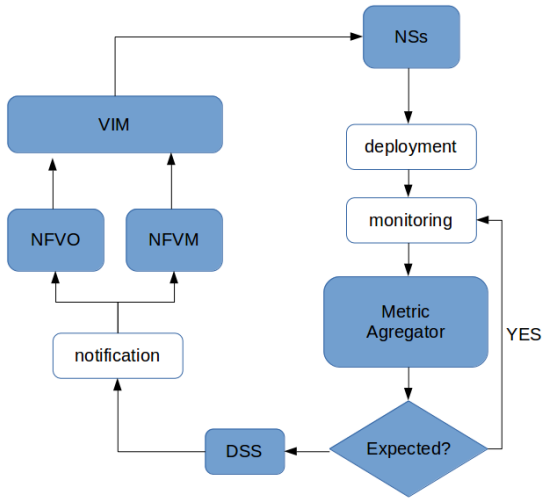


Fig. 6. QoS feedback loop

VII. CONCLUSIONS

This paper has analyzed the evolution of mobile networks to cloud architectures. Proposes an hybrid-cloud coexistence until centralized cloud solutions were extensively applied. It describes the proposed provisioning models and analyzes the components in the proposed cloud RAN architecture in which QoS mechanisms should be applied in order to achieve the KPI agreed with the VSCNOs.

As a result, a vertical management system interaction is needed to reciprocally manage QoS of the overall system in a coordinated manner. QoS mechanism implemented in each level should has its own actuation point in order to improve the QoS locally and therefore improve the global user expedience. In addition an upstream information flow of the decisions taken locally is needed to allow higher levels to achieve global optimization. And also a downstream flow is needed to fix the degrees of freedom at each level the optimization could be made.

VIII. ACKNOWLEDGEMENTS

The research leading to these results has been supported by the EU funded H2020 5G-PPP projects SESAME (Grant Agreement n 671596) and ESSENCE project (Grant Agreement no 761592) and by the Spanish Ministerio de Economia y Competitividad (MINECO) under grant TEC2016-80090-C2-2-R (5RANVIR).

REFERENCIAS

[1] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (c-ran): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, Jan 2015.

[2] B. Blanco, J. O. Fajardo, I. Giannoulakis, E. Kafetzakis, S. Peng, J. Pérez-Romero, I. Trajkovska, P. S. Khodashenas, L. Goratti, M. Paolino, E. Sfakianakis, F. Liberal, and G. Xilouris, "Technology pillars in the architecture of future 5g mobile networks: Nfv, mec and sdn," *Computer Standards & Interfaces*, vol. 54, pp. 216 – 228, 2017, sI: Standardization SDN&NFV. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0920548916302446

[3] 3GPP, "Network Sharing: Architecture and functional description," 3rd Generation Partnership Project (3GPP), TS 23.251, 12 2013.

[4] "Virtualization in Small Cell Networks," *Small Cell Forum*, 2015.

[5] ETSI, "Network Functions Virtualization(NFV): Management and Orchestration," *Small Cell Forum*, 2014.

[6] S.Sharma, "Evolution of as-a-Service Era in Cloud. A review on as-a-Service framework." *Center for Survey Statistics and Methodology, Iowa State University*, USA 2015.

[7] J. O. Fajardo, F. Liberal, I. Giannoulakis, E. Kafetzakis, V. Pii, I. Trajkovska, T. M. Bohnert, L. Goratti, R. Riggio, J. G. Lloreda, P. S. Khodashenas, M. Paolino, P. Bliznakov, J. Perez-Romero, C. Meani, I. Chochliouros, and M. Belesioti, "Introducing mobile edge computing capabilities through distributed 5g cloud enabled small cells," *Mobile Networks and Applications*, vol. 21, no. 4, pp. 564–574, Aug 2016. [Online]. Available: https://doi.org/10.1007/s11036-016-0752-2

[8] B. Blanco, J. O. Fajardo, and F. Liberal, *Design of Cognitive Cycles in 5G Networks*. Cham: Springer International Publishing, 2016, pp. 697–708. [Online]. Available: https://doi.org/10.1007/978-3-319-44944-9_62

[9] "OpenStack Live-migration." [Online]. Available: https://docs.openstack.org/admin-guide/compute-configuring-migrations.html

[10] P. Sayyad, B. Blanco, I. Taobada, M.-A. Kourtis, G. Xilouris, I. Giannoulakis, E. Jimeno, I. Trajkovska, J. O. Fajardo, E. Kafetzakis, J. Garcia, F. Liberal, A. Whitehead, and M. Wilson, "Service management and orchestrarion over multi-tenant cloud-enabled ran," March 2017.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6504

A Study on the Energetic Viability of Single Board Computers for Cloud Computing Scenarios

Pedro Verdugo, Joaquín Salvachúa, Gabriel Huecas
Grupo de Internet de Nueva Generación,
Departamento de Ingeniería Telemática,
Escuela Técnica Superior de Ingenieros de Telecomunicación,
Universidad Politécnica de Madrid
Avenida Complutense, 30. 28040 Madrid.
pmverdugo@dit.upm.es, joaquin.salvachua@upm.es, gabriel.huecas@upm.es

***Abstract*—The following document explores the viability of the usage of consumer-grade, ARM-based single board computers as a power saving alternative to the traditional monolithic x64-full-server based approach. By taking advantage of several capabilities provided by such devices, such as low cost, low power consumption and low on-time, the authors finally propose a scalable, energy-efficient, ARM-based cloud infrastructure. To that end, we start analyzing the current offerings in terms of capabilities, net cost, processing power and power consumption, comparing them with the relevant server-oriented offerings. We subsequently explore the adequacy of several metrics to model on-budget raw data processing, considering full-system wattage under nominal usage conditions. The low initial investment and long-term affordability of this approach results in quite a relevant case of application to Edge Cloud computing scenarios.**

***Palabras Clave*—cloud computing; energy efficiency; green datacenter; microprocessor**

I. INTRODUCTION

As current Big Data loads continue to increase, datacenter processing power is constantly required to scale exponentially. Therefore, it is of the utmost importance for the current infrastructures to ensure that such increase in volume is performed in an energy-efficient way. As a result, major industry players are turning to customized hardware options, often different from the traditional monolithic x64-full-server based approach.

In the present study we will investigate the energy and cost viability of ARM architecture processors for the deployment of cloud based Big Data analysis datacenters as opposed to more traditional systems.

For starters, we will classify the current hardware offerings in terms of selected variables applicable to the investigation field in the subject matter, such as *power consumption*, *price* and *processing power*.

A higher order classification will be made possible by subdividing the detected offerings in terms of actual datacenter volume, providing three levels of performance maturity.

These first order variables will later be put into context by means of an analysis of the relevant metrics to use, where we will explore the validity of the data provided by current manufacturer documentation, extending and adapting them to our constrained field of study.

A case will be made for the selection of the main metrics employed along this case of study, which will namely consist of energy-related **Data per Joule** , time-related **Data per Second** and cost-related **Total Cost of Ownership**.

Once clarified and refined, these detected useful metrics will be used to obtain real world values with which to compare the previously selected offerings. These results will be presented in a graphic form and interpreted in terms of significance.

Finally, we will summarize the value of the original contribution as presented, as well as point out future open avenues of investigation.

II. LOW COST DEVICES VS. TRADITIONAL SYSTEMS

The last few years have seen a great deal of effort being poured into the development of embedded processors, mainly driven by the need for low power-high performance systems in the cellphone market. As a side effect, nowadays ARM (Advanced RISC Machine) architecture based devices are ubiquitous as access devices, and also as the core of embedded systems in all kind of sensor networks and home appliances.

Table I
PROCESSOR CHARACTERISTICS

Grade	*User*		*MicroServer*		*Datacenter*	
Processor	*Amlogic S805*$^+$	*Xeon E5404*$^+$	*Samsung Exynos5422*$^+$	*Core i7-4790K*$^+$	*Cavium ThunderX*	*Xeon E74890*
ISA	ARM Cortex-A5	x86-64	ARM Cortex-A15	x86-64	ARMv8-x64	x86-64
Number of cores	4	8	8	4	48	15
Frequency (GHz)	1.5	2.5	2.00	4.00	2.5	2.80
Dhrystone GIPS	1.57	12.10	1.78	33.435	12.53	270.73
DGIPS (Total)	9.42	96.85	14.24	133.74	601.65	4061
TDP (W)	2.3	80	4.25	88	80	155
Idle Power (W)	0.73	30	1.75	42	24	67.5
Price (USD)	37	382.84	74	699	750	5649.95

There is a case to be made, therefore, for the use of ARM machines as datacenter processing nodes[1].

A. SPECIFIC HARDWARE COMPARISON

To set the testbed for the following sections, in table I we will break down the specifics of the two base processors families used for comparison, indicating their Instruction Set Architecture (from now on, **ISA**), processing power and price[1].

Single core Dhrystone MIPS have been considered as a de-facto standard for processing power calculations, avoiding the shortcomings of regular manufacturer-provided MIPS data for different system architectures. For the studied architectures, a common term of reference will be billion of instructions per second, or **GIPS**.

As illustrated with the data presented in table I, in its current state the ARM architecture offers some promising characteristics that we can relate to their most common x86 counterpart for a set number of cores:

- **Performance**: In a first approach, x86 processor performance is extremely superior to that of an ARM core.
- **Price**: The cost of a server or computer grade x86 processor and board is quite superior to an ARM system.
- **Power Consumption**: As previously mentioned, the ARM power usage is unequivocally inferior to its x86 counterpart.

III. RELEVANT METRICS

A. CLASSIC POWER METRICS

From 1982[2], classic relevant metrics for server workloads have been based on the **Thermal Design Power** (TDP), or thermal design point, defined as the maximum amount of heat generated during typical computer operation [3]. This clearly insufficient concept[2] as a measure of computer processing power has been overly relegated to a back plane, as the main cpu manufacturers tend to introduce new mainly subjective and inexact measurement to try and solve these constraints.

On the one side, in 2009 AMD proposed the **Average CPU Power** (ACP) [4], with scarce application results.

[1]All prices valid for the current date, given in USD and retrieved from *Amazon* or the manufacturer site as applicable.
indicates measured values.

In the same vein, Intel's own recently introduced[5] **Scenario Design Power** (SDP) is defined as an operating mode of certain mobile processors, revamping the TPD concept to set another metric with a lower thermal point, without any practical application or formal definition whatsoever.

B. ENERGY METRICS

As stated by Hennessy[2], processor performance metrics must necessarily be tied to energy, and not power measurements. This ensures the ability to compare in the same grounds different processor architectures as well as different families from the same one. The introduced energy metrics that will be used along the rest of this publication, are defined as follows:

- **Data Processed Per Second** (DPS): Understood as the amount of CPU processed data in a given time (in our case, one second).
- **Data Processed Per Joule** (DPJ): Defined as the amount of data the CPU is able to process with a given energy budget of 1 Joule.
- **Energy-Delay Product** (EDP): As introduced by Horowitz [6] in the transistor performance environment and subsequently expanded by Laros III[7], defines in our specific environment the time taken by the processor to output a given amount of data for a set energy budget. Because it relates to the output processing latency due to i/o artifacts, we will not consider it in our processor-only context.

C. COST CALCULATIONS

The study of cloud computing setup costs has been of wide interest throughout the literature, and there seems to be an agreement as the usage of the **Total Cost of Ownership** (TCO) as a de-facto standard. For the purposes of this paper, we will base our calculations on the TOC formulae presented in [8] as referred to **Infrastructure as a Service** (IaaS) setups, with further operational cost refinements as detailed by [9].

n	Number of nodes in cluster
t	Runtime (Hours)
C_{pi}	Provisioning Cost per node ($)
C_{ei}	Total Electricity Cost per node ($)
C_h	Electricity Cost per hour (KWh)
P_f	Full Power Usage per node (W)
P_i	Idle Power Usage per node (W)

U Usage Factor (%)

Equation 1 will set the main cost formula to apply to our setup:

$$TCO = \sum_{i=1}^{n}(C_{pi} + C_{ei}) \quad (1)$$

Where C_e can be furtherly detailed as follows:

$$C_e = t * C_h * (U * P_f + (1 - U) * P_i) \quad (2)$$

In our case of study, we will limit the energy aspects to a given set of cpu-intensive tasks (Mesos based MapReduce tasks), but in the interest of completeness we must remark that datacenter and input/output related costs are of the utmost importance to the correct applicability of the following calculations.

D. ENERGY MINIMIZATION STRATEGIES

As detailed in [10], there are several layers of energy efficiency mechanisms to consider when designing a cloud computing system, from the hardware perspective to the Data Center level, including the OS and virtualization levels.

For our pretended setup, we must take into consideration the workload for which our setup will be used. For Big Data-oriented case, the following techniques have been proposed for saving energy in Hadoop MapReduce deployments[11]:

- **Covering Set** (CS): By powering off all non-essential nodes, current jobs are delegated to a given subset of nodes, sufficient to cover the task at hand. This implies an increase in the DPS and the time of task completion, but at a lower DPJ, given that less nodes are active.
- **All-In Strategy** (AIS): This strategy proposes to use all the processing power available at a given time, thus reducing the aforementioned DPS metric and increasing the DPJ to achieve the maximum available performance.
- **Berkeley Energy Efficient MapReduce** (BEEMR): Emerges as an improvement over the CS strategy, for real-time processing MapReduce systems; this time defining interactive zones, where real time processing is required and where all nodes will be active, and batch zones, that will be permitted to enter low power states depending on task load.

IV. TCO RESULTS

Most of the proposed capacity allocation algorithms and concepts can be applied without further modifications to embedded system architectures. From the aforegiven strategies, we will implement an **AIS** approach, in which tasks will be completed as soon as possible in order to maximize off-time.

In table II, we will set the values of the previously defined parameters for our calculations. The values as presented are based on mean cost calculations as illustrated in Martens[8].

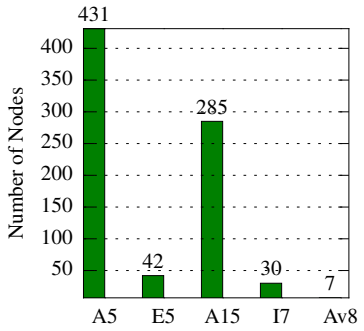


Figure 1. Nodes per DPJ budget

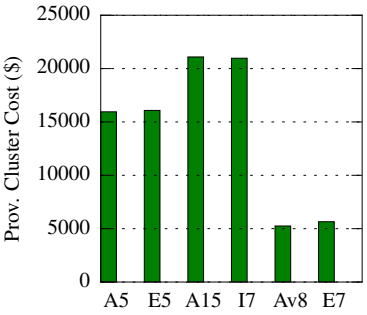


Figure 2. Initial Cost per DPJ budget

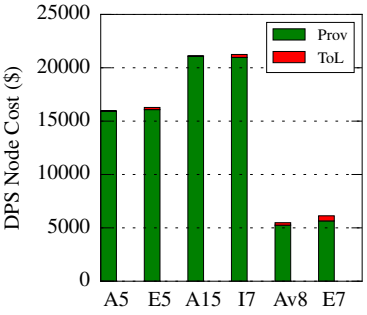


Figure 3. Node Cost per DPS

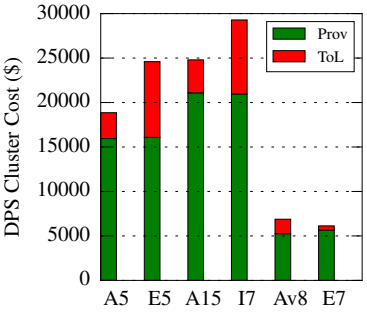


Figure 4. Cluster Cost per DPS

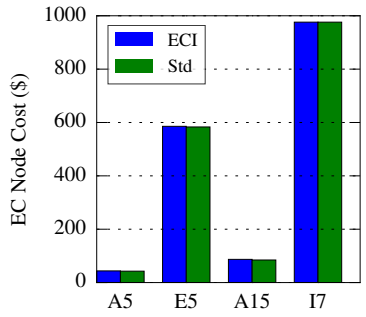


Figure 5. Node Cost with ECI

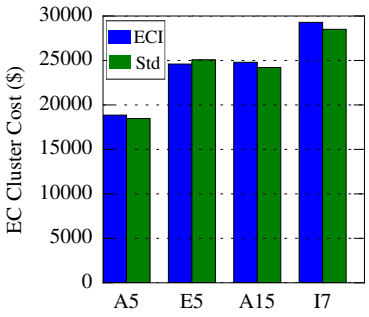


Figure 6. Cluster Cost with ECI

A. ENERGY-CONSTRAINED BUDGET

The calculations for this section will consider costs for a fixed DPJ budget set by the most powerful processor available, the Intel Xeon E7 4890. From here, we will determine how many nodes would be needed to reach said DPJ budget, as well as the provisioning costs for the obtained setup.

In figure 1 we can see how 431, 42, 285,30 and 7 nodes are respectively needed to achieve the Xeon performance level.

We also observe in figure 2 the relative initial cost advantage when provisioning user-grade infrastructure, as opposed to the high cost of deploying a microserver-grade datacenter.

B. TIME-CONSTRAINED BUDGET

In this section, we will analyze the costs associated with a given DPS budget, for a fixed time to task completion of 4 years of full use server lifetime.

Both figure 3 and 4 will represent the aforementioned systems in their X-axes, while the Y-axis will provide the results for the total cost (C_e from the previous formula 1) for the element of study in the given time period, marked in the graph as Time of Life (ToL). The base of the bar (darker colour) represents the provisioning costs (C_{pi}, marked as Prov in the graph) as opposed to the electrical cost (C_{ei} in lighter colour, upper bar).

Table II
TEST DATA

Time of Life (ToL)	4 years
Electric Power Cost (KWh)	$0.11
Mean Usage Factor	0.65

Table III
GOOGLE TCO (4 YEARS)

Grade	*User*	*MicroServer*	*DataCenter*
Google TCO	*New Startup*	*Static Enterprise App*	*Mature App*
Total Cost ($)	6258,76	153,861.6	267,632.84

Figure 3 illustrates the total cost incurred for each node for the given period of time As for figure 4, the total cost for a cluster comprised of the previously calculated number of nodes is shown.

As we can clearly see, Intel processors power usage for a cluster is quite elevated compared with their ARM counterparts (2 to 3 times).

1) ENERGY CONTROLLER INSERTION: As proposed in [12], the insertion of a power controller node with the only task of turning on or off the required nodes as needed reducing idle power consumption, renders a negligible increase in provisioning costs (figure 5) for a relatively positive increase in energy efficiency in the case of user and microserver-grade processor clusters, as seen in Figure 6 where ECI values (green, with legend ECI) are related to the previously studied case (blue, with legend Std).

C. COST-CONSTRAINED BUDGET

This last case of study will model an initial capital inversion based on the three **Google TCO Platform Calculator** (https://cloud.google.com/pricing/tco) profiles that most closely resemble our system partition, which we will detail in table III. We must note that these prices are taken as a fixed reference for processor performance comparison, not as total datacenter costs.

In the first graph from figure 7 we see how even with a limited budget we can keep a considerable number of working ARM nodes for the given timeframe. The second graph from figure 7 closely follows the behaviour introduced in the first one, and shows how a huge number of low power cores at full usage can be more efficient than a reduced number of power-hungry ones. The third graph from figure 7 is consistent with the description in [13], and shows the performance price to pay for this increase in the number of nodes, that is, a decrease in data throughtput clearly limited by the processing capabilities of each individual node.

V. CONCLUSION

As a brief recapitulation of the exposed data as well as a comparison with the current literature, we can summarize our findings as follows:

As stated in previous works [14], ARM-based architectures are not univocally superior to traditional datacenter infrastructures neither in raw data processing nor in standard energy usage. In a first instance this seems to oppose the optimistic findings of Svandeldt-Winter[15], but we must consider the excellent improvements to x86 energy state management in the last few years, which have

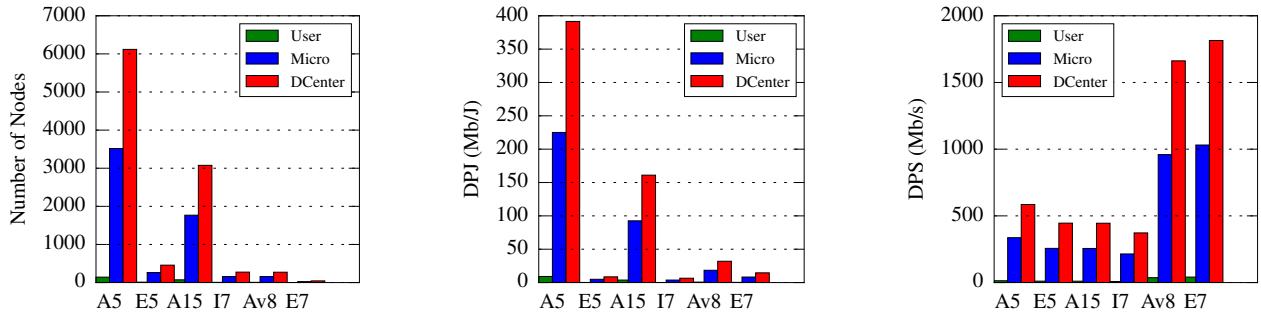


Figure 7. Cost-Driven Budget

increased the DPJ in at least 2 orders of magnitude over the previous generation[5]. However, depending on the task at hand, there seems to be a case to be made for the proposed system in whichever situation involves CPU Intensive workloads, such as the case of small, random database accesses[16]. A generalization of these ideas, based on a similar hardware platform, can also be seen in Cecowski [17] as the proposal for a modular, ARM-based datacenter. Furthermore, in the same trend of our current proposal, Big Data workloads over ARM infrastructures can already be simulated thanks to the work of Kecskemeti[18]

From a cost-based approach, we have improved on [19] by considering typical warehouse time-of-life power consumption, discovering that the accumulated electricity costs for a cluster system clearly cut on the data therein presented.

From an energy-based standpoint, and extending on the outstanding analysis of Tudor [20], we have generalized the energy studies to n-machine cluster structures. This has clearly shown the performance degradation caused by the linear scale of the underlying support hardware (ram bus speed, network, storage), which will hit on the system's performance under I/O Intensive workloads, as is the case with sequential database scans. As pointed in [21], a valid solution for these constraints is the integration of more cpu cores by board, which seems to be consistent with the current market direction.

Cluster job scheduling has also received attention as denoted in [22], as well as physical thermal design [23] and data migration considerations [24]. However insightful these technologies may be for a practical infrastructure deployment, in a first instance they escape the applicable premises for our study.

There are common shortcoming for ARM-based systems pointed at in all the researched literature related to the operating system and software layer, that we'll newly establish here:

- It's necessary to ensure the usage of a parallellization oriented OS (coreos, ranchos) to keep the underlying hardware performance degradation under check.
- The usage of a parallel task optimized management environment (mesos, nomad, hadoop) is also of paramount importance to ensure scale-growth.
- The compile-time optimization of the running code for the specific processor architecture is probably the most determinant and most often underlooked feature that can improve ARM cluster performance.

To conclude, the authors concur on the interest and feasibility of the proposed reference infrastructure, generalizing the DPS and DPJ energy studies to n-machine clusters based on ARM processors, and, given the presented results, also consider the need to consolidate the presented data with the promising current advances in the ARM64 architecture [25].

VI. ACKNOWLEDGEMENTS

The current work has been partially funded by the Ministerio de Educación, Cultura y Deporte of Spain.

REFERENCES

[1] C. Pahl, S. Helmer, L. Miori, J. Sanin, and B. Lee, "A Container-Based Edge Cloud PaaS Architecture Based on Raspberry Pi Clusters," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 2016, pp. 117–124.

[2] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*. Elsevier, 2012.

[3] C. Gough, I. Steiner, and W. Saunders, "CPU Power Management," in *Energy Efficient Servers*. Apress, 2015, pp. 21–70. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4302-6638-9_2

[4] AMD, "ACP — The Truth About Power Consumption Starts Here," AMD, Tech. Rep., 2009, 00000.

[5] Intel, *Intel Brings Core Down to 7W, Introduces a New Power Rating to Get There: Y-Series SKUs Demystified*, 2013, 00000. [Online]. Available: http://www.anandtech.com/show/6655/intel-brings-core-down-to-7w-introduces-a-new-power-rating-to-get-there-yseries-skus-demystified

[6] M. Horowitz, T. Indermaur, and R. Gonzalez, "Low-power digital design," in *Low Power Electronics, 1994. Digest of Technical Papers., IEEE Symposium*. IEEE, 1994, pp. 8–11, 00362. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=573184

[7] J. H. Laros III, K. Pedretti, S. M. Kelly, W. Shu, K. Ferreira, J. Vandyke, and C. Vaughan, "Energy delay product," in *Energy-Efficient High Performance Computing*. Springer, 2013, pp. 51–55, 00005. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4471-4492-2_8

[8] B. Martens, M. Walterbusch, and F. Teuteberg, "Costing of cloud computing services: A total cost of ownership approach," in *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE, 2012, pp. 1563–1572, 00050. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6149074

[9] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 8, no. 3, pp. 1–154, 2013, 01105. [Online]. Available: http://www.morganclaypool.com/doi/abs/10.2200/S00516ED2V01Y201306CAC024

[10] A. Beloglazov, "Energy-efficient management of virtual machines in data centers for cloud computing," 2013, 00028. [Online]. Available: https://minerva-access.unimelb.edu.au/handle/11343/38198

[11] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/cpe.1867/full

[12] N. Maheshwari, R. Nanduri, and V. Varma, "Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 119–127, 2012, 00061. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167739X1100135X

[13] M. Malik and H. Homayoun, "Big data on low power cores: Are low power embedded processors a good fit for the big data workloads?" in *Computer Design (ICCD), 2015 33rd IEEE International Conference on*. IEEE, 2015, pp. 379–382, 00001. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7357128

[14] D. Loghin, B. M. Tudor, H. Zhang, B. C. Ooi, and Y. M. Teo, "A performance study of big data on small nodes," *Proceedings of the VLDB Endowment*, vol. 8, no. 7, pp. 762–773, 2015, 00006. [Online]. Available: http://dl.acm.org/citation.cfm?id=2752945

[15] O. Svanfeldt-Winter, S. Lafond, and J. Lilius, "Cost and energy reduction evaluation for ARM based web servers," in *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*. IEEE, 2011, pp. 480–487. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6118745

[16] B. Kantarci, L. Foschini, A. Corradi, and H. T. Mouftah, "Inter-and-intra data center VM-placement for energy-efficient large-scale cloud systems," in *Globecom Workshops (GC Wkshps), 2012 IEEE*. IEEE, 2012, pp. 708–713, 00019. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6477661

[17] M. Cecowski, G. Agosta, A. Oleksiak, M. Kierzynka, M. v. d. Berge, W. Christmann, S. Krupop, M. Porrmann, J. Hagemeyer, R. Griessl, M. Peykanu, L. Tigges, S. Rosinger, D. Schlitt, C. Pieper, C. Brandolese, W. Fornaciari, G. Pelosi, R. Plestenjak, J. Cinkelj, L. Cudennec, T. Goubier, J. M. Philippe, U. Janssen, and C. Adeniyi-Jones, "The M2dc Project: Modular Microserver DataCentre," in *2016 Euromicro Conference on Digital System Design (DSD)*, 2016, pp. 68–74.

[18] G. Kecskemeti, W. Hajji, and F. P. Tso, "Modelling Low Power Compute Clusters for Cloud Simulation," in *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, Mar. 2017, pp. 39–45.

[19] Z. Ou, B. Pang, Y. Deng, J. K. Nurminen, A. Yla-Jaaski, and P. Hui, "Energy-and cost-efficiency analysis of arm-based clusters," in *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*. IEEE, 2012, pp. 115–123. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6217412

[20] B. M. Tudor and Y. M. Teo, "On understanding the energy consumption of arm-based multicore servers," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 41. ACM, 2013, pp. 267–278. [Online]. Available: http://dl.acm.org/citation.cfm?id=2465553

[21] N. Rajovic, L. Vilanova, C. Villavieja, N. Puzovic, and A. Ramirez, "The low power architecture approach towards exascale computing," *Journal of Computational Science*, vol. 4, no. 6, pp. 439–443, Nov. 2013, 00046. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877750313000148

[22] S. Zikos and H. D. Karatza, "Performance and energy aware cluster-level scheduling of compute-intensive jobs with unknown service times," *Simulation Modelling Practice and Theory*, vol. 19, no. 1, pp. 239–250, 2011, 00048. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1569190X10001309

[23] R. T. Kaushik and K. Nahrstedt, "T: a data-centric cooling energy costs reduction approach for big data analytics cloud," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society Press, 2012, p. 52, 00027. [Online]. Available: http://dl.acm.org/citation.cfm?id=2389067

[24] C. Ghribi, M. Hadji, and D. Zeghlache, "Energy Efficient VM Scheduling for Cloud Data Centers: Exact Allocation and Migration Algorithms." IEEE, May 2013, pp. 671–678, 00041. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6546155

[25] K. Keipert, G. Mitra, V. Sunriyal, S. S. Leang, M. Sosonkina, A. P. Rendell, and M. S. Gordon, "Energy-Efficient Computational Chemistry: Comparison of x86 and ARM Systems," *Journal of chemical theory and computation*, vol. 11, no. 11, pp. 5055–5061, 2015, 00001. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b00713

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6574

Aplicación de técnicas de detección de anomalías a escenarios de ciudades inteligentes

Irene Romero, Carolina Alonso, Víctor A. Villagrá, Luis Vázquez, Pilar Holgado
Departamento de Ingeniería de Sistemas Telemáticos
Universidad Politécnica de Madrid
iromero@dit.upm.es, carolina.alonso.lopez@alumnos.upm.es. villagra@dit.upm.es, lvazquez@dit.upm.es, pholgado@dit.upm.es

***Resumen*- Una de las grandes preocupaciones en la actualidad de las empresas es la detección y prevención temprana de ataques de ciberseguridad. Para ello, existen los Sistemas de Detección de Intrusiones, herramientas que cuentan con sensores virtuales y que basan su detección en el análisis del tráfico de red. El problema surge cuando se dan ataques que estos sistemas no detectan. Una de las soluciones existentes a esta problemática es acudir a la minería de datos e intentar detectar anomalías en grandes volúmenes de datos, no pertenecientes únicamente al tráfico de red sino datos que puedan provenir de diversas fuentes. En este artículo se propone una solución enmarcada en el proyecto DHARMA haciendo uso de la técnica de agrupamiento, dentro de la disciplina de la minería de datos, en concreto del algoritmo DBSCAN.**

***Palabras clave* – DBSCAN, minería de datos, agrupamiento, detección de anomalías.**

I. INTRODUCCIÓN

Todos los ámbitos de las tecnologías de la información tienen un concepto en común: la innovación. A diario encontramos nuevos lenguajes de programación, nuevas y útiles librerías para aquellos lenguajes ya existentes, e incluso nuevos paradigmas de desarrollo. Y esto también se cumple en el campo de la ciberseguridad: nuevas vulnerabilidades, nuevos métodos de ataque y, por supuesto, formas de luchar contra ellos. Por otro lado, las nuevas tecnologías son clave en el desarrollo de soluciones de software innovadoras. En concreto, el análisis estadístico y la minería de datos son el segundo conocimiento más valorado en Linkedin para encontrar trabajo en 2017 [1], y la razón detrás de ello es el hecho de que son tecnologías emergentes cuyas posibilidades de aplicación todavía no están totalmente explotadas e incluso se desconocen.

Dado el auge de las comunicaciones a distancia y de Internet en concreto, una de las grandes preocupaciones de las empresas y de las organizaciones es la detección y prevención de ataques de ciberseguridad. Que una persona al otro lado del mundo pueda infiltrarse en tu sistema sin que seas consciente puede resultar extremadamente problemático. Además, no solo es necesario saber que tus sistemas están siendo atacados sino que hacerlo en el momento oportuno, en etapas tempranas del ataque, puede resultar clave.

La minería de datos toma parte en este apartado: la mayoría de las empresas, y sobre todo aquellas que trabajan online, cuentan volúmenes de datos recogidos durante años, muchas veces sin explotar, de los que se puede extraer información. Se puede aplicar al problema anterior y analizar qué sucesos pueden indicar que un ataque está en proceso, o incluso anticiparse y detectarlo en etapas no críticas, de forma que se puedan tomar las salvaguardas apropiadas antes de que haya causado daño.

Definir cuáles son los indicios de un ciberataque no supone una tarea simple. Existen sucesos evidentes, como la presencia de un nuevo usuario en nuestros sistemas accediendo a información crítica, pero la problemática surge en relacionar sucesos que puedan considerarse irrelevantes con los datos disponibles y detectar así una posible anomalía. Es este el problema para el que trata de plantear solución este artículo.

En este documento, la detección de sucesos anómalos se enmarca dentro de un proyecto de evaluación y gestión dinámica de riesgos denominado DHARMA [2], [3] (Dynamic Heterogeneous threAts Risk Management and Assessment).

El análisis y gestión de riesgo tradicionales se encargan de identificar activos y sus vulnerabilidades, amenazas sobre estos y probabilidad de ocurrencia (obteniendo el riesgo sobre un activo) y salvaguardas para reducir este riesgo. Para ello se siguen distintas metodologías como MAGERIT, ISO/IEC 27005, OCTAVE, etc. No obstante, estas metodologías son estáticas, pues únicamente se ejecutan en momentos determinados, y el contexto puede variar completamente entre ejecuciones.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

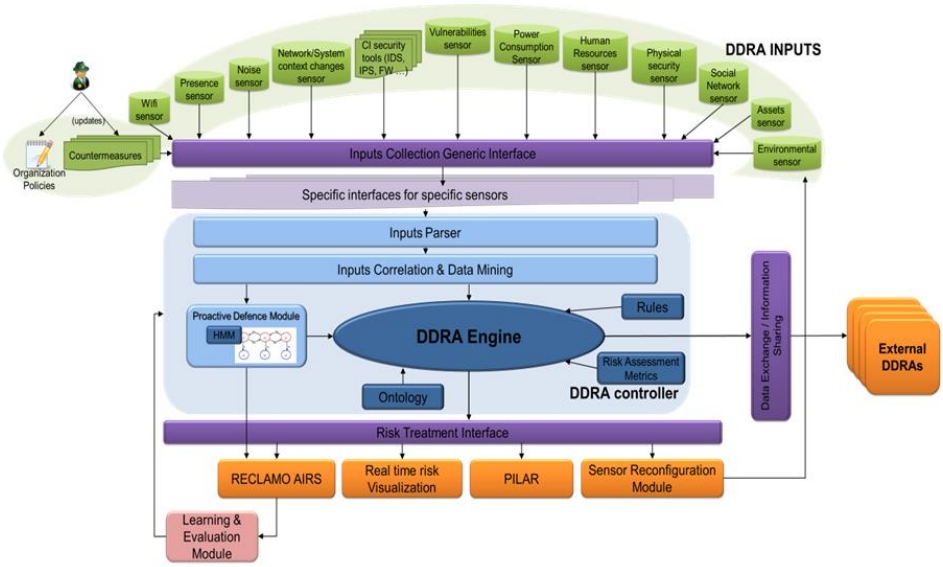


Fig. 1 Arquitectura DHARMA

Este sistema se encarga de dar respuesta a incidentes o ataques en un corto plazo de tiempo, pudiendo responder a esos cambios de contexto de una forma dinámica. Para ello se ha propuesto una arquitectura en la que el contexto de la organización se modela mediante un gran número de sensores heterogéneos (wifi, Bluetooth, presencia, estado de activos, crispación laboral, ambientales, actividad en redes sociales, etc.), como se observa en la figura 1. Los datos obtenidos por estos sensores serán analizados, correlados (el tema principal de este artículo) y enviados a un motor de análisis de riesgo que calculará el riesgo instantáneo de la organización y desplegará las contramedidas pertinentes (visualización en tiempo real, reconfiguración de los sensores, alimentación de un Sistema de Respuesta Automática a Intrusiones o herramientas clásicas de análisis de riesgo como PILAR, etc.).

La propuesta estudiada en este artículo en encuentra en el nivel "Inputs Correlation and Data Mining" dentro de la arquitectura DHARMA, como se observa en la figura 1.

La aplicabilidad de esta arquitectura y técnica se explorará a través del caso de uso de ciudades inteligentes.

II. Técnicas de detección de anomalías con minería de datos

Una anomalía es una observación significativamente diferente al resto de observaciones, de lo que se deduce que dicha observación puede haber sido generada por un mecanismo diferente del resto de observaciones.

La detección de anomalías surge como un área de trabajo importante en los algoritmos de aprendizaje. En algún caso la capacidad de detección de anomalías deriva de forma natural de los algoritmos de agrupamiento, que se estudiarán a continuación.

El agrupamiento es una disciplina de la minería de datos que consiste en la clasificación de las entradas de un set de datos en distintos grupos, pero sin tener ningún tipo de información a priori sobre éstos. Es decir, trata de buscar algún tipo de relación desconocida que pueda indicar que un subconjunto de nuestros datos tenga algo en común entre sí.

En este caso se explotará una de sus utilidades: la detección de anomalías. Al hacer agrupamiento, lo normal es obtener una serie de clústeres y, en función del tipo de algoritmo, ruido. El ruido está formado por aquellos puntos que no cumplen las condiciones necesarias para formar parte de ningún otro clúster; normalmente están demasiado lejos de otros puntos. De esta forma, se puede definir el ruido como puntos que están fuera de la norma, del comportamiento del resto del set de datos. En el caso de datos de uso de redes Wifi, una anomalía sería que el tiempo de conexión de un usuario fuese mayor a un día o que consumiese más ancho de banda del que debería.

Por otro lado, hay distintos métodos para agrupar datos: por distribución estadística, asignando un centroide a cada grupo y calculando la distancia a éste, basados en la densidad de puntos en el espacio bidimensional… A continuación se explica el funcionamiento de algunos de estos métodos.

A. Agrupamiento por centroide

En este modelo se asigna un centroide a cada clúster, de forma que todos los elementos que pertenecen a ese clúster estén más cerca de su centroide que del centroide de otros clústeres. El punto asignado al propio centroide depende de los elementos que haya en el clúster, y su valor no tiene por qué ser un punto del set de datos.

Ejemplo: K-means. Se trata de un algoritmo de agrupamiento por centroide en el que el número de clústeres está prefijado, y viene dado por el valor k. De esta forma, para k=2 tendremos dos clústeres. El esquema que sigue es el siguiente:

1. Asignar k centroides de forma aleatoria.
2. Asignar cada elemento del set de datos al centroide que proporcione la mínima suma de distancias al centroide.
3. Recalcular los centroides.
4. Repetir hasta que converja.

Se considera que el algoritmo converge cuando durante el paso 3 los centroides no cambian, o cuando la distancia entre centroides de una iteración a la siguiente es menor que cierto parámetro de precisión predefinido.

B. Agrupamiento por distribución

Este modelo está basado en el ajuste de cada clúster a una distribución estadística, de forma que cada punto se asigna al clúster al que tiene la máxima probabilidad de pertenecer.

Ejemplo: algoritmo esperanza-maximización. Se trata de un método de agrupamiento por distribución en dos pasos: durante el primero, llamado paso de "esperanza", se calcula una función de la media de la función de verosimilitud de los actuales parámetros; durante el segundo, el de "maximización", se maximiza dicha función para buscar una solución óptima.

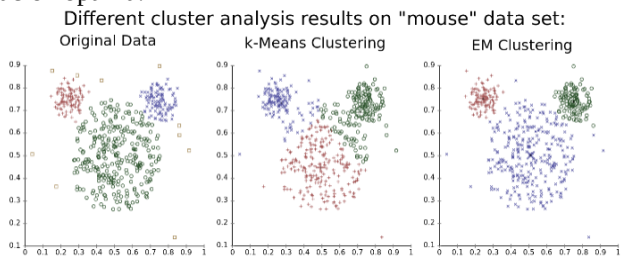


Fig. 2 Comparación del resultado de aplicar k=3 means respecto a esperanza-maximización

C. Agrupamiento por densidad

En este modelo se considera un grupo como una zona del espacio en el que la densidad de puntos es mayor que en el resto. Tienen la ventaja de que, si en una zona no hay suficiente densidad de puntos, los que se encuentren en ella se consideran ruido, que se puede interpretar como anomalías.

Ejemplo: DBSCAN: representa las siglas en inglés de "Agrupamiento Espacial Basado en Densidad de Aplicaciones con Ruido". En este algoritmo contamos con dos parámetros: **minPts** (número mínimo de puntos) y **ε** (épsilon, que en el código se representa como eps). Por otro lado, los puntos se clasifican en tres tipos: núcleos, no-núcleos y ruido. Un **núcleo** es un punto que tiene a una distancia épsilon a al menos minPts-1 puntos, por lo que juntos forman un vecindario de radio ε. Un **no-núcleo** es un punto vecino de un núcleo, pero que no es núcleo por sí mismo. Y finalmente, un punto es considerado **ruido** si no es vecino de ningún núcleo.

De esta manera, en DBSCAN un grupo de puntos forman un clúster si todos ellos son núcleos o vecinos de al menos un núcleo, y los puntos que no forman parte de ningún grupo de tamaño minPts son ruido. El esquema que sigue el algoritmo es el siguiente:

1. Elegir un punto aleatorio no visitado y moverlo a la lista de visitados.
2. Calcular su vecindario en un radio ε.
    a. Si no está formado por minPts elementos, etiquetar como ruido y volver a 1.
    b. Si está formado por al menos minPts elementos, iniciar un clúster y añadir a sus vecinos a la lista de puntos candidatos a formar parte de éste.
        i. Para cada punto candidato, si no ha sido visitado, calcular sus vecinos y comprobar si es núcleo.
            1. Si lo es, añadir sus vecinos a la lista de candidatos.
        ii. Para cada punto candidato, si no forma parte de ningún clúster, añadirlo al actual.
3. Si quedan puntos sin visitar, volver a 1.

A pesar de que tiene la ventaja de permitir encontrar anomalías, su principal punto en contra es que no es determinista. Un punto no-núcleo puede ser vecino de varios núcleos que a su vez formen parte de distintos clústeres, por lo que en el resultado final puede estar en cualquiera de ellos. Sin embargo, esta desventaja no afecta a esta solución, ya que en ella se utilizará DBSCAN sólo para obtener el ruido en un set de datos. DBSCAN también tiene una serie de características que pueden resultar interesantes al elegir un algoritmo de detección de anomalías: es determinista en cuanto a ruido y no requiere que haya una correlación alta entre variables, sólo que

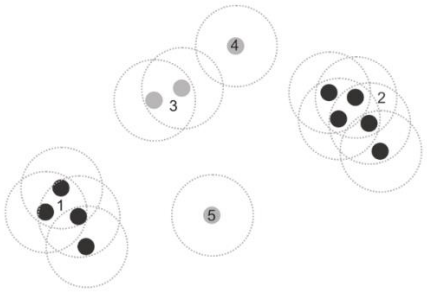


Fig. 3 Ejemplo de distintos tipos de puntos tras ejecutar DBSCAN para $minPts = 3$ y ϵ representada por las líneas discontinuas. 1 y 2 representan dos grupos, mientras que 3, 4 y 5 no cumplen la condición y se consideran ruido.

las variaciones en el set de datos sean suaves. En cuanto a sus desventajas. En cuanto a sus desventajas, dependiendo de la implementación puede ser lento, con una complejidad temporal entre O(*nlogn*) y O(n^2), sucediendo lo mismo en cuanto a memoria, con complejidad entre O(n) y O(n^2)[4]. También implica tener un conocimiento previo de los datos, ya que es complicado determinar los valores adecuados de ϵ y minPts al vuelo, y la utilidad de su aplicación a variables nominales es discutible, ya que no siempre es inmediato entender qué mide una distancia en este tipo de valores.

Finalmente, a pesar de ser relativamente antiguo, al haber sido publicado originalmente en 1996, DBSCAN es un algoritmo que sigue utilizándose en un amplio campo, recibiendo incluso en 2014 el *Test of Time Award* de *KDD*[5], y habiendo probado anteriormente[6-7] su utilidad en la detección temprana de riesgos de seguridad.

III. Caso de uso: Ciudades Inteligentes

El término "Ciudad Inteligente" (o Smart City) es el nombre que se da a una ciudad que utiliza las tecnologías de la información y el internet de las cosas para facilitar servicios a sus residentes y mejorar el funcionamiento global de la ciudad. Se basan en la presencia de sensores por distintas zonas, como pueden ser cámaras, sensores de humedad o de presencia en las aceras.

Las tecnologías de las Smart Cities tienen gran cantidad de aplicaciones, principalmente enfocadas a la prestación de servicios eficientes y útiles a sus habitantes. Son un modo de mejorar el nivel de vida de la población, y el hecho de que a día de hoy todo el mundo esté conectado en todo momento, es especialmente fácil encontrar aplicaciones con las que podrán interactuar personalmente.

La tendencia a evolucionar hacia ciudades inteligentes, dada la velocidad de crecimiento de las áreas urbanas, es cada

vez más acusada. Teniendo en cuenta que se prevé que el 70% de los seres humanos habiten en centros urbanos en 2050, es necesario avanzar hacia ciudades que permitan una gestión menos manual y más sostenible. De esta forma, todos los elementos que conforman la ciudad han de evolucionar hacia plataformas inteligentes para lograr una gestión eficiente y crear ciudades formadas por una red compleja de sensores que proporcionarán información clave para el funcionamiento y la toma de decisiones.

Esta compleja red de sensores que forman las plataformas de ciudades inteligentes pueden ser utilizadas para múltiples ámbitos: uno de ellos es la ciberseguridad, o cómo aprovechar el gran volumen de datos que estos sensores proporcionan para poder correlar sus anomalías con posibles intrusiones en entornos de ciberseguridad en sistemas y plataformas albergadas en estas ciudad inteligente. Se trata de un enfoque holístico: la combinación de detección de anomalías en el entorno físico (proporcionado por los sensores de la ciudad inteligente) y entorno lógico (proporcionado por los sensores propios de ciberseguridad como sistemas de detección de intrusiones, etc.)

El uso de técnicas de detección de anomalías en las Smart Cities resulta de mucha utilidad para extraer información valiosa o detectar comportamientos anómalos que puedan generarse en estas ciudades. Dada la cantidad de sensores con los que contará una Smart City, es necesario implementar una técnica que permita analizar el elevado volumen de datos generado y, de esta forma, que la gestión de dicha ciudad sea también inteligente. Los comportamientos anómalos detectados podrían pertenecer a ataques y la sola gestión manual de un operario no sería suficiente para detener dicho ataque.

El escenario donde se enmarca este trabajo es una plataforma de procesamiento de datos procedentes de la City of the Future de la Universidad Politécnica de Madrid. Esta iniciativa, creada y mantenida tanto por docentes como alumnos, está centrada en la investigación, desarrollo e innovación en el campo de las City Sciences, y aborda temas tan amplios como el control de tráfico, la gestión de emergencias o las *smart grids*.

La City of the Future es el nombre que se le da al conjunto de sensores y fuentes de datos que tiene la UPM por toda la ciudad de Madrid, con mayor presencia en el área de Ciudad Universitaria y Moncloa. Está formada por sensores de clima, de tráfico, de conexiones Bluetooth, etc. Supone un proyecto enfocada a la aplicación de las tecnologías de la información para crear una Smart City en Madrid. Al tratarse de un proyecto universitario, su principal objetivo es la investigación y el desarrollo de técnicas nuevas en un ámbito docente, proporcionando a los estudiantes acceso a proyectos diversos con los que enriquecer sus conocimientos, además de ayudar a la prestación de servicios inteligentes a los habitantes de la propia ciudad.

Este trabajo se centra en el tratamiento de la información generada por sistemas ya implementados para obtener información relevante mediante minado de datos.

Dado que los datos procederán de fuentes heterogéneas, será importante crear una plataforma que pueda tratar con datos de distintos tipos (no sólo numéricos), de gran tamaño o de distintos formatos. También será fundamental que se trate de un sistema multiplataforma y ligero para el usuario, además de poder hacer un volcado de los resultados para posibles aplicaciones futuras.

Como ya se ha comentado, la aplicación estará enfocada al tratamiento de los datos generados por estas aplicaciones, de los que se intentará extraer información subyacente. El proyecto más relevante es el Smart CEIM.

Como se puede ver en la figura 4, la plataforma Smart CEIM está formada por una red de sensores y actuadores conectados a la plataforma de almacenamiento y al cuadro de control mediante IP.

Los datos recogidos por dicha plataforma arrojan información ambiental e información sobre las redes WiFi y su uso, por lo que sólo detectaremos comportamiento anómalo en dichos datos, identificando posibles amenazas en estas dimensiones.

Los datos ambientales provienen de sensores ambientales colocados en las bibliotecas de las distintas escuelas, así como en exteriores (en zonas resguardadas para evitar daños por lluvia o sol). Estos sensores proporcionan valores en tiempo real de temperatura ambiente, humedad, nivel de luz, batería del dispositivo, monóxido de carbono, dióxido de nitrógeno y nivel de ruido. Además, cada valor incluye una marca de tiempo que permite conocer cómo varía cada variable a lo largo del día.

En cuanto a los sensores WiFi, se cuenta con los datos de uso de redes WiFi en las bibliotecas de las escuelas de Aeronáuticos, Industriales y Telecomunicaciones. En la escuela de Telecomunicaciones, además, se contará con acceso a los datos de varias subredes repartidas por los edificios. Dichos sensores están colocados en zonas de mucho tránsito, como bibliotecas, ya que su finalidad original es la monitorización de personas.

La información que proporcionan estos sensores incluye la relativa a los dispositivos conectados (dirección MAC, canal al que se han conectado, potencia, etc.) así como estadísticas que pueden ser de utilidad (número total de dispositivos conectados, número total de conexiones simultáneas a la red, etc.).

Así pues, la detección de anomalías se realizará únicamente sobre estos dos conjuntos de datos, permitiendo detectar comportamientos anómalos en el uso de las redes WiFi o en los datos ambientales; ya que ambos pueden ser indicadores de un problema mayor.

IV. VALIDACIÓN

La implementación de este sistema se ve dividida en distintos módulos, que se exponen a continuación:

A. Preprocesado

Como se ha comentado, la aplicación cuenta con un apartado de preprocesado, mediante el cual se consiguen dos cosas: en primer lugar, darle información al algoritmo sobre cómo son los datos con los que va a tratar, y en segundo lugar modificar los datos para que DBSCAN proporcione información más relevante, sin modificar su significado.

El flujo básico de la aplicación es la siguiente: un usuario sube datos en cualquier momento al servidor, indicando en la pantalla de preprocesado los parámetros necesarios, como el tipo de las variables. Estos parámetros se almacenan en un archivo de configuración, que se utilizará cada vez que el usuario decida entrar de nuevo en la aplicación. En la vista principal se muestra una lista con todos los sets de datos disponibles y sus archivos de configuración correspondientes, de forma que cuando el usuario lo desee, puede seleccionarlos y ejecutar los cálculos que ha indicado.

Una vez recibidos los resultados, se pasa a la vista de visualización de resultados, en la que el usuario puede detectar las anomalías entre cada par de variables.

B. Detección de anomalías

Para la detección de anomalías se ha utilizado R, dado su versatilidad en lo que a minería de datos se refiere.

En cuanto a herramientas adicionales, se han utilizado varias librerías para funciones no disponibles por defecto en R: en primer lugar se encuentra *devtools*, utilizada para crear e instalar funciones personalizadas dentro del entorno de ejecución. En segundo lugar se encuentra *OpenCPU*, que proporciona acceso a dichas funciones mediante peticiones HTTP. La ventaja de usar estas dos herramientas yace en el hecho de que se está realizando minería de datos, por lo que conviene dedicar su propio servidor a esta herramienta, evitando así quedarse sin recursos, tanto con potencia de procesamiento como con memoria.

Entrando en la implementación de la detección de anomalías, hay que tener en cuenta una serie de consideraciones:

- En primer lugar, el algoritmo DBSCAN sólo puede recibir valores completos. En caso de recibir datos *null*, *undefined* o vacíos, salta un error y para la ejecución. Esto se soluciona fácilmente en R mediante la función *complete.cases(datos)*, que filtra los datos incompletos.
- En segundo lugar, limpiar los datos de valores no completos tiene una consecuencia: una vez se tengan los resultados, éstos no serán aplicables al set de datos original. Por meros estándares de calidad, es conveniente que los resultados se generen de forma concordante con los datos recibidos. La solución para este problema consiste en considerar que estos datos incompletos son anomalías en sí, e intercalar los resultados de DBSCAN con estas consideraciones en el orden adecuado.
- En tercer lugar, los valores deben estar normalizados. Es uno de los errores más comunes cuando se agrupa utilizando medidas de distancia: no se tiene en cuenta que distintas variables representan métricas distintas, por lo que, si una tiene un rango mucho mayor, la medida de distancia sólo será representativa para ella. Como se utiliza DBSCAN para detectar anomalías con pares de variables, esto hará que se obtengan resultados erróneos al tratar con un par de variables de rangos muy distintos. Normalizándolas entre 0 y 1 este apartado queda solucionado.
- En cuarto lugar, como se ha comentado, DBSCAN hace uso de dos parámetros: $minPts$ y ϵ. Por comodidad, siendo lo más común, se fija el valor de $minPts$ en 3, de forma que podamos calcular ϵ en función de los datos con los que se esté tratando en cada momento. Para ello se utiliza una función de la desviación estándar de ambas variables.
- Finalmente, no se detectarán anomalías para variables nominales, pues es de dudosa utilidad medir la distancia entre este tipo de variables, por lo que directamente se descartan. Por simplificación del código, simplemente se oculta el botón correspondiente en la vista, de forma que no haya que realizar comprobaciones de los tipos de todas las variables en cada llamada a la función

De esta manera, el esquema que sigue la función de detección de anomalías es la siguiente:

- Lectura de datos
- Filtrado de valores incompletos
- Normalización
- Cálculo de ϵ a partir de las desviaciones estándar
- Ejecución de DBSCAN y obtención de anomalías
- Ampliación de resultados a set de datos completos
- Devolución de resultados

C. Detección de anomalías en tiempo real

El sistema permite también la detección de anomalías en tiempo real, dentro del proyecto DHARMA, que utiliza esta propuesta como un indicador de posibles ataques en fases tempranas.

Esta nueva idea sigue el siguiente esquema: se cuenta con varios flujos de datos constantes, cuya frecuencia de actualización depende de las fuentes. En el momento en el que se reciban datos nuevos, un *watchdog* los detecta y preprocesa para ser almacenados en distintos sets de datos. Dónde se almacenan depende de parámetros predefinidos de actividad, esto se debe a que no es lo mismo tener tráfico alto en un sistema en jueves por la mañana, en horario laboral, que en la noche del sábado al domingo. Es importante diferenciar entre períodos de alta y baja actividad, así que mediante el preprocesado se indicarán los rangos de cada uno. Una vez se tiene esta información, se analiza y se almacena en los sets de datos correspondientes, cambiado el formato de las marcas temporales para favorecer la ejecución del algoritmo de agrupamiento. Finalmente, en el momento en el que se ha terminado de procesar toda la información, se ejecuta DBSCAN para todos los sets de datos con nuevas entradas, y, en el caso de que haya nuevas anomalías, se informa a DHARMA como si se tratase de un indicio inicial de ataque.

Teniendo en cuenta estas consideraciones, el preprocesado tiene dos fases: una inicial en la que se crea un fichero de configuración con información sobre los datos (nombres de variables y sus tipos, cuáles son marcas temporales y sus formatos, y los periodos de alta y baja actividad), y una segunda que tiene lugar cada vez que se reciben datos nuevos, en la que se procesan los datos, ajustando su formato para que sea más adecuada su visualización y para almacenarlos donde corresponda.

Cabe destacar que el preprocesado en esta implementación es distinto al anterior: tiene lugar en el momento en el que se configura el servidor para recibir datos de una fuente, ya que hasta que no se cuente con la información de preprocesado no se podrá ejecutar el código de detección de anomalías.

En cuanto a los cambios de formato, un ejemplo es el de las marcas temporales. En primer lugar, se pasan a formato UNIX, dado que éste permite tener variaciones constantes entre instantes consecutivos. Esto se debe a que los sets de datos con los que se ha trabajado durante el desarrollo del proyecto siguen formatos similares a ISO 8601, en los que se emplean directamente la fecha y la hora. La desventaja de estas configuraciones es que los rangos que sigue cada uno de los elementos no son decimales: la hora varía entre 00 y 23, el mes entre 1 y 12, etc. y a la hora de representarlos en un eje temporal se producen saltos que harán que al ejecutar DBSCAN se identifiquen como grupos distintos valores que realmente son cercanos.

Otra consideración en lo que a cambio de formato se refiere es cómo asegurarse de que se utilizan correctamente los rangos de actividad. Su finalidad es que, en el momento de detectar anomalías, estemos comparando los nuevos datos

con otros equivalentes, como puede ser datos de uso de redes Wifi en una universidad durante el curso. Dados los horarios utilizados en estos tipos de instituciones, cabría deducir que el tráfico durante todos los martes debería ser similar a las mismas horas. Por tanto, el formato de la marca de tiempo se ha cambiado para indicar sólo la hora.

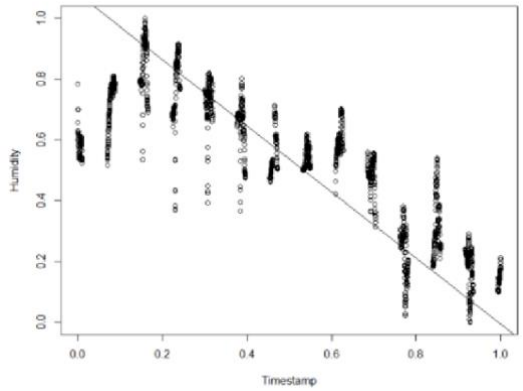


Fig. 5 Ejemplo de visualización de datos con eje temporal con formato ISO 8601

Una vez realizados estos procesos, los datos se encuentran ya en un formato adecuado para la detección de anomalías.

D. Postprocesado

Una vez se han recibido los datos de R, hay que realizar un postprocesado. La razón por la que sucede esto es que mediante DBSCAN se van a detectar anomalías en todo el set de datos, incluso anomalías antiguas que ya no son motivo de alarma.

Para ello, una vez recibidos los resultados se comparan con los anteriores, que estarán almacenados. En caso de haber nuevas anomalías, se notifica a DHARMA. Por otro lado, en todos los casos se almacenan los nuevos resultados, para utilizarlos en la próxima llamada.

También se tiene en cuenta el caso inicial, en el que no hay resultados antiguos. La decisión tomada para este apartado es simple: se notifica a DHARMA de todas las anomalías presentes, ya que, si en el sistema no se encuentra en proceso un ataque, será DHARMA quien lo deducirá.

V. PROTOTIPO: APLICACIÓN WEB

En un principio, la idea a implementar consiste en una aplicación web que permite al usuario subir sus sets de datos, realizar el preprocesado manualmente y, una vez almacenados, calcular la correlación entre distintas variables, visualizarlas en un diagrama de dispersión y detectar anomalías.

Para ello se ha hecho uso de las siguientes tecnologías:

- HTML, CSS, AngularJS y D3.js para el front end.
- Node.js y Express para el back end.
- R para el motor de cálculo.

Por otro lado, R no sólo se utiliza mayormente para realizar minado de datos, sino que, en el caso de querer ampliar las funcionalidades de la aplicación, permitiría tener acceso a gran cantidad de funciones pre implementadas y ampliar el alcance de manera rápida.

En cuanto al diseño de la aplicación, la forman tres vistas que se detallarán a continuación.

A. Vista principal

En la vista principal de la aplicación se permite al usuario seleccionar un set de datos existente o subir uno nuevo.

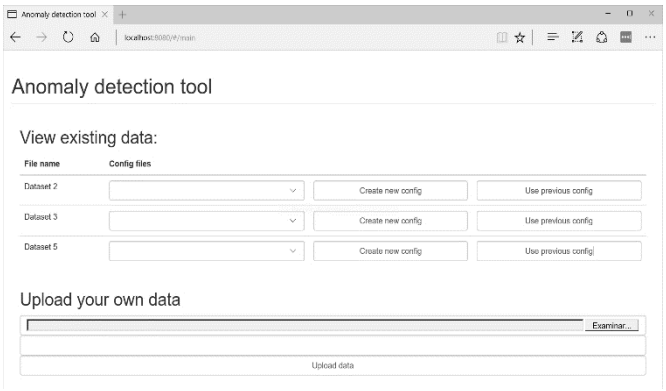


Fig. 6 Vista principal

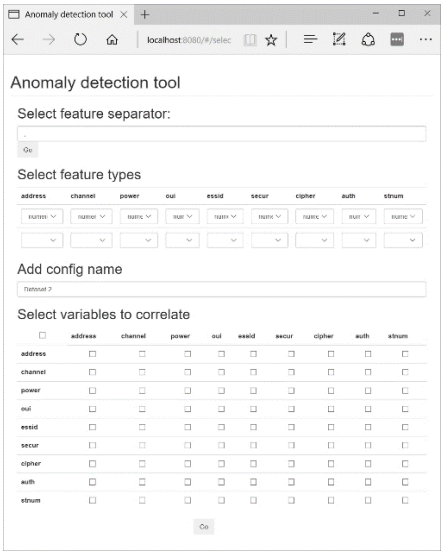


Fig. 7 Vista de selección de parámetros de preprocesado

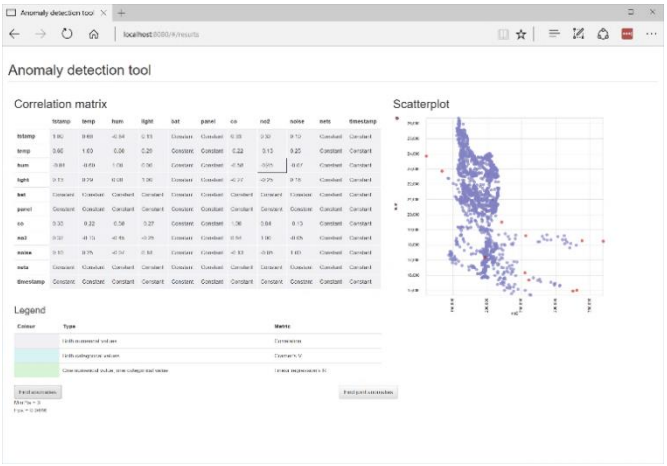


Fig. 8 Vista de visualización de resultados con las anomalías representadas en rojo

B. Vista de preprocesado

En esta vista se permite al usuario definir ciertos parámetros para el correcto funcionamiento del algoritmo de agrupamiento.

C. Vista de visualización de resultados

Aquí el usuario puede visualizar la matriz de correlación de su set de datos, junto con un diagrama de dispersión al hacer hover sobre sus distintos valores. Haciendo click en cualquiera de ellos podrá bloquear la gráfica y seleccionar la opción de detectar anomalías.

VI. CONCLUSIONES

En este artículo se ha analizado la aplicación de un algoritmo de detección de anomalías, en concreto de un algoritmo de agrupamiento. La aplicabilidad del mismo se ha estudiado a través del caso de uso de ciudades inteligentes donde la gestión de la misma ha de ser inteligente y son necesarios algoritmos como el empleado.

Una vez terminada la implementación, los resultados obtenidos tras probarlos son positivos. No se puede hablar de términos como precisión, ya que no se ha contado con información sobre si un dato es anómalo o no, pero tras la ejecución del algoritmo y mediante la visualización se puede observar cómo aquellos datos que distan de los demás se han detectado correctamente como anomalías.

Esto es también aplicable a la solución en tiempo real, para la que se han proporcionado unos parámetros más laxos en la implementación de DBSCAN. La razón es su propia aplicación: al tratarse de una herramienta para detección temprana de ataques, es importante no dejar pasar ningún falso negativo, de forma que es preferible que salten falsas alarmas, que DHARMA se encargará de filtrar posteriormente.

Como se ha comentado a lo largo del documento, detectar ataques en fases tempranas puede resultar crítico en sistemas en los que la seguridad es importante, y utilizar tecnologías nuevas, como es el minado de datos, puede resultar extremadamente útil y resultar en soluciones innovadoras con múltiples aplicaciones.

Finalmente, ha destacado la importancia de investigar en el ámbito del minado de datos: que un algoritmo tan antiguo haya servido para añadir funcionalidades extra en un ámbito como la detección de ataques de ciberseguridad en ciudades inteligentes es un indicador de que las posibilidades de crecimiento aplicando otros conceptos del minado de datos pueden resultar muy satisfactorias.

Como línea futura, se propone la extensión de dicha técnica de detección de anomalías no solo a los datos procedentes de los sensores ambientales y de WiFi, sino a todos los sensores presentes en la ciudad inteligente (correlando la información entre ellos) y de esta forma completar el módulo de minería y correlación de datos para que arroje información sobre comportamientos anómalos en todo el sistema.

VII. AGRADECIMIENTOS

Este trabajo ha sido financiado en parte con el apoyo del MINECO español (proyecto DHARMA, Dynamic Heterogeneous Threats Risk Management and Assessment, con código TIN2014-59023-C2-2-R) y por la Comisión Europea (FEDER/ERDF).

REFERENCIAS

[1] Catherine Fisher, “Linkedin Unveils the Top Skills That Can Get You Hired In 2017”, en *https://blog.linkedin.com/2016/10/20/top-skills-2016-week-of-learning-linkedin*, Octubre 2016.

[2] Proyecto DHARMA: http://dharma.inf.um.es/

[3] Pilar Holgado, Víctor A. Villagrá: "Sistema de detección de fases de ataque basado en modelos ocultos de Markov," en *Proceedings of the II Jornadas Nacionales de Investigación en Ciberseguridad*, pp. 25-28, Granada (Spain), 15-17 June 2016.

[4] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu: “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, en *KDD-96*, 1996.

[5] KDD, “2014 SIGKDD Test of Time Award” in *ACM SIFKDD*, August 2014.

[6] Xiaohua Yan, Joy Ying Zhang, “Early Detection of Cyber Security Threats using Structured Behavior Modeling” in ACM Transactions on Information and System Security, January 2013.

[7] Li Xue-yong, Gao Guo-hong, Sun Jia-xia: “A New Intrusion Detection Method Based on Improved DBSCAN”, en *WASE International Conference on Information Engineering*, August 2010.

[8] Pilar Holgado, Víctor A. Villagrá: "Evolving from a static toward a proactive and dynamic risk-based defense strategy” en *Proceedings of the I Jornadas Nacionales de Investigación en Ciberseguridad*, pp. 129-136, León (Spain), 14-16 September (2015).

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6581

Entorno de simulación distribuida de redes basado en la nube computacional

Sergio Serrano Iglesias, Eduardo Gómez Sánchez, Miguel L. Bote Lorenzo,
Juan I. Asensio Pérez, Manuel Rodríguez Cayetano
Departamento de Teoría de la Señal, Comunicaciones e Ingeniería Telemática
Universidad de Valladolid
ETSI de Telecomunicación, Paseo de Belén 15, 47011 Valladolid
sergio@gsic.uva.es, edugom@tel.uva.es, migbot@tel.uva.es, juaase@tel.uva.es, manrod@tel.uva.es

***Resumen*—Las simulaciones de barridos de parámetros tienen un gran potencial en el estudio de redes telemáticas, especialmente en contextos docentes. Sin embargo, el elevado tiempo necesario habitualmente para completar este tipo de simulaciones es una limitación importante para su uso. En este artículo se propone DNSE3, un entorno que permite la ejecución distribuida de tareas de simulación en el simulador ns-3 dentro de un entorno de nube computacional, a través de una arquitectura de servicios RESTful. El sistema se ha diseñado para ser autoescalable, aprovisionando y liberando dinámicamente recursos de la nube computacional en función de la carga de simulaciones demandada, y garantizando un reparto equitativo de los recursos entre los distintos usuarios. Además, DNSE3 se ha implementado reutilizando servicios presentes en *middlewares* de nube populares, y ha sido evaluado mediante pruebas sintéticas. La implementación de DNSE3 ha demostrando un correcto comportamiento funcional y un rendimiento considerablemente superior a otras alternativas cuando el número de simulaciones es muy elevado.**

***Palabras Clave*—Nube computacional, simulación, barrido de parámetros, escalado, entorno distribuido, REST**

I. INTRODUCCIÓN

La simulación es una herramienta de gran ayuda para el estudio de redes telemáticas. Ofrece un mecanismo de evaluación de protocolos y despliegues que no son fácilmente caracterizables mediante modelos analíticos. Del mismo modo, permite estudiar el comportamiento de las redes sin tener que recurrir a instalaciones reales, generalmente costosas [1], [2]. Entre los usos que se dan a estas simulaciones se encuentran la investigación y desarrollo de nuevos estándares y protocolos, así como la caracterización del comportamiento del despliegue de una infraestructura ante la llegada de diferentes patrones de tráfico [2]. Algunos de los simuladores de redes más empleados y conocidos son el ns-2 [3], muy utilizado en investigación; el ns-3 [4], siguiente versión del ns-2; OMNet++; [5] y Riverbed Modeller [6].

La simulación de redes es también una aproximación muy utilizada en entornos académicos. Gracias a las simulaciones, los alumnos pueden reforzar los conocimientos adquiridos en las sesiones teóricas mediante la replicación de los escenarios planteados. Las simulaciones, a su vez, permiten estudiar redes que no se pueden poner fácilmente en marcha con equipamiento real (por falta de recursos o potenciales problemas de seguridad); o, incluso, estudiar redes que sí están disponibles, pero en situaciones que son difíciles de reproducir en la realidad.

Para el análisis de redes es muy habitual recurrir a las denominadas simulaciones de barrido de parámetros, en las cuales algunos de sus parámetros toman valores dentro de un rango definido por el usuario para ver cómo se ve afectado el modelo de simulación. Entre los usos que se dan a este tipo de simulaciones se incluye la revisión del desempeño de la red en función de los valores de los distintos factores que entran en juego en la comunicación (velocidades binarias de enlaces, tamaños de ventana, etc.) o la optimización de dichos factores para alcanzar determinados requisitos deseables de calidad de servicio.

Normalmente la ejecución de un barrido de parámetros implica la realización de múltiples simulaciones individuales, tantas como número de combinaciones diferentes de valores de parámetros sean posibles dentro de los rangos definidos por el usuario. En un ordenador convencional, estas simulaciones se ejecutarán secuencialmente. Aunque el tiempo de ejecución de una única simulación puede ser relativamente corto, en una de barrido de parámetros éste se ve incrementado por el gran volumen de procesos necesarios para completarla, lo que alarga la espera hasta disponer de los resultados finales. Dentro del contexto educativo, este retardo puede afectar negativamente al desarrollo de las prácticas de laboratorio, que disponen de un tiempo limitado en sus sesiones de trabajo.

Dado que cada una de las simulaciones en las que se descompone un barrido de parámetros constituye una tarea individual e independiente del resto, se puede re-

ducir el tiempo requerido para su finalización mediante la ejecución de varias simulaciones en paralelo, ya sea explotando las capacidades de computación paralela de un ordenador (múltiples núcleos) o de un sistema distribuido (múltiples ordenadores) [7]. En la computación paralela en un ordenador se reparten los diferentes trabajos en hilos y procesos de ejecución diferentes, con el fin de sacar el máximo rendimiento de los recursos de los que dispone la máquina. Algunos de los simuladores de redes, como el ns-3 u OMNet++, incorporan esta técnica para la ejecución de las simulaciones. Aunque se pueden conseguir mejoras en el rendimiento, la escalabilidad de esta solución se ve limitada por el número de núcleos o procesadores instalados en el único ordenador donde se realiza la simulación [7].

Por su parte, en la computación distribuida se utilizan múltiples ordenadores para repartir los diferentes trabajos a realizar. Así, la escalabilidad no está limitada por los recursos de computación de un ordenador, ya que pueden incorporarse ordenadores adicionales. Por el contrario, los límites a la mejora del rendimiento los define la capacidad de la red y la necesidad de sincronizar los diferentes procesos, en el caso de que haya dependencias [7]. Como se ha mencionado, las simulaciones resultantes de un barrido son totalmente independientes, no existiendo este coste de sincronización. Por ello, si una simulación se completa en un tiempo T, N simulaciones podrían completarse en N ordenadores en un tiempo marginalmente superior a T. Sin embargo, y hasta donde sabemos, los simuladores de redes telemáticas existentes no hacen uso de esta técnica.

Dada esta problemática, los autores de este artículo desarrollaron el DNSE (*Distributed Network Simulation Environment*) [8], un entorno que permitía la ejecución de simulaciones de barrido del simulador ns-2 dentro de una infraestructura de *grid* computacional. Este entorno tuvo una buena acogida entre los profesores y alumnos de la E.T.S. de Ingenieros de Telecomunicación de la Universidad de Valladolid. A pesar del buen funcionamiento que mostraba, el sistema presentaba una serie de problemas ligados a la propia infraestructura de *grid* en la que funcionaba. El escalado del sistema se tenía que administrar manualmente, lo que propiciaba situaciones en las que o bien el entorno disponía de recursos aprovisionados sin utilizarse, o bien no se alcanzaban tiempos de respuesta suficientemente bajos que mejorasen la calidad de servicio. Otro problema importante era la considerable dificultad técnica y administrativa que suponía la puesta en marcha de un *grid* en el que participaran múltiples organizaciones administrativas con el objetivo de asegurar la cantidad de recursos necesarios para soportar el uso del DNSE a gran escala. Un problema adicional que existía en la época en la que se estuvo utilizando fue la falta de madurez que ofrecían los *middlewares* disponibles para el desarrollo de aplicaciones orientadas al grid.

En los últimos años ha surgido un paradigma que consigue solucionar algunos de estos problemas: la nube computacional. La nube computacional ha cobrado gran popularidad con la oferta de nubes públicas, entre las que se encuentran Google App Engine [9], Microsoft Azure [10] o Amazon Web Services (AWS) [11]. Las denominadas nubes de infraestructura como servicio (*Infrastructure as a Service*)[12] presentan una serie de características que cubren las carencias del *grid*, como son: el aprovisionamiento de recursos bajo demanda y su gestión remota, lo que permite un despliegue rápido de máquinas virtuales y simplifica su administración; y la monitorización de los servicios y recursos empleados y el escalado rápido de éstos [12]. Estas dos últimas características, junto con la capacidad de ofrecer recursos aparentemente infinitos [12], permiten realizar un seguimiento del sistema y ajustar el número de máquinas desplegadas a las necesidades computacionales. Además, gracias al interés comercial generado por este paradigma [12], el *middleware* disponible es mucho más estable y robusto, existiendo varias alternativas, tanto en la nube pública como para desplegar nubes privadas, con interfaces de servicios muchas veces compatibles.

En este artículo se presenta el DNSE3 (*Distributed Network Simulation Environment 3*), un rediseño importante inspirado en el sistema anterior, que introduce cambios significativos: por un lado, cambia la infraestructura del *grid* por la nube computacional gracias a las ventajas que aporta; por otro, aprovecha la evolución del simulador ns-2 a ns-3, con sus correspondientes mejoras de rendimiento [2]; finalmente, introduce funcionalidades detectadas como necesarias en el estudio anterior.

A lo largo de este artículo se presentarán los siguientes apartados. En primer lugar se revisarán los trabajos que es posible encontrar en la literatura con propuestas relacionadas con la realización de simulaciones de forma distribuida. A continuación, se tratará el diseño del DNSE3, discutiendo los principales requisitos y la propuesta de arquitectura resultante. La siguiente sección estará enfocada en los principales problemas que debe resolver el DNSE3 y la forma en la que se han tratado. Finalmente se discutirá sobre las conclusiones y el trabajo futuro.

II. ESTADO DEL ARTE

En la literatura se pueden encontrar diferentes proyectos que han permitido el uso distribuido de los simuladores de diferentes campos de estudio. Uno de ellos es el DNSE [8], que, según se mencionó en la sección anterior, distribuía las simulaciones de barrido de parámetros de redes telemáticas para el simulador ns-2 dentro de un *grid* computacional. Aparte de mejorar el tiempo de respuesta de las simulaciones, ofrecía una GUI (*Graphic User Interface*) para facilitar tanto la gestión de los trabajos de simulación como el visionado de las animaciones generadas a partir de la herramienta NetAnim, ofrecida de forma conjunta con el ns-2.

Otro ejemplo de despliegue en el *grid* fue DSoG (*Distributed Simulation on Grid*) [13], que facilita el estudio y modelado de motores de aviación. Esta propuesta, a diferencia de las mencionadas en esta sección, no pretende optimizar la ejecución de las simulaciones. En su lugar,

ofrece un entorno colaborativo en el que sus usuarios puedan compartir sus datos y modelos, de tal forma que se dispone de una amplia biblioteca con multitud de recursos con los que generar los modelos de simulación finales.

Si nos centramos ahora en aquellos proyectos que hagan uso de la nube, se pueden encontrar dos entornos orientados a la simulación de tráfico automovilístico: SEMSim CS (*Scalable Electro-Mobility Simulation Cloud Service*) [14], usado principalmente para el estudio del escenario en el que se reemplazasen todos los vehículos actualmente existentes por una alternativa eléctrica, y C^2SuMo (*Cloud-based, Collaborative ans Scale-up Modelling and Simulation Framework for STEM Education*) [15], un entorno educativo para estudiantes de educación superior con el que estudiar simulaciones ambientadas en el mundo real. En ambas propuestas se aprovecha la flexibilidad de la nube para mejorar el tiempo de respuesta de las simulaciones.

Aunque existen otros simuladores que hacen uso de la nube computacional, no existen propuestas que empleen la nube computacional enfocadas al estudio de las redes telemáticas y, los que sí están enfocados, utilizan otro tipo de infraestructuras. Con la propuesta de este artículo se espera cubrir esta carencia.

III. DISEÑO DEL DNSE3

En esta sección se cubre el diseño del DNSE3. En primer lugar se presentan los requisitos más importantes que debe cumplir para utilizarse en un ambiente multiusuario con acceso concurrente, para, finalmente, presentar la arquitectura empleada.

A. *Requisitos del DNSE3*

El DNSE3 es un entorno de simulación distribuida de redes para ser usado en contextos académicos. Debe satisfacer un conjunto de funcionalidades básicas para que los alumnos gestionen simulaciones de manera sencilla, y debe funcionar de manera robusta y eficiente, para ser un ayuda y no un impedimento al aprendizaje. A continuación se detallan brevemente los requisitos impuestos en el diseño del DNSE3.

En cuanto a los **requisitos funcionales**, el sistema debe:

- **Gestionar los proyectos de simulación**. Los usuarios deben ser capaces de crear proyectos de simulación, en los que se definen el modelo a emplear en las diferentes simulaciones, los parámetros soportados por el modelo y los resultados que se generan tras su ejecución. Toda esta información es necesaria para que el sistema pueda ejecutar adecuadamente las simulaciones y se puedan recuperar los resultados deseados.
- **Ejecutar simulaciones de barrido de parámetros**. Además de la ejecución de simulaciones individuales, se debe permitir a los usuarios la libre configuración de los parámetros y la selección de los resultados para el barrido.
- **Controlar la ejecución de las simulaciones**. El usuario podrá elegir, en todo momento, si quiere iniciar la ejecución de una simulación previamente creada o detenerla, en caso de que quiera modificar alguno de sus parámetros o ficheros recolectados. Los usuarios pueden confundirse en los valores proporcionados y deben ser capaces de alterar dichos valores sin tener que esperar a que se termine la ejecución completa y descartar los resultados generados. Por otro lado, en caso de detectar un fallo en la ejecución de las simulaciones, se deberá notificar al usuario, detener la ejecución de todas las tareas asociadas y esperar a que el usuario lo resuelva antes de continuar con su ejecución.
- **Recuperar los resultados**. Tras la ejecución de las simulaciones, los ficheros de resultados se deben almacenar de forma persistente, incluso después de que el usuario haya recibido una copia. En el caso de los resultados de los barridos, el sistema debe aclarar al usuario qué resultado se corresponde con qué combinación de parámetros.

Por otra parte, existen una serie de **requisitos no funcionales** que condicionan el diseño y las decisiones tecnológicas acerca del DNSE3 y que se enumerarán a continuación.

- **Escalabilidad del sistema**. El tiempo requerido para la ejecución de las simulaciones de barrido no deberá crecer de forma proporcional al número de simulaciones derivadas de estas. Por este motivo, será necesario utilizar múltiples máquinas virtuales que ejecutarán las simulaciones en paralelo. Para conseguir optimizar tanto los tiempos de respuesta como los recursos utilizados, el sistema debe ser autoescalable. Para ello, deberá contar con una cantidad de recursos suficiente y ser capaz de determinar el número de recursos necesarios para las simulaciones solicitadas, ya sea aprovisionado o liberándolos, con el menor impacto posible en el resto de funciones.
- **Reparto de recursos**. El sistema será empleado por diferentes alumnos, y se debe garantizar un reparto equitativo de los recursos disponibles entre todos ellos. Será necesario incorporar mecanismos que impidan que un alumno (de manera intencionada o no) acapare todos los recursos del sistema e impida el progreso de las tareas del resto de usuarios.
- **Tolerancia a fallos**. El sistema debe reaccionar a los fallos en la infraestructura (se pierde alguna máquina virtual) o de alguno de sus servicios (se bloquea un proceso de simulación) ocultando estos eventos al usuario final y replanificando las tareas, de manera que en todo caso la degradación sea paulatina (*graceful degradation*).
- **Clientes de acceso**. Se debe disponer de clientes, tanto en línea de comandos (orientados a pruebas o usuarios avanzados, y de los que no se hablará en este artículo) como a través de una interfaz web, que permitan el control remoto de las operaciones del sistema.

B. Arquitectura del sistema

El DNSE3 presenta una arquitectura orientada a servicios (SOA, *Service Oriented Architecture*) que propone la separación de las diferentes funciones de un sistema en diferentes servicios con una interfaz de acceso expuesta a cualquier clase de cliente. Con este tipo de arquitectura obtendremos un sistema más sencillo de construir, mantener y escalar [16]. Para las aplicaciones desplegadas en una nube de computación, este tipo de arquitectura ofrece ventajas en su administración. Una vez configurados los servicios que funcionarán en cada una de las instancias, se puede generar una instantánea que mantenga su configuración en ese momento y, cuando se produzca un fallo en alguna de las máquinas, se puede crear una nueva máquina a partir de esa instantánea. De igual manera, cuando sea necesario replicar servicios para conseguir escalar hacia arriba, sólo será necesario levantar nuevas máquinas virtuales a partir de estas imágenes.

Cada uno de los servicios que forman parte de la aplicación DNSE3 ha sido diseñado siguiendo el estilo arquitectural REST (*REpresentational State Transfer*), que propugna una arquitectura orientada a recursos (ROA, *Resource Oriented Architecture*) y un conjunto de interfaces homogéneo. De las diferentes características de REST, como son el uso de una interfaz uniforme o la navegación a través de sus recursos, la que más nos interesa ahora es la carencia de estado de peticiones previas o *statelessness*. Los servicios RESTful, según este principio, no mantendrán información relacionada con ninguna petición previa de cualquiera de sus clientes, sino que estos deberán mantener esa información. Esto facilita el equilibrado de carga, al poder dirigir las peticiones a diferentes réplicas de un servicio y, por ello, la tolerancia a fallos y la escalabilidad.

La propuesta final de la arquitectura del DNSE3 se compone de un total de siete servicios, mostrados en la figura 1, entre los que nos podemos encontrar el *servicio de orquestación*, encargado de recibir las peticiones de los usuarios y coordinar al resto del sistema mediante el envío de tareas de simulación al servicio de; el *servicio de colas*, que gestiona el reparto de tareas de simulación; el *servicio de simulación*, capaz de ejecutar las simulaciones de los usuarios; el *servicio de informe*, que se encarga de preparar los resultados de las simulaciones para que puedan ser utilizados posteriormente por los usuarios; el *servicio de almacenamiento*, que ofrece un almacenamiento persistente compartido por el resto de servicios; el *servicio de monitorización*, que recolecta métricas que muestran el uso que se hace del sistema, y el *servicio de escalado*, capaz de ajustar el número de instancias de simulación a la demanda de trabajos de los usuarios.

La propuesta final de la arquitectura del DNSE3 se compone de un total de siete servicios, mostrados en la figura 1, entre los que nos podemos encontrar el *servicio de orquestación*, encargado de recibir las peticiones de los usuarios y coordinar al resto del sistema mediante el envío de tareas de simulación al servicio de; el *servicio de colas*, que gestiona el reparto de tareas de simulación; el *servicio*

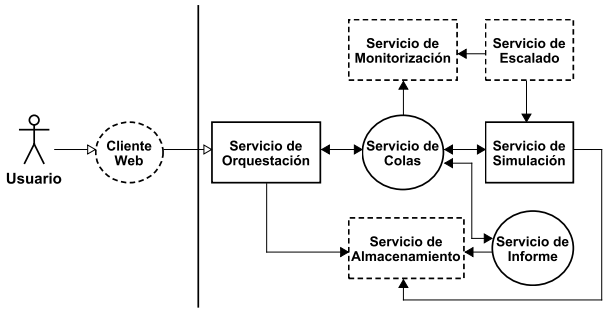


Figura 1. Servicios que componen el DNSE3 y las interacciones que establecen.

de simulación, capaz de ejecutar las simulaciones de los usuarios; el *servicio de informe*, que se encarga de preparar los resultados de las simulaciones para que puedan ser utilizados posteriormente por los usuarios; el *servicio de almacenamiento*, que ofrece un almacenamiento persistente compartido por el resto de servicios; el *servicio de monitorización*, que recolecta métricas que muestran el uso que se hace del sistema, y el *servicio de escalado*. Este último servicio es de vital importancia para el correcto desempeño de la aplicación, al ser el encargado de ajustar el número de instancias de simulación a la demanda de trabajos de los usuarios mediante la definición de una serie de reglas basadas en las métricas del servicio de monitorización.

Cada uno de los servicios previamente mencionados presentan diferentes grados de acoplamiento con la aplicación, lo que facilita su agrupación basada en la facilidad que presentan para ser integrados en otros proyectos diferentes. De esta forma, los servicios de orquestación y simulación presentan las funciones más ligadas a la aplicación (el primero gestiona el flujo de esta aplicación, el segundo envuelve e invoca al ns-3) y ofrecen mayor dificultad para ser utilizado en otros ámbitos. Por otro lado, los servicios de colas e informe presentan funciones más genéricas y fácilmente extensibles (podrían servir para, con facilidad, desplegar otro sistema paralelo y escalable y en otras infraestructuras diferentes de la nube). Finalmente, es interesante observar que los servicios de almacenamiento, monitorización y escalado son de uso habitual en diversas aplicaciones desplegadas en esta clase de entornos y son ofrecidos de forma nativa por muchos *middlewares* de nube, por lo que, a diferencia de los anteriormente mencionados, no necesitan ser desarrollados *ad-hoc* sino que se accederá a su API pública. Esta distinción de los servicios se ve reflejada en la figura 1, perteneciendo los servicios representados con un rectángulo con trazo sólido, un círculo con trazo sólido y un rectángulo con trazo discontinuo, respectivamente, a cada una de las tres agrupaciones mencionadas.

IV. IMPLEMENTACIÓN DE LOS SERVICIOS

A. *Aspectos tecnológicos*

Los entornos de nube computacional presentan una gran heterogeneidad. Por una parte, se presentan diferentes modelos de nube en función de la abstracción que nos ofrezcan de su infraestructura, como son IaaS, PaaS (*Platform as a Service*) o CaaS (*Container as a Service*), entre otros. Por otra parte, tanto las instancias desplegadas como las máquinas anfitrionas no están sujetas a un tipo de arquitectura o sistema operativo concreto. De entre los diferentes modelos de nube, se ha optado por un diseño orientado a IaaS para el DNSE3, al ofrecer un mayor nivel de gestión y configuración requeridos por algunos servicios, como sucede con el de simulación.

El despliegue se ha realizado en una nube privada que utilizaba OpenStack [18] como sistema gestor de la infraestructura. Este *middleware*, muy utilizado en nubes privadas, ofrece un entorno IaaS compatible con AWS, con el que se ofrece la posibilidad de extender a una nube híbrida, orientada a cubrir picos de demanda cuando los recursos de la nube privada no sean suficientes. Entre los servicios de OpenStack destacan Swift y Cinder, para el almacenamiento distribuido basado en objetos y volúmenes, respectivamente; Ceilometer, para la publicación de métricas; y Heat que, junto a Ceilometer, permiten gestionar las políticas de escalado del sistema. Todos estos servicios han sido integrados dentro del DNSE3 por medio de la invocación a su API REST.

Dada la libre configuración de entornos en la nube, se ha decidido desarrollar los diferentes servicios del DNSE3 en Java, al tratarse de un lenguaje multiplataforma muy popular que facilita la libre elección del entorno empleado. Además, existen diferentes *APIs* (*Application Programming Interface*) para el despliegue de servicios RESTful en Java. En el DNSE3 se ha empleado la API Restlet [17], que permite el desarrollo tanto de servidores RESTful como clientes REST, está disponible para diferentes plataformas (Java SE, Java EE, OSGI...) y ofrece un gran abanico de funcionalidades adicionales como son las implementaciones de seguridad para la integración de otros servicios.

Otro aspecto importante en una arquitectura SOA es la representación de la información intercambiada en las interfaces de servicio. Para reducir el volumen del tráfico cursado, se utilizarán documentos JSON (*JavaScript Object Notation*) para el intercambio de mensajes. Este formato de documento es mucho más liviano que otro tipo de lenguajes de marcado, como sucede en XML (*eXtensible Markup Language*). Esta simplicidad y ligereza conlleva la pérdida de la validación de la estructura del documento (que sí está disponible en XML). Este problema no supone un impacto mayor gracias a la disposición de librerías como GSON o Jackson, que permiten la conversión de los datos contenidos en el documento a objetos de clases conocidas o desarrolladas.

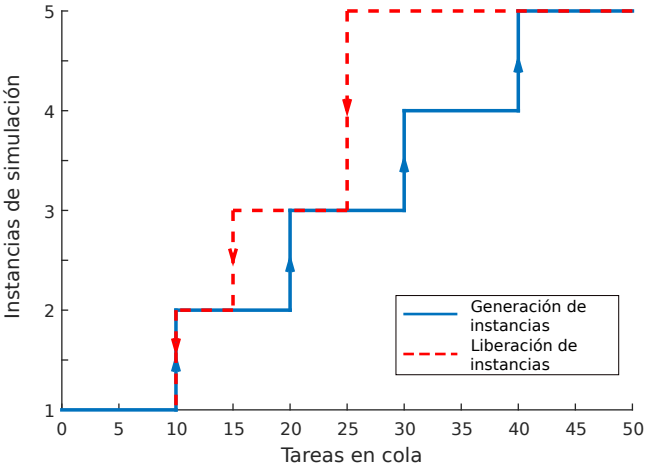


Figura 2. Representación de cómo las reglas de escalado definidas van haciendo aumentar el número de instancias disponibles según aumentan los trabajos pendientes en la cola de simulación (línea azul, progreso de izquierda a derecha), y de cómo esas mismas reglas van eliminando instancias al reducirse el número de trabajos pendientes (línea roja, progreso de derecha a izquierda).

B. *Escalado del sistema*

Esta función es la más importante del DNSE3 y la que aporta mayor valor añadido respecto a la versión anterior. El sistema debe ser capaz de acomodar el número de recursos desplegados a la demanda de trabajos de simulación solicitada por los usuarios. Para que este ajuste sea lo más eficiente posible, se debe lograr, por un lado, evitar continuas fluctuaciones en el número de instancias utilizadas y, por otro, mantener la mayor ocupación posible.

El tiempo de arranque de las máquinas no es despreciable y podría suceder que, una vez se haya lanzado una nueva máquina, se decida eliminar otra de las desplegadas al haberse reducido la demanda en ese tiempo de carga. Esta situación supondría un desperdicio de recursos de la máquina anfitriona, con un impacto negativo en el rendimiento.

Así, en lugar de eliminar las instancias tan pronto como se reduzca la carga, interesa que las instancias sólo sean eliminadas cuando la carga se reduzca considerablemente. Este sobreaprovisionamiento temporal permitirá mejorar el tiempo de respuesta y reaccionar rápidamente a nuevas subidas de carga y, en cualquier caso, desaparecerá cuando el número de tareas descienda de manera importante.

La política de escalado empleada en el DNSE3 para determinar cuándo y cómo se efectúa este escalado y que sea capaz de cumplir los requisitos anteriores se basa en el número de trabajos que debe efectuar cada una de las instancias de simulación, tal y como se ilustra en la figura 2. Hasta que esta proporción no supera un determinado umbral (en nuestro caso 10 simulaciones por instancia), no se aprovisionarán nuevos recursos de simulación. En el caso contrario, mientras no se reduzca la carga de cada instancia al 50 % del umbral anterior (5 simulaciones por instancia), se mantendrá el número de recursos desplegados. Una vez se reduzca, se liberan la

mitad de las instancias desplegadas.

Una ventaja de esta política es su sencillez de implementación, Los servicios de escalado de los entornos de nube permiten incluir políticas basadas en reglas *IF-THEN*, en las que se comparan los valores almacenados en el servicio de monitorización con un umbral previamente fijado, o bien la agrupación de éstas para generar reglas más complejas. En nuestro caso sólo se necesitan 2 reglas de comparación: si $\frac{tareas}{instancias} > 10$, entonces $instancias++$; y, si $\frac{tareas}{instancias} < 5$, entonces $instancias = 0{,}5 \times instancias$, redondeando al entero mayor. En las reglas anteriores el término $tarea$ agrupa tanto a las simulaciones pendientes de ejecutar como las que están siendo procesadas por las diferentes instancias de simulación. Para que estas métricas puedan aplicarse, el servicio de colas deberá enviar a Ceilometer una métrica que contenga el reparto equitativo de las tareas solicitadas (dato conocido al contener todo el listado de trabajos) entre las instancias de simulación (que deberá de solicitar de forma periódica a Heat).

Cuando la carga de trabajo decrece a los niveles indicados por la política de escalado, se procede a la eliminación de las instancias de simulación sobrantes. En el caso de que estas máquinas estén ejecutando alguna simulación, se intentará esperar a que ésta se complete sin solicitar un nuevo trabajo antes de eliminarla, aunque puede darse la situación de que se interrumpa dicho trabajo. En este último escenario, los trabajos asociados volverían a estar disponibles una vez haya expirado su tiempo de reserva al no recibir actualizaciones del servicio de simulación.

C. *Asignación de trabajos de simulación*

Un aspecto fuertemente ligado al escalado es el lugar en el que se realiza la planificación de recursos. Si el servicio de colas actuase de *scheduler* centralizado, debería conocer qué servicios de simulación están disponibles, cuáles se están apagando, si se están levantando nuevos, y sus URI de acceso. Como se puede prever, esto supone un aumento de complejidad en el servicio.

En su lugar, se ha seguido un modelo de asignación *work-stealling*, en el que, en lugar de asignar los trabajos a los diferentes trabajadores, son estos los que solicitan el trabajo que deben realizar. Cuando se inicia una nueva instancia, ésta conoce de antemano la dirección de acceso al servicio de colas para poder preguntar por el siguiente trabajo a realizar. El servicio de simulación está siempre intentando trabajar, así que o está ocupado, o demanda nuevos trabajos al servicio de colas. Además, periódicamente informa al servicio de colas de su actividad, de manera que cuando deja de comunicarse con él, el servicio de colas puede asumir que ha caído. Como en los servicios RESTful no se mantiene información de las conexiones previas, este diseño permite que el servicio de colas sólo se tenga que preocupar de mantener la cola de trabajos y determinar el trabajo siguiente a procesar, tareas que no dependen del número de instancias de simulación que haya, lo que simplifica enormemente la escalabilidad.

En la determinación de la siguiente tarea a ejecutar se realiza una rotación (*round-robin*) de dos niveles entre los trabajos pendientes, ilustrada en la figura 3, con el fin de evitar tanto el acaparamiento del sistema por parte de un usuario (de forma intencionada o no) como la recolección de un mismo trabajo por parte de dos instancias diferentes. El primer nivel de rotación va alternando de usuario propietario de los trabajos, de forma que dos trabajos solicitados consecutivos no pueden pertenecer a un mismo usuario (salvo que no haya trabajos pendientes de otro usuario), mientras que el segundo nivel recorre los trabajos pendientes de ejecutar del usuario escogido en el nivel anterior. Dentro de los trabajos pendientes se incluyen aquellos no asignados y los asignados anteriormente a un servicio que ha caído (que ha dejado de reportar su actividad).

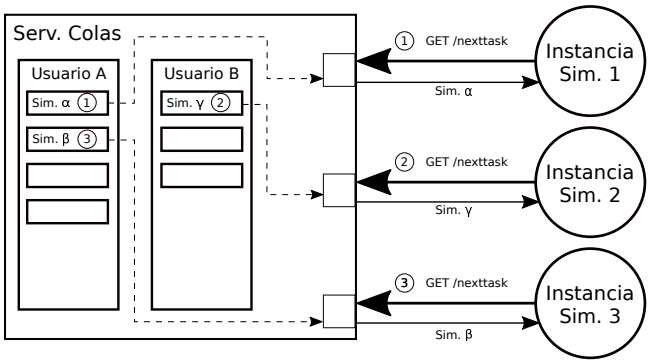


Figura 3. Ilustración de la asignación del trabajo siguiente mediante la rotación *round-robin* de los usuarios propietarios de los trabajos.

V. INTERFAZ DE USUARIO

Aunque los servicios RESTful, gracias a su interfaz uniforme, que en el caso de los servicios web se corresponde con los métodos de del protocolo HTTP (*Hyper-Text Transfer Protocol*), permiten usar cualquier cliente web (por ejemplo, un navegador) para ser invocados, es complejo que el usuario final prepare y consuma los cuerpos de peticiones y respuestas. Es por ello que se ha optado por desarrollar un cliente del DNSE3 que es una aplicación web con una interfaz de usuario ilustrada en la figura 4. Este cliente se encarga de recibir y procesar las peticiones de los usuarios finales, adecuándolas a la interfaz REST del servicio de orquestación y de reproducir las respuestas recibidas en representaciones HTML visualmente agradables para devolver al navegador. Para que esta aplicación sea usable desde cualquier navegador web y tipo de dispositivo, se ha desarrollado usando el *framework* Bootstrap [19], que ofrece soporte con los dispositivos móviles como *smartphones* o tabletas, muy populares entre los alumnos en los últimos años.

Adicionalmente, la seguridad se puede gestionar de manera delegada: el servicio de orquestación acepta todas las peticiones provenientes del cliente web, mientras que éste es el encargado de validar las credenciales de los usuarios. Como no se aceptan en el servicio de orquestación peticiones de otros orígenes, resulta más sencillo proteger al DNSE3 de ataques externos.

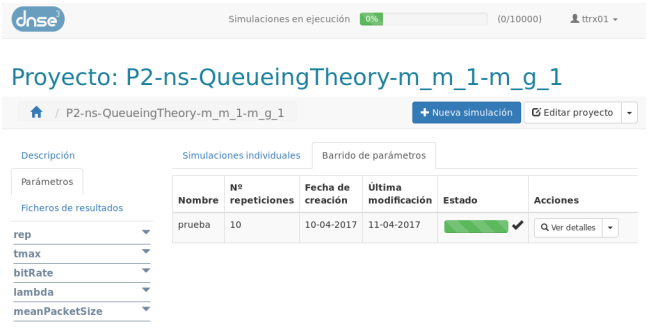


Figura 4. Intrefaz web del DNSE3

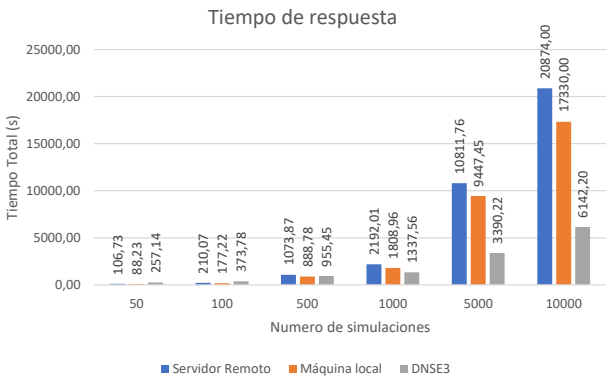


Figura 5. Tiempos de respuesta ofrecidos por cada uno de los sistemas evaluados ante diferentes tamaños de simulaciones a ejecutar

VI. EVALUACIÓN DEL RENDIMIENTO

Una vez se ha completado el desarrollo de una versión de pruebas del sistema, se ha evaluado si podría ser desplegada en un entorno de producción. Las pruebas realizadas incluyen el acceso concurrente de varios usuarios, que no debe comprometer la calidad de servicio (QoS, *Quality of Service*) del resto de usuarios; la publicación de simulaciones; la recolección de los resultados; y el reparto de los recursos según se comentó en la sección C.

A falta de una evaluación con usuarios reales, en lo que respecta a temas de rendimiento, se han efectuado una serie de pruebas sintéticas para evaluar el escalado automático de las instancias de simulación. Estas pruebas han consistido en la ejecución de un barrido de parámetros sintético consistente en varias simulaciones de duración más o menos semejante. Estas pruebas se han repetido en tres sistemas: el servidor del laboratorio de la asignatura (ordenador con procesador Intel(R) Xeon(R) E5-2620 de 4 núcleos a 2 GHz y 8GB de memoria RAM), que los alumnos podrían usar fuera del horario de prácticas; los ordenadores del laboratorio (ordenadores con procesador Intel(R) Core(TM) i5-2400 de 4 núcleos a 3,10 GHz y 4 GB de memoria RAM), sólo accesibles en horario lectivo; y el DNSE3. En los dos primeros casos se han buscado momentos en los que no había otros usuarios corriendo procesos en el sistema. En el caso del DNSE3 se partía de una situación de reposo (un único servicio de simulación creado) en cada experimento, y el servicio de escalado podría llegar a crear hasta un máximo de 15 instancias simultáneas de servicios de simulación.

La figura 5 muestra los tiempos requeridos para la ejecución completa de todas las simulaciones, para barridos de parámetros sintéticos consistentes en un distinto número de trabajos. El tiempo típico para completar una simulación individual ha rondado los 2-3 segundos tanto en las máquinas locales como en el servidor del laboratorio, mientras que en el DNSE3 este tiempo se incrementaba hasta los 7 segundos. Este incremento se debe a las comunicaciones realizadas entre los diferentes servicios: la publicación de la tarea en el servicio de colas, la recogida de la tarea por parte del servicio de simulación, la obtención del modelo de simulación almacenado en el servicio de almacenamiento y la posterior carga de los resultados y la notificación del estado de ejecución al servicio de orquestación.

En vista a los resultados obtenidos, se pueden distinguir dos situaciones diferentes. En primer lugar, si el número de simulaciones a ejecutar es bajo, en barridos que den lugar hasta unas 500 simulaciones, el DNSE3 no presenta una ventaja comparativa: por una parte, el número de simulaciones es bastante bajo y puede completarse secuencialmente en un tiempo pequeño. Aunque el servicio de escalado aprovisiona nuevas instancias, estas tardan en arrancar unos 75 segundos, por lo que llegan a asumir pocas tareas, obteniéndose un beneficio muy limitado de la paralelización. Por otra parte, a medida que aumenta el número de simulaciones totales a ejecutar, las nuevas instancias levantadas por el servicio de escalado llegan con todavía muchos servicios pendientes de procesar en la cola, así que en un momento dado se están realizando un gran número de ejecuciones en paralelo, compensando sobradamente la sobrecarga de la distribución, llegándose a alcanzar una reducción del 65 % del tiempo de respuesta con barridos de 5000 simulaciones.

Aún así, como en el entorno de pruebas se ha restringido el número de máximo de servicios de simulación, a partir de un punto el tiempo de respuesta se incrementa de forma lineal. Para evitar esta situación, se puede hacer uso de una nube híbrida y así conseguir recursos virtualmente infinitos.

VII. CONCLUSIONES Y TRABAJO FUTURO

El uso de los simuladores de redes telemáticas es beneficioso para la investigación y análisis de despliegues de infraestructura y protocolos de comunicaciones. Aun así, las simulaciones de barrido de parámetros tienen un coste computacional excesivo, lo que limita su utilización especialmente en entornos educativos, con recursos computaciones y tiempo escasos.

Para tratar este problema, este artículo ha propuesto el DNSE3, una aplicación distribuida orientada a servicios que, aprovechando las ventajas de la nube computacional, introduce mecanismos de autoescalado para aprovisionar recursos dinámicamente en función del número de simulaciones demandadas por los alumnos. Su diseño también ha tenido en cuenta otros aspectos relevantes, como el reparto

equitativo de los recursos entre usuarios y la tolerancia a fallos.

Para ilustrar su funcionamiento, se ha programado un prototipo de la aplicación con una interfaz de usuario web. Posteriormente, se ha evaluado el rendimiento del escalado sometiendo al DNSE3 a barridos de simulación sintéticos de distintos tamaños, observándose que cuando el número de simulaciones individuales es bastante elevado el tiempo de respuesta obtenido por el DNSE3 es hasta un 65 % menor que el alcanzado con las otras alternativas disponibles en la ETSIT de Valladolid. Además, se ha observado que el comportamiento era funcionalmente correcto, y que se lanzaban y detenían servicios de simulación en función del número de tareas en la cola, siguiendo las reglas de escalado definidas.

Aun así, las pruebas realizadas hasta ahora se han hecho en situaciones controladas por los autores de este trabajo. Es, por lo tanto, necesario repetir este estudio con usuarios reales (alumnos) en un contexto real (muchos alumnos compitiendo simultáneamente por los recursos). Este nuevo estudio permitirá validar los resultados obtenidos anteriormente y, además, recoger el patrón de peticiones de simulación llevado a cabo por los usuarios reales, lo que a su vez facilitará la propuesta de nuevos experimentos sintéticos más realistas en el futuro. También se aprovecharán estas pruebas para conocer la opinión de los usuarios acerca de aspectos de usabilidad e interfaz de usuario, así como para indagar en el impacto que el uso de esta aplicación tiene en sus procesos de aprendizaje.

Otra vía de trabajo futuro se refiere a la exploración de alternativas a las políticas de escalado descritas en la sección B, para estudiar el coste/beneficio de utilizar políticas más agresivas o el mantenimiento de un número mínimo de instancias de simulación permanentemente disponibles.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por los proyecto TIN2014-53199-C3-2-R del Ministerio de Economía y Competitividad, y VA082U16 de la Junta de Castilla y León, con cofinanciación FEDER. Los autores agradecen al resto del grupo de investigación GSIC/EMIC por su apoyo e interés en el proyecto.

REFERENCIAS

[1] A. M. Law and W. D. Kelton, *Simulation modeling and analysis*. McGraw-Hill, 1991.

[2] E. Weingartner, H. vom Lehn, and K. Wehrle, "A Performance Comparison of Recent Network Simulators," in *2009 IEEE International Conference on Communications*, June 2009, pp. 1–5.

[3] "The network simulator - ns-2," 2017. [Online]. Available: http://www.isi.edu/nsnam/ns/

[4] "The network simulator - ns-3," 2015. [Online]. Available: https://www.nsnam.org/

[5] "Omnet++ discrete event simulator," 2017. [Online]. Available: https://omnetpp.org/

[6] *Riverbed Modeler*, Riverbed Technology, 2015. [Online]. Available: https://www.riverbed.com/es/products/steelcentral/steelcentral-riverbed-modeler.html

[7] R. M. Fujimoto, "Research Challenges in Parallel and Distributed Simulation," *ACM Transactions on Modeling and Computer Simulation*, 2016.

[8] M. L. Bote-Lorenzo, J. I. Asensio-Pérez, E. Gómez-Sánchez, G. Vega-Gorgojo, and C. Alario-Hoyos, "A grid service-based Distributed Network Simulation Environment for computer networks education," *Computer Applications in Engineering Education*, 2010.

[9] *Google App Engine*, Alphabet Inc., 2017. [Online]. Available: https://cloud.google.com/appengine/?hl=es

[10] *Microsoft Azure*, Microsoft Corporation, 2017. [Online]. Available: https://azure.microsoft.com/es-es/

[11] *Amazon Web Servicies*, Amazon Inc., 2017. [Online]. Available: https://aws.amazon.com/es/

[12] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," *Communications of the ACM*, 2010.

[13] Y. Cao, X. Jin, and Z. Li, "A distributed simulation system and its application," *Simulation Modelling Practice and Theory*, 2007.

[14] D. Zehe, A. Knoll, W. Cai, and H. Aydt, "SEMSim Cloud Service: Large-scale urban systems simulation in the cloud," *Simulation Modelling Practice and Theory*, 2015.

[15] F. Caglar, S. Shekhar, A. Gokhale, S. Basu, T. Rafi, J. Kinnebrew, and G. Biswas, "Cloud-hosted simulation-as-a-service for high school STEM education," *Simulation Modelling Practice and Theory*, 2015.

[16] S. Vinoski, "REST Eye for the SOA Guy," *IEEE Internet Computing*, 2007.

[17] *Restlet Framework*, Restlet, Inc., 2017. [Online]. Available: https://restlet.com/open-source/

[18] *OpenStack*, OpenStack Foundation, 2017. [Online]. Available: https://www.openstack.org/

[19] "Bootstrap," 2017. [Online]. Available: http://getbootstrap.com/

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6618

Marco para el Análisis e Inferencia de Conocimiento en Redes 5G

Marco Antonio Sotelo Monge, Jorge Maestre Vidal, Luis Javier García Villalba
Grupo de Análisis, Seguridad y Sistemas (GASS)
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, España
masotelo@ucm.es, jmaestre@ucm.es, javiergv@fdi.ucm.es

***Resumen*—Este trabajo propone un marco de análisis de información para la inferencia de conocimiento en el contexto del proyecto SELFNET, cuya finalidad es diagnosticar el estado de la red y predecir problemas potenciales que afecten la operatividad de la red, facilitándose también el proceso de toma de decisiones en redes 5G. Esta propuesta proporciona capacidades para el descubrimiento de hechos, reconocimiento de patrones, razonamiento y predicción con el objetivo de inferir conductas sospechosas que puedan ser mitigadas a través del despliegue de medidas de respuesta, tanto de forma reactiva como proactiva. Además, este marco de análisis utiliza una metodología basada en casos de uso, donde el operador es capaz de personalizar los parámetros de operación y las reglas para la inferencia de conocimiento. La propuesta ha sido evaluada sobre un caso de uso, donde se demuestra cómo a partir de una configuración sencilla es posible abastecer a las capas de inteligencia del conocimiento necesario para mejorar la toma de decisiones que permita adaptar la red a cambios en el volumen de datos monitorizados.**

***Palabras Clave*—5G, Conciencia Situacional, Inferencia de Conocimiento, NFV, Predicción, SDN.**

I. INTRODUCCIÓN

El crecimiento del número de dispositivos conectados a las infraestructuras móviles actuales y la demanda de servicios en línea han generado nuevos desafíos en la gestión de las infraestructuras de redes y telecomunicaciones. Los sistemas actuales requieren de respuestas rápidas frente a problemas típicos de red, tales como caída de enlaces o congestión, con el objetivo de garantizar la Calidad de Servicio (QoS) y la Calidad de Experiencia (QoE) de los usuarios finales. Se busca también que el tiempo de recuperación de servicios y los costes de capital (capex) y de operación (opex) disminuyan [2] cada vez más. Actualmente, la personalización de los servicios de red se ve limitada por la rigidez de las arquitecturas tradicionales debido a que tienen una fuerte dependencia de la configuración manual de dispositivos, así como de un lento proceso de estandarización (desde su diseño hasta su implementación) para la generación de nuevos productos y servicios. Según el reporte de la comisión Europea "Future Internet" la próxima generación de redes móviles (5G) superará dichas limitaciones, y se pronostican sus primeros resultados para el año 2020 [1]. Éstas tienen por objetivo el proporcionar un entorno de comunicación confiable y de alto rendimiento que asegure la provisión eficiente de servicios, y que a su vez garantice el cumplimiento de los Acuerdos de Nivel de Servicio (SLAs) [3]. La nueva generación se caracterizarará por indicadores de desempeño muy superiores a los ofrecidos por las generaciones predecesoras, considerando entre ellos las altas capacidades de transmisión, baja latencia, altos niveles de QoE/QoS (evaluados según métricas objetivas [4]), gran densidad de dispositivos por área geográfica, etc. Por lo tanto, 5G debe ser capaz no sólo de responder automáticamente a situaciones imprevistas que comprometan la operatividad de la red, sino también debe ofrecer un modelo de control heterogéneo y unificado [5]. Para alcanzar los objetivos planteados, 5G propone la integración de diversas tecnologías como Redes Definidas por Software (SDN) [6], Virtualización de Funciones de Red (NFV) [7], computación en la nube [8], Inteligencia Artificial (IA) [9], Redes Auto-Organizadas (SON) [10], entre otras. La combinación de dichas tecnologías facilitará el despliegue y gestión de nuevos servicios en un entorno abierto y con capacidades mejoradas en la utilización del espectro, virtualización y compartición de recursos, eficiencia energética, etc. SDN ha emergido como una nueva arquitectura en la cual el control de los dispositivos de red está centralizado en el controlador SDN, gestionando el comportamiento de la red a través de aplicaciones SDN. Asimismo, diversas funciones virtualizadas de red (NFVs) podrían ser desplegadas automáticamente y a demanda desde el controlador en un entorno virtualizado. El análisis del tráfico que circula a través de la red permitirá por lo tanto recurrir a esquemas de análisis más complejos que

harán posible inferir síntomas que proyecten escenarios de red que afecten el cumplimiento de los indicadores de desempeño en una red 5G, y den lugar a procesos complejos de toma de decisiones que se traduzcan en medidas de actuación. Con este objetivo, el presente trabajo propone la creación de un Marco para el Análisis e Inferencia de Conocimiento en Redes 5G que cumpla los requerimientos antes mencionados.

El trabajo se estructura en cinco secciones, siendo esta introducción la primera. La sección II describe los trabajos relacionados y el proyecto SELFNET. En la sección III se presenta el modelo de análisis para la inferencia de conocimiento en 5G. La sección IV describe los resultados preliminares del modelo propuesto. Finalmente, en la sección V se presentan las conclusiones.

II. TRABAJOS RELACIONADOS

La comunidad científica, académica, operadores de telecomunicaciones, proveedores de servicio, entre otros, han unido esfuerzos para el desarrollo de soluciones que mejoren la gestión y el análisis de información en redes 5G, tomando las ventajas de SDN, NFV, SON e inteligencia artificial. Tal es el caso del consorcio 5G Americas [11], o las iniciativas para la adopción de 5G en Asia [12], entre otras. En Europa, el consorcio 5G-PPP [13] promueve el desarrollo de proyectos de investigación que brinden soluciones a los desafíos de 5G a través de la sinergia de diversas áreas del conocimiento. En la Tabla I se describen proyectos relevantes en el ámbito de 5G.

En particular, el proyecto SELFNET (Self-Organized Network Management in Virtualized and Software Defined Networks) [14] ha sido creado con el fin de proporcionar un marco de gestión autónomo que provea capacidades de auto-organización en redes móviles 5G. Este proyecto se encuentra en desarrollo y es financiado por el programa marco Horizonte 2020. SELFNET hace uso de los principios de SDN y NFV para gestionar de forma inteligente y autonómica diversas funciones de red con el propósito de dar solución automática a problemas comunes de red, tales como el retardo en la transmisión, la caída de enlaces, etc. Para ello, SELFNET [15] integra un paradigma de autogestión de la red basado en el uso de algoritmos de predicción, reconocimiento de patrones, técnicas de minería de datos, entre otros, para identificar el comportamiento de una red móvil 5G, y determinar las mejores acciones que mitiguen de forma automática los problemas de red detectados. SELFNET se basa en la definición de tres grandes casos de uso: la autoprotección (self-protection), la autosalud (self-healing) y la autooptimización (self-optimization), en los que se distinguen escenarios sobre los que el sistema requiera procesos de toma de decisiones y el despliegue de acciones de respuesta. Para este propósito, un conjunto de sensores y actuadores SDN/NFV (por ejemplo, Sistemas de Detección de Intrusión (IDS), servidores de Inspección Profunda de Paquetes (DPI), anti-malware, etc.) serán desplegados para analizar el estado de la red y detectar situaciones que sean controladas de forma reactiva o proactiva. Los sensores y actuadores pueden ser desplegados tanto en la infraestructura física como virtual de SELFNET.

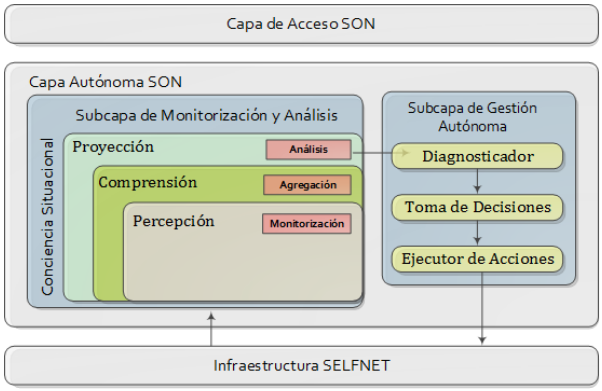


Fig. 1. Modelo de Conciencia Situacional en SELFNET.

En SELFNET la inteligencia de red es provista por la capa Autónoma SON. Esta capa recopila métricas relacionadas con el comportamiento de la red y utiliza esta información para inferir el estado de la misma. Luego, decide las acciones a ser ejecutadas para cumplir los objetivos del sistema. En particular, SELFNET proporciona un sistema de Monitorización, Agregación y Correlación, y Análisis (Subcapa de Monitorización y Análisis), que facilita el conocimiento del contexto de operación a través del modelo de consciencia situacional propuesto por Endsley [16] (ver Fig. 1). Los procesos de SELFNET previos a la toma de decisones son descritos a continuación:

- *Monitorización*. El principal objetivo de la monitorización es recopilar información proveniente de la infraestructura de red física y virtual, en forma de métricas de bajo nivel y de eventos, para facilitar su acceso a capas superiores.
- *Agregación y correlación*. En esta etapa las métricas de bajo nivel sirven de entrada a procesos de agregación con el objetivo de reducir el alto volumen de datos recibido, obteniéndose métricas agregadas sobre algún domino particular del sistema. Por otra parte, los eventos recibidos son correlacionados con el fin de filtrar información no sensitiva o redundante (por ejemplo, eliminando alertas repetidas) y también de obtener una visión global de la red en función del contexto de operación.
- *Análisis*. En el componente de análisis se lleva a cabo la identificación de escenarios que potencialmente amenacen la operatividad de la red a partir de la información sensada. Cuando estos escenarios son inferidos son enviados a la Subcapa de Gestión Autónoma, la cual utiliza diferentes algoritmos y técnicas avanzadas de inteligencia para determinar la causa del problema (Diagnosticador). Luego se define la mejor estrategia de reacción (Toma de Decisiones), para que finalmente se lleven a cabo las acciones de respuesta o mitigación del problema (Ejecutor de Acciones).

Este artículo se centra en el módulo de análisis y la inferencia de conocimiento.

Tabla I
PROYECTOS 5G.

Proyecto	Tecnología	Descripción
CROWD [17]	SDN, NFV	Este proyecto tiene el objetivo de aumentar la capacidad en el parámetro de densidad de las redes de acceso inalámbrico en redes heterogéneas. Se enfoca en garantizar la QoE de los usuarios móviles, la optimización de los recursos y el consumo de energía.
T-NOVA [18]	SDN, NFV	Este proyecto se enfoca en el despliegue de Funciones de Red como Servicio (NFaaS) sobre infraestructuras de red virtualizadas. T-Nova diseña e implementa una plataforma de gestión y orquestación automatizada para la provisión, configuración, monitorización y optimización de recursos virtualizados, basada en SDN y NFV.
UNIFY [19]	SDN, NFV	El proyecto tiene como objetivo desarrollar una plataforma dinámica para la creación de servicios. Unify permite el despliegue dinámico y automático de los servicios en entornos cloud (red, computación y almacenamiento). De manera similar, el orquestador incluye algoritmos de optimización para la asignación óptima de dichos recursos a lo largo de la infraestructura.
COGNET [20]	SDN, NFV, Aprendizaje de máquina	Este proyecto se centra en el mejoramiento de las tareas de monitorización y la gestión automática de red, para lo cual predice la demanda de recursos y luego cambia su configuración basada en el análisis actual de la red (detección de errores, fallos de seguridad).
SELFNET [15]	SDN, NFV	El proyecto provee un marco de gestión autónoma de red, basado en el despliegue inteligente de sensores y actuadores tanto en la infraestructura física como virtual. Las funciones pueden ser basadas en SDN o NFV.
CHARISMA [21]	SDN, NFV	Este proyecto permite el despliegue inteligente de servicios para redes de acceso de radio en entornos cloud (C-RAN).
5G-ENSURE [22]	SDN, NFV	El proyecto se enfoca en la evaluación de riesgos y su mitigación para cubrir los requerimientos de seguridad de 5G. 5G-Ensure cubre un rango amplio de problemas de seguridad, desde el aseguramiento de dispositivos físicos hasta de recursos SDN o NFV.

III. MARCO PARA EL ANÁLISIS E INFERENCIA DE CONOCIMIENTO EN 5G

El componente de análisis tiene el objetivo de inferir conocimiento sobre la base de la información proveniente de los niveles inferiores de la arquitectura (monitorización, agregación y correlación), para identificar situaciones potenciales (síntomas) que desencadenen procesos avanzados de diagnóstico y toma de decisiones en el componente de inteligencia de SELFNET (ver Fig. 2). A continuación se describen cada uno de sus componentes y su relación con el resto de elementos del sistema.

A. Componentes

Se distiguen siete pasos principales para el procesamiento de información: Integración de Casos de Uso [O], Descubrimiento [DIS], Reconocimento de Patrones [PR], Predicción [FT], Umbrales Adaptativos [ATh], Inferencia de Conocimiento [KI] y Notificación [N].

- *Integración de Casos de Uso*. En este conjunto de acciones se definen los descriptores que permiten la elaboración de la ontología de información, tanto a nivel procedimental como factual, necesaria para configurar el marco de análisis en función de los requisitos de los casos de uso. A partir de ellos se definen desde los objetos O a tener en cuenta, hasta la reglas Ru de inferencia de conocimiento.
- *Descubrimiento [DIS]*. Esta operación permite enlazar las capas de agregación y de análisis de SELFNET para que la primera envíe periódicamente datos agregados construidos a partir de las observaciones extraídas por los sensores desplegados a lo largo del escenario de monitorización. El módulo de análisis traduce dichas observaciones en hechos (*Fa*) que se añaden a la memoria de trabajo, para que en lo posterior se lleven a cabo las tareas de reconocimiento de patrones, predicción o umbrales adaptativos.
- *Reconocimiento de Patrones [PR]*. En este proceso se identifican ciertos patrones (*PR*) adquiridos o conocidos anteriormente sobre hechos (*Fa*) relacionados con datos agregados o eventos, y se generan nuevos hechos (*Fa(PR)*) derivados del conocimiento resultante de de su estudio. Con este propósito, diferentes tareas internas son ejecutadas: estudio de los datos de entrada (tanto los datos de entrenamiento como las observaciones realizadas), decisión de la estrategia de datos más adecuada para cada contexto, características de extracción, construcción de modelos o regresiones, etc. La selección de métodos de reconocimiento de patrones es adaptada a las necesidades de cada escenario de seguridad. El marco de análisis de SELFNET se enfoca en dos acciones fundamentales: la identificación de firmas de eventos conocidos previamente y la detección de anomalías.
- *Predicción [FT]*. En esta etapa se lleva a cabo el cálculo de las métricas de predicción (*Ft*), expresadas como hechos (*Fa(Ft)*), asociadas a cada escenario a partir de las observaciones provistas por la fase de agregación y correlación. Este proceso implica diferentes etapas: gestión del historial de datos requeridos para construir el modelo de predicción, decisión del algoritmo de predicción más adecuado, y la evaluación de los resultados que faciliten el aprendizaje basado en las decisiones previas. El pronóstico de situaciones de red facilita la optimización de recursos, el despliegue de acciones proactivas y permite anticipar la identificación de riesgos en el entorno de monitorización. Este componente incluye una batería de algoritmos de predicción (alisamiento exponencial, medias móviles, modelo autorregresivo integrado de media móvil (ARIMA), etc.) que son evaluados con distintos parámetros de ajuste para determinar el método de predicción aplicado a la serie temporal que

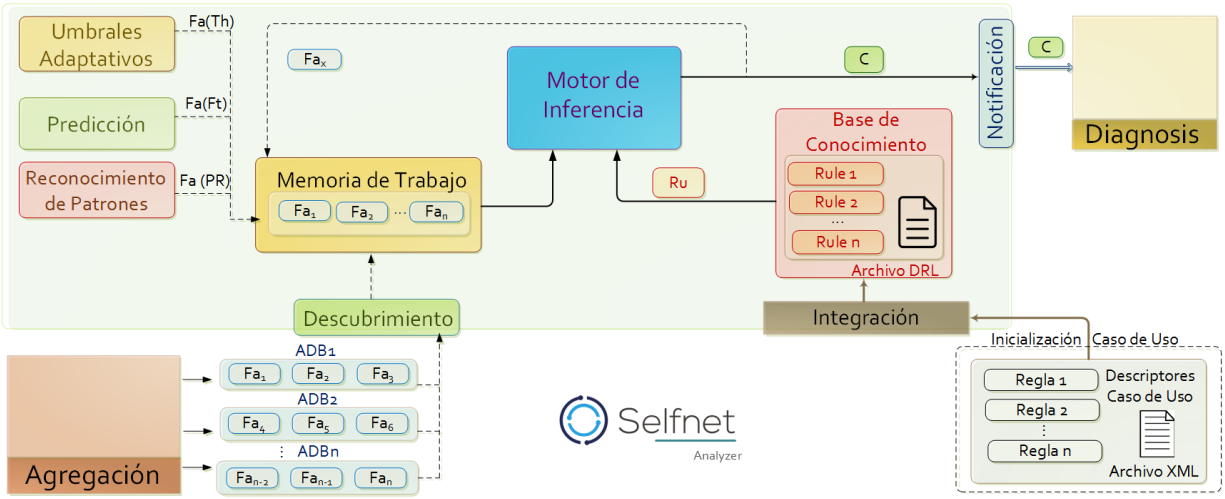


Fig. 2. Inferencia de Conocimiento en SELFNET

minimiza el error de estimación. Con el método seleccionado, el marco de análisis de SELFNET predice los horizontes requeridos para una serie temporal dada.

- *Umbrales adaptativos [ATh]*. En este proceso se llevan acabo las acciones que permiten definir cuando los errores de predicción deben ser tenidos en cuenta. De este modo se construyen umbrales (*Th*) y se genera nuevos hechos (*Fa(Th)*) a partir de ellos. Su elaboración implica diferentes pasos, tales como el análisis y la extracción de las características principales de los datos de entrada, decisión de los algoritmos más adecuados, o modelado y estimación de los valores en los umbrales. La principal aplicabilidad de los umbrales adaptativos en SELFNET es que considera el contexto de operación en la inferencia de nuevos hechos relacionados con el filtrado, reduciendo así las tasas de falsos positivos de sus sensores.
- *Inferencia de Conocimiento [KI]*. En esta tarea se aplican las reglas (*Ru*) de producción contenidas en la base de conocimiento con el objetivo de deducir nuevo conocimiento. Las reglas son generadas por el operador del sistema, y se determinan por cada caso de uso. El motor de inferencia opera bajo un esquema de encadenamiento hacia adelante (*forward chaining*), es decir, considera primero los hechos conocidos con anterioridad e infiere nuevos hechos hasta que sea capaz de inferir conclusiones (*C*) [25], [26]. Además, es importante tener en cuenta que la implementación más sencilla del motor de inferencia considera reglas de separación (modus ponens) basadas en lógica proposicional. Las reglas pueden ser adaptadas a una representación diferente de incertidumbre, tal como la lógica difusa, conjuntos aproximados, o redes Bayesianas.
- *Notificación [N]*. Esta última etapa tiene por objetivo actuar como enlace entre el conocimiento adquirido por el módulo de Análisis y la subcapa de inteligencia de SELFNET (Subcapa Autónoma SON).

Adicionalmente, el marco de análisis provee un módulo de Interfaz de Usuario que permite al operador configurar las reglas de inferencia por cada caso de uso, así como un módulo de Notificación cuya función es enviar las conclusiones obtenidas al componente de inteligencia.

B. *Entradas y salidas*

Desde el componente de agregación llegan métricas agregadas y eventos correlacionados. Los módulos internos del marco de análisis (reconocimiento de patrones, predicción y umbrales adaptativos) actúan como proveedores internos de nuevos hechos que sirven como entradas en la memoria de la base de conocimiento. Asimismo, el conjunto de reglas para un caso de uso en particular, desde la interfaz de usuario, complementa el conjunto de entradas necesarias. Las conclusiones del proceso de inferencia de conocimiento son expresadas como reportes que se envían como salida al componente de inteligencia de SELFNET.

IV. EXPERIMENTACIÓN Y RESULTADOS PRELIMINARES

A. *Experimentación*

La experimentación se conduce sobre un caso de uso de ejemplo, implementado para la obtención de resultados. El objetivo es recorrer las etapas definidas en el marco de análisis para la generación de conocimiento; las cuales incluyen un mecanismo básico de reconocimiento de patrones, predicción de datos sobre una serie temporal, construcción de intervalos de predicción adaptativos, y aplicación de reglas de producción con el fin de generar conclusiones que denoten el estado de la red.

1) Conjuntos de Muestras:

Se describen a continuación los dos conjuntos de muestras usados en la experimentación. El primero, con fines de validación del componente de predicción. El segundo, para la generación de conocimiento (conclusiones) en el motor de inferencia.

Competición M3

Para evaluar la precisión del componente de predicción se

Tabla II
CARACTERÍSTICAS DE LAS SERIES TEMPORALES EN LA COMPETICIÓN M3.

Frecuencia	Series Temporales	Observaciones	Horizontes
Anual	645	19	6
Cuatrimestral	756	44	8
Mensual	1428	115	18
Otra	174	63	8

cuenta con un conjunto de muestras de propósito general usado en la competición M3 [23]. Esta competición reúne un conjunto de 3003 series temporales sobre las que se aplicaron diversos métodos de predicción en distintos horizontes ($t+1$, $t+2$, ..., $t+18$), según el número de observaciones contenidas en ellas. Las características y horizontes evaluados en cada caso se resumen en la Tabla II. La precisión de los métodos de predicción aplicados en esta prueba se compara luego con los resultados obtenidos por SELFNET.

Tráfico de red UCM 2011

Se cuenta además con un conjunto de muestras de tráfico real elaborado a partir de capturas proporcionadas por el Centro de Cálculo y Procesamiento de Datos de la Universidad Complutense de Madrid (UCM). Esta colección consta de diferentes trazas de tráfico monitorizado en la facultad de informática de la UCM durante varios días, en diferentes periodos de tiempo, a lo largo del año 2011 [27]. En la experimentación realizada han sido compactadas en formato pcap [27], y son usadas para predecir el volumen de tráfico de red en los siguientes horizontes temporales. Por cada paquete contenido en el archivo pcap se obtiene como métrica el número de bytes transmitidos. Este valor corresponde a una métrica de bajo nivel en la capa de monitorización de SELFNET. El número de bytes transmitidos se acumula (mediante un promedio) en intervalos de 5 segundos generando así una serie temporal en la que cada observación promediada, expresada en Kilobytes, representa una métrica agregada. Esta operación corresponde a la capa de agregación de SELFNET y la serie temporal producida sirve como información de entrada para el marco de análisis propuesto.

2) Caso de Uso: Comportamiento anómalo del volumen de tráfico: El objetivo de este caso de uso es inferir si el volumen de tráfico de la red presenta un comportamiento anómalo. Los distintos componentes del marco de análisis contribuyen a la generación de hechos en la memoria de trabajo, sobre los que el motor de inferencia aplicará las reglas de producción configuradas. El proceso se inicia cuando se recibe la serie temporal del volumen de tráfico. Cada vez que se añade un nuevo elemento a la serie, el componente de reconocimiento de patrones (PR) evalúa si existe una tendencia creciente o decreciente en las observaciones, e introduce nuevos hechos en la memoria de trabajo que registren dicho patrón. En la siguiente etapa de procesamiento de información se lleva a cabo su predicción, estableciéndose un horizonte temporal de 8 observaciones para la generación de las estimaciones que serán insertadas en la memoria de trabajo. Una vez conocidas las métricas pronosticadas, se definen los intervalos de predicción (PI) para cada nueva observación que se añade a la serie temporal, y se registran dichos intervalos en la memoria de trabajo del sistema experto. El PI se compone de un umbral superior (U. Sup) y uno inferior (U. inf) que se generan a través de la comparación entre el valor observado y el valor pronosticado, según se detalla en [24]. El patrón identificado, las predicciones, y los umbrales, permiten al motor de inferencia determinar si el tráfico de red es anómalo cuando éste se encuentra fuera del PI. Además, el motor de inferencia evalúa el patrón de la observación para conocer si el volumen de tráfico anómalo se acerca o aleja del intervalo de predicción (congestión o subutilización de la red, respectivamente). Por ejemplo, si una observación ha excedido el umbral superior y su tendencia es creciente, se generará un síntoma de congestión con tendencia positiva (+), lo que podría desencadenar el despliegue de acciones de mitigación específicas desde el componente de inteligencia de SELFNET, tales como la ampliación del ancho de banda o la instanciación de nuevas funciones de red; por el contrarion, si el síntoma determina una tendencia negativa (-), es posible que se decida eliminar instancias de funciones de red a fin de mejorar la Calidad de Servicio (QoS) y la Calidad de Experiencia (QoE), así como de reducir el consumo de recursos.

B. Resultados

Los resultados se han dividido en dos secciones. En la primera, se validaron las capacidades de predicción del marco propuesto. En la segunda sección, se presentan los resultados obtenidos tras la implementación del caso de uso.

1) Evaluación del módulo de Predicción:

La precisión de los resultados obtenidos por el componente de predicción ha sido comparada con la precisión obtenida por diversos métodos de predicción aplicados en la competición M3. La precisión fue evaluada mediante el cálculo del Error Porcentual Absoluto Medio Simétrico (SMAPE) [23], cuyo valor fluctúa entre 0 y 200%, siendo éste el error de predicción en cada caso. Dado que los errores se calcularon en varias series temporales (por ejemplo 756 en datos cuatrimestrales), se obtuvo el valor promedio del SMAPE por cada horizonte de predicción y por cada método aplicado. Nótese que cuanto menor es el valor del error SMAPE promedio, más precisa es la predicción que evalúa. En la experimentación se comparó el SMAPE promedio obtenido por SELFNET con el mínimo valor de SMAPE promedio hallado en la competición M3 para un determinado horizonte temporal, encontrándose que SELFNET obtuvo, en general, mejores resultados de predicción en todas las series temporales analizadas y en casi todos los horizontes evaluados. En las 645 series temporales anuales, SELFNET mostró un mejor resultado en todos los horizontes temporales. Estos resultados se resumen en la Tabla III.

En las 756 series cuatrimestrales, SELFNET fue superado sólo en el horizonte t+1 por el método de predicción *PP-autocast*, cuyo SMAPE promedio fue de 4.8, siendo

Tabla III
VALORES SMAPE DE M3 Y SELFNET EN LAS SERIES ANUALES

Método	1	2	3	4	5	6
M3 mejor	7.6	11.8	16.1	18.2	13.4	22.7
SELFNET	6.9	6.6	7.6	7.2	8.5	9.4

Tabla IV
VALORES SMAPE DE M3 Y SELFNET EN LAS SERIES CUATRIMESTRALES

Método	1	2	3	4	5	6	8
M3 mejor	4.8	6.6	7.4	8.8	9.4	10.9	12
SELFNET	5.3	5.2	4.5	4.7	4.4	4.8	4.9

inferior al 5.3 que obtuvo SELFNET. En todos los demás casos, la precisión de SELFNET fue superior. Estos resultados se resumen en la Tabla IV.

En las 1428 series mensuales, SELFNET fue superado por M3 sólo en dos ocasiones. En el horizonte t+2, el método ForecastPro obtuvo un SMAPE promedio de 10.7 (menor a SELFNET en 0.5), mientras que en el horizonte t+4 el método Theta obtuvo un SMAPE promedio de 12.4, inferior al de SELFNET en 0.1. En todos los demás casos, los resultados de SELFNET fueron mejores. Estos resultados se resumen en la Tabla V.

En las otras 174 series temporales, SELFNET mostró un mejor resultado en todos los horizontes temporales, con excepción de t+1 en donde el método Autobox 2 de M3 obtuvo un SMAPE promedio de 1.6 que fue inferior a SELFNET en 0.2. Estos resultados se resumen en la Tabla VI.

2) Evaluación del caso de uso: La serie temporal Tráfico de red UCM 2011 cuenta con 31 elementos. Por ello, se hicieron las predicciones para 8 horizontes temporales debido a que el tamaño de esta serie temporal es cercano al tamaño promedio de las series cuatrimestrales en la competición M3. No obstante, este parámetro es configurable en el marco de análisis de SELFNET. La detección del patrón respondió a un algoritmo básico que comparó la observación actual con la anterior para determinar si existe una tendencia creciente o decreciente en los datos de la muestra. Los patrones identificados se muestran en la Tabla VII.

Para hallar las predicciones se restaron ocho elementos a la serie temporal original, los que fueron luego pronosticados y comparados con los valores originales para estimar la precisión. Para el tráfico de red UCM, el componente de predicción de SELFNET seleccionó el algoritmo de Suavizamiento Exponencial Simple (SES), con α aproximado a 0.7 y un SMAPE medio de 5.342725 (obtenido de los ocho horizontes), como el algoritmo

Tabla VI
VALORES SMAPE DE M3 Y SELFNET EN LAS OTRAS SERIES TEMPORALES

Método	1	2	3	4	5	6	8
M3 mejor	1.6	2.7	3.8	4.3	5.3	5.1	6
SELFNET	1.8	2.3	2.2	2	2.3	1.5	2.4

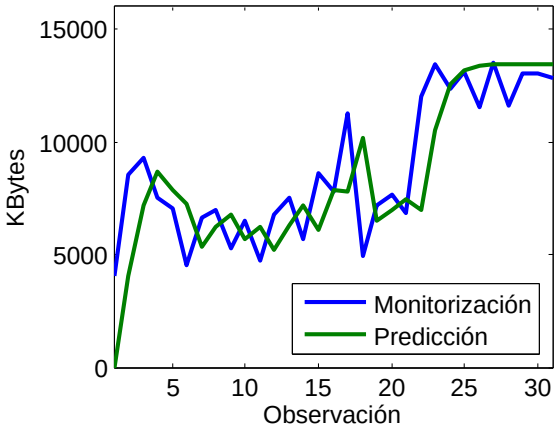


Fig. 3. Serie Temporal Real y Serie Pronosticada para UCM 2011.

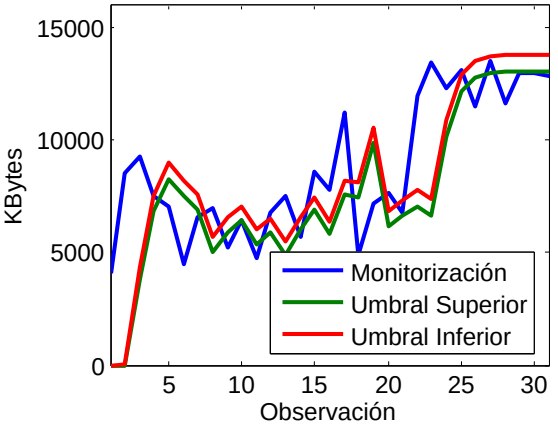


Fig. 4. Observaciones y Umbrales Calculados para UCM 2011.

que mejores resultados obtuvo tras la comparativa con otros métodos de predicción contenidos en su batería de algoritmos. Los resultados se muestran en la Tabla VII. A su vez, la serie original y la serie pronosticada se grafican en Fig. 3.

A partir de los resultados pronosticados SELFNET estimó el intervalo de predicción, construyendo para ello los umbrales adaptativos superior e inferior con un valor

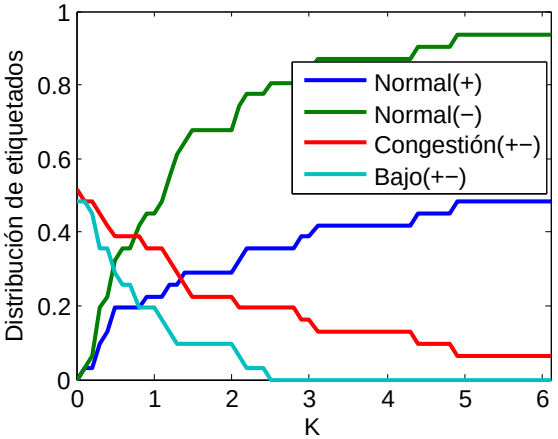


Fig. 5. Resultados al variar K para UCM 2011.

Tabla V
VALORES SMAPE DE M3 Y SELFNET EN LAS SERIES ANUALES

Método	1	2	3	4	5	6	8	12	15	18
M3 mejor	11.2	10.7	11.7	12.4	11.8	12.2	12.6	13.2	16.2	17.8
SELFNET	11	11.2	11.7	12.5	11.6	11.4	10.6	9.6	11	12.7

Tabla VII
RESULTADOS OBTENIDOS POR EL MARCO DE ANÁLISIS EN EL CASO DE USO EVALUADO

N	KPI(Volumen) (KB)	Patrón	Predicción	U. Sup.	U. Inf.	Síntoma
1	4063	Crece	0.0	0.0	0.0	Congestión (+)
2	8557	Crece	4062.5	215.8	-215.8	Congestión (+)
3	9264	Crece	7208.4	5125.3	2999.8	Congestión (+)
4	7498	Decrece	8647.0	8592.8	5824.0	Normal (-)
5	7030	Decrece	7842.8	10149.4	7144.6	Tráfico escaso (-)
6	4519	Decrece	7273.7	9217.2	6468.4	Tráfico escaso (-)
7	6616	Crece	5345.1	8623.1	5924.3	Normal (+)
8	6979	Crece	6235.1	6717.3	3972.9	Congestión (+)
9	5251	Decrece	6756.2	7545.4	4924.7	Normal (-)
10	6469	Crece	5702.3	8061.5	5450.8	Normal (+)
11	4739	Decrece	6239.3	6953.9	4450.7	Normal (-)
12	6779	Crece	5188.8	7441.1	5037.6	Normal (+)
13	7535	Crece	6302.1	6356.8	4020.9	Congesti?n (+)
14	5671	Decrece	7165.4	7431.1	5173.1	Normal (-)
15	8631	Crece	6119.5	8269.8	6061.0	Congestión (+)
16	7772	Decrece	7877.4	7270.3	4968.7	Congestión (-)
17	11239	Crece	7803.5	9062.0	6692.8	Congestión (+)
18	4926	Decrece	10208.6	9197.4	6409.7	Tráfico escaso (-)
19	7194	Crece	6510.5	11598.9	8818.4	Tráfico escaso (+)
20	7648	Crece	6988.9	7895.4	5125.6	Normal (+)
21	6809	Decrece	7450.2	8366.9	5610.8	Normal (-)
22	12001	Crece	7001.1	8948.4	5952.0	Congestión (+)
23	13443	Crece	10501.4	8479.2	5523.0	Congestión (+)
24	12350	Decrece	12560.3	11993.3	9009.4	Congestión (-)
25	13118	Crece	13177.9	14068.5	11052.0	Normal (+)
26	11534	Decrece	13363.2	14656.9	11698.9	Tráfico escaso (-)
27	13512	Crece	13418.8	14853.1	11873.3	Normal (+)
28	11630	Decrece	13435.5	14881.9	11955.7	Tráfico escaso (-)
29	13026	Crece	13440.5	14895.5	11975.5	Normal (+)
30	13015	Decrece	13442.0	14895.9	11985.1	Normal (-)
31	12856	Decrece	13442.4	14887.6	11996.4	Normal (-)

de ajuste K igual a 1 (Fig. 4). Con los umbrales definidos y el patrón previamente conocido, el motor de inferencia genera conclusiones sobre el estado de la red. Si la observación analizada excede el umbral superior o es menor al umbral inferior, se infieren anomalías de congestión o subutilización de la red, respectivamente, además de indicar si la tendencia es creciente (+) o decreciente (-). Las conclusiones obtenidas para cada observación se muestran en la última columna de la Tabla VII. Como puede verse, para las últimas ocho observaciones (horizonte temporal pronosticado), el motor de inferencia notifica una anomalía de congestión creciente para la observación 24, y dos de tráfico escaso y decreciente correspondiente a las observaciones 26 y 28. Es importante tener en cuenta que la variación del parámetro K efecta directamente al nivel de restricción bajo el que operan los umbrales adaptativos. Tal y como se muestra en Fig. 5, en los valores más bajos de K aumenta la distribución de síntomas reportados relacionados con la identificación de congestión en la red o bajadas representativas del volumen de datos promedio en el tráfico monitorizado; en concreto, cuando K se aproxima a 0, éstos suponen más de la mitad de los informes emitidos. Sin embargo, a medida que decrede el nivel de restricción, disminuye su frecuencia de aparición, llegando a representar aproximadamente el 10% de los etiquetados realizados. Nótese que en la actualidad existen diferentes estrategias de calibrado y adaptación de variables de ajuste similares, a los diferentes problemas relacionados con adaptar sensores a escenarios de monitorización no estacionales, recópilandose en [28] algunas de las más relevantes. Pero dada su complejidad y a menudo, dependencia de los casos de uso, profundizar en ello está fuera del alcance de este artículo, estableciéndose de esta manera una interesante linea de trabajo futuro.

V. CONCLUSIONES

El trabajo presentado propone un marco de análisis para la inferencia de conocimiento en redes 5G, y se enmarca en el desarrollo del proyecto de financiación Europea SELFNET. Su principal objetivo es generar conocimiento, a partir de la información monitorizada y agregada en etapas previas de procesamiento, con el cual es posible generar conclusiones, expresadas como síntomas, acerca del estado actual y futuro de la red. Las métricas agregadas y eventos provenientes de los niveles de infraestructura física y virtual sirven además para identificar o desencadenar alertas de seguridad sobre un dominio de red específico. Para la adquisición de conocimiento, el marco

de análisis distingue las etapas de integración de casos de uso, descubrimiento, reconocimiento de patrones, predicción, cálculo de umbrales adaptativos, inferencia de conocimiento y notificación. El conocimiento obtenido sobre el contexto de operación permite que la capa de gestión autónoma (inteligencia) mejore y optimice las tareas de diagnóstico y toma de decisiones relacionadas con la gestión de incidencias, facilitando así el despliegue automático de contramedidas proactivas y reactivas que busquen mitigar los problemas de red detectados.

La aproximación realizada ha sido probada sobre un caso de uso real, donde se ha procesado información procedente de capturas de tráfico monitorizadas en la red de la Universidad Complutense de Madrid. Con este fin se ha procedido a su instanciación, y el componente de predicción ha sido validado a partir del estándar funcional resultante de la competición M3, demostrándose su capacidad de ofrecer pronósticos precisos. En los resultados obtenidos es posible observar cómo desde un ajuste muy sencillo, es posible inferir conocimiento para adaptar la red y su configuración a la evolución del volumen de tráfico observado. Además, el camino trazado permite distinguir diferentes líneas de investigación orientadas a su mejora, como por ejemplo el diseño de estrategias de autoajuste adaptadas a los escenarios 5G, la definición de síntomas más precisos, o la identificación de métodos que mejoren su robustez frente a técnicas de evasión, como por ejemplo los métodos de imitación (mimicry). También debe considerarse la necesidad de ampliar la experimentación a casos de uso más complejos, que presenten una mayor similitud con circunstancias potencialmente observables en redes de 5G. Es importante descatar que a día de hoy no existen colecciones de muestras ni trazas de tráfico con estas características, por lo que habrá que esperar a que estas tecnologías alcancen un mayor nivel de madurez para ser capaces de evaluar con precisión tanto el trabajo descrito en este artículo, como el resto de iniciativas de investigación en 5G.

AGRADECIMIENTOS

This work was funded by the European Commission Horizon 2020 Programme under Grant Agreement number H2020-ICT-2014-2/ 671672 SELFNET (A Framework for Self-Organized Network Management in Virtualized and Software Defined Networks).

REFERENCIAS

[1] J.C. Hourcade, Y. Neuvo, R. Saracco, W. Wahlster, R. Posch, "Future Internet 2020: Visions of an Industry Expert Group", Panel Report, May 2009.

[2] P.K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, A. Benjebbour, "Design Considerations for a 5G Network Architecture", IEEE Communications Magazine, vol. 52, n. 11, pp. 65-75, November 2014.

[3] L.I. Barona López, A.L. Valdivieso Caraguay, M.C. Sotelo Monge, L.J. García Villalba, "Key Technologies in the Context of Future Networks: Operational and Management Requirements", Future Internet, vol. 9, n. 1, pp. 1-15, October 2016.

[4] "ITU-T P-1201: Parametric non-intrusive assessment of audiovisual media streaming quality". http://handle.itu.int/11.1002/1000/11727 (Accessed 14 July 2017).

[5] A. Belmonte Martin, L. Marinos, E. Rekleitis, G. Spanoudakis, N. Petroulakis, "Threat Landscape and Good Practice Guide for Software Defined Networks/5G", January 2016.

[6] D. Kreutz, F.M. Ramos, P.E. Verissimo, C.E. Rothenberg, S. Azodolmolky, S. Uhlig, "Software-defined Networking: A Comprehensive Survey", in Proc. of the IEEE, vol. 103, n. 1, pp. 14-76, December 2015.

[7] S. Abdelwahab, B. Hamdaoui, M. Guizani, T. Znati, "Network Function Virtualization in 5G", IEEE Communications Magazine, vol. 54, n. 4, pp. 84-91, April 2016.

[8] Q. Zhang, L. Cheng, R. Boutaba, "Cloud Computing: state-of-the-art and Research Challenges", Journal of Internet Services and Applications", vol. 1, n. 1, pp. 7-18, May 2014.

[9] A. Imran, A. Zoha, A. Abu-Dayya, "Challenges in 5G: How to Empower SON with Big Data for Enabling 5G", IEEE Network, vol. 28, n. 6, pp. 27-33, December 2014.

[10] N. Baldo, L. Giupponi, J. Mangues-Bafalluy, "Big Data Empowered Self Organized Networks", in Proc. of the 20th European Wireless Conference, pp. 1-8, Barcelona, Spain, May 2014.

[11] "5G Americas", http://www.5gamericas.org/es/ (Accessed 01 May 2017).

[12] "IMT-2020 (5G) Promotion Group", http://www.imt-2020.cn/en/introduction (Accessed 01 May 2017).

[13] W. Mohr, "The 5G Infrastructure Public-Private Partnership", ITU GSC-19 Meeting, 2015.

[14] "Self-Organized Network Management in Virtualized and Software Defined Networks (SELFNET)", Project reference: H2020-ICT-2014-2/671672, Funded under: H2020. http://www.selfnet-5g.eu (Accessed 01 May 2017).

[15] P. Neves, R. Calé, M.R. Costa, C. Parada, B. Parreira, J. Alcaraz-Calero, Q. Wang, J. Nightingale, E. Chirivella-Perez, W. Jiang, "The SELFNET Approach for Autonomic Management in an NFV/SDN Networking Paradigm", International Journal of Distributed Sensor Networks, pp. 1-17, December 2015.

[16] M.R. Endsley, "Design and Evaluation for Situation Awareness Enhancement", in Proc. of 32nd Annual Meeting on Human Factors Society, vol. 32, no. 1, pp. 97-101, October 1988.

[17] "CROWD Project (2013) Connectivity Management for EneRgy Optimised Wireless Dense Networks", Funded under: FP7-ICT. Project Reference: 318115, http://www.ict-crowd.eu/ (Accessed 01 May 2017).

[18] "T-NOVA Project (2013) Network Functions as-a-Service over Virtualised Infrastructures", Funded under: FP7-ICT. Project Reference: 619520. http://www.t-nova.eu/ (Accessed 01 May 2017).

[19] "UNIFY Project (2013) Unifying Cloud and Carrier Networks". Funded under: FP7-ICT. Project Reference: 619609. http://www.fp7-unify.eu/ (Accessed 01 May 2017).

[20] L. Xu, H. Assem, I.G.B. Yahia, T.S. Buda, A. Martin, D. Gallico, M. Biancani, A. Pastor, P. Aranda, M. Smirnov, "CogNet: A Network Management Architecture Featuring Cognitive Capabilities", in Proc. of the European Conference on Networks and Communications, pp. 325-329, June 2016.

[21] "CHARISMA Project (2014) Converged Heterogeneous Advanced 5G Cloud-RAN Architecture for Intelligent and Secure Media Access". Funded under: H2020-ICT-2014-2. Project Reference: 671704. http://www.charisma5g.eu/ (Accessed 01 May 2017).

[22] "5G-Ensure Project (2014) Enablers for Network and System Security and Resilience". Funded under: H2020-ICT-2014-2. Project Reference: 671562. http://www.5gensure.eu/ (Accessed 01 May 2017).

[23] S. Makridakis, M. Hibon, "The M3-Competition: results, conclusions and implications", International Journal of Forecasting, vol. 16, n. 4, pp. 451-176, December 2000.

[24] S. Makridakis, S.C. Wheelwright, R.J. Hyndman, "Forecasting: Methods and Applications", John Wiley & Sons, 1998.

[25] C. Forgy, "Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem", Artificial Intelligence, vol. 19, n. 1, pp. 17-37, September 1982.

[26] Drools, http://https://www.drools.org (Accessed 01 May 2017).

[27] L.J. García Villalba, A.L. Sandoval Orozco, J. Maestre Vidal, "Advanced Payload Analyzer Preprocessor", Future Generation Computer Systems, doi.org/10.1016/j.future.2016.10.032, November 2016.

[28] G. Ditzler, M. Roveri, C. Alippi, R. Polikar, "Learning in Nonstationary Environments: A Survey" IEEE Computational Intelligence Magazine, vol. 10, no. 4, pp. 12-25, november 2015.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemáica (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6608

Simulación genérica a nivel de sistema para soluciones avanzadas de gestión de recursos

Paula Rodríguez, Paula Sarasúa, Luis Diez, Ramón Agüero
Departamento de Ingeniería de Comunicaciones
Universidad de Cantabria
Plaza de la Ciencia, s/n. 39005 Santander
{paula.rodriguez, paula.sarasua}@alumnos.unican.es, {ldiez, ramon}@tlmat.unican.es

***Resumen*—A pesar del notable esfuerzo llevado a cabo por la comunidad científica en el ámbito de la tecnología LTE, aún no existe una metodología globalmente aceptada para analizar este tipo de redes. Igualmente, no se conoce una única solución que responda a todos los requisitos que se pueden plantear a la hora de acometer su análisis, de modo que se emplean diferentes herramientas y soluciones, cada una con sus ventajas e inconvenientes. Una de las limitaciones más importantes es la dificultad para evaluar escenarios con un número elevado de elementos de red, lo que podría reflejar la situación de redes heterogéneas (HetNets). En otros casos, no se presta mucha atención a las características del nivel de servicio, suponiendo habitualmente que el sistema está saturado (*full-buffer*). Este trabajo presenta un entorno de simulación flexible y genérico (GWNSyM - *Generic Wireless Network System Modeler*), que permite el despliegue de escenarios complejos, y el análisis de diferentes técnicas y soluciones de gestión, así como nuevas arquitecturas de red. La herramienta se valida analizando una red muy heterogénea, con un elevado número de usuarios. Sobre este escenario se han analizado diferentes técnicas de acceso, incluyendo DUDe (Downlink-Uplink Decoupling), que plantea un cambio sustancial frente a las soluciones tradicionales de selección de acceso.**

***Palabras Clave*—Network Modeling, Simulation, LTE/LTE-A, DUDe**

I. INTRODUCCIÓN

De acuerdo a las previsiones actuales [1], la demanda de tráfico en redes móviles inalámbricas se incrementará de forma notable en los próximos años. Esto es debido, entre otras razones, a la consolidación de servicios de gran capacidad, tales como el *streaming* de vídeo o los juegos *online*, que compartirán los recursos de las redes con otros más tradicionales, como la navegación *web* o las descargas de ficheros.

Aunque a día de hoy las tecnologías 4G no están totalmente consolidadas, la comunidad investigadora ya está dirigiendo sus esfuerzos a la definición de las bases de la 5G, que daría soporte a la previsible heterogeneidad de servicios. Así, se prevé que en el futuro coexistan redes de diferentes tecnologías, y que la cooperación entre ellas se lleve a cabo de manera natural. Por ejemplo, se espera que las estrategias de densificación mediante *small-cells* jueguen un papel muy importante en los próximos años [2], ya que pueden proporcionar un importante aumento de la capacidad. Otras técnicas que se han incluido recientemente en las especificaciones del 3GPP son las referidas a la cooperación entre elementos de red, o técnicas Cooperative Multi-Point (CoMP), o el desacoplamiento de los enlaces ascendente y descendente de las conexiones [3], o Downlink-Uplink Decoupling (DUDe).

Además de las soluciones que se sitúan en las capas inferiores de las redes celulares (gestión de recursos), en las capas superiores las técnicas de virtualización [4], Network Function Virtualization (NFV) y Software Defined Networks (SDN), se presentan como elementos clave de los despliegues de red en los próximos años [5].

A pesar de los claros avances que proporcionan estas soluciones, también originan nuevas problemáticas, que requieren un estudio y análisis apropiados. Para ello, la comunidad investigadora se centra habitualmente en escenarios específicos y casos de uso concretos, extraídos del amplio abanico de posibilidades que aparecen a raíz de los nuevos conceptos de red.

Uno de los primeros problemas a la hora de comenzar un análisis concreto es la elección de la herramienta, o conjunto de herramientas, que pueden ser usadas a fin de llevar a cabo un estudio adecuado, en función del nivel de abstracción requerido. Las plataformas de simulación juegan un papel fundamental, debido a su versatilidad y coste. En este sentido, existen básicamente tres alternativas principales: (1) simuladores a nivel de enlace, (2) a nivel de sistema, o (3) simuladores de red, que permiten acometer análisis más detallados. La primera opción se centra habitualmente en el último salto (inalámbrico) de la comunicación, y facilita la evaluación de técnicas de enlace, como la estimación de canal, técnicas Multiple-Input Multiple-Output (MIMO) o soluciones relacionadas

con el Adaptive Modulation and Coding (AMC). Dentro de este grupo *Vienna LTE Simulator* [6] destaca como una de las soluciones mayoritariamente adoptadas. La segunda alternativa permite una mayor flexibilidad, aunque como contrapartida suele requerir alguna simplificación, con la consiguiente pérdida de precisión. La mayoría de los simuladores a nivel de sistema se basan en desarrollos propietarios (normalmente basados en MATLAB), aunque en algunos casos se usan algunas de las pocas herramientas específicas disponibles, donde nuevamente destaca el simulador LTE Vienna [6]. Finalmente, en el tercer grupo existen varias plataformas, aunque la que seguramente está recibiendo más atención últimamente es `ns-3` [7], y su extensión LTE-EPC Network Simulator (LENA) [8], que se encuentra en constante evolución. En el caso de simuladores de red, la mayor limitación son los tiempos para realizar los análisis, que se debe al alto grado de detalle de su implementación y modelos.

La Tabla I proporciona una comparativa detallada de las tres alternativas identificadas, en base a algunas de sus principales características. Una de las limitaciones que comparten las soluciones existentes es la dificultad para incluir o modelar nuevas técnicas y modelos de red, ya que su implementación suele ser bastante rígida, habitualmente centrada en escenarios concretos.

En base a las características de las soluciones de simulación existentes, surge la cuestión recurrente de qué herramienta se debería usar, de acuerdo al escenario y topología de red de interés, ya que no existe una solución que sea la idónea para todos los casos. El simulador *Vienna LTE* [6] está implementado en Matlab, y se centra en las capas inferiores, por lo que normalmente no es capaz de reflejar de manera apropiada diferentes patrones de servicio, adoptando habitualmente modelos de saturación o *full-buffer*. Por otro lado, los tiempos de simulación son bastante elevados, de modo que los análisis normalmente no cubren periodos de tiempo prolongados. Como consecuencia, en muchas ocasiones se opta por desarrollos propietarios, lo que requiere invertir un tiempo elevado en su desarrollo. Además, dado que se trata de soluciones ad-hoc, es complicado replicar los experimentos, así como integrarlos en otros entornos.

A fin de proporcionar una mejor respuesta, en este trabajo se presenta Generic Wireless Network System Modeler (GWNSyM), una plataforma flexible para la simulación de sistemas complejos. Esta herramienta se ha diseñado de modo genérico, para que sea fácilmente extensible con nuevas funcionalidades o soluciones de red. En este trabajo se presentan los principales aspectos tanto del diseño de la herramienta como de su implementación. Además, a modo de prueba de concepto, se presentará el análisis de diferentes técnicas de selección de acceso, haciendo uso de la herramienta.

El resto del artículo se estructura como sigue: en la Sección II se describe la funcionalidad básica de la herramienta, haciendo especial hincapié en los aspectos más relevantes de su implementación. A continuación, la Sección III presenta el escenario que se va a evaluar, los modelos implementados, así como los resultados obtenidos. Finalmente, la Sección IV enumera las principales conclusiones de este trabajo, así como las lineas futuras de investigación que surgen al aprovechar las posibilidades ofrecidas por la herramienta.

II. METODOLOGÍA DE SIMULACIÓN

Como se ha comentado anteriormente, el objetivo principal de GWNSyM es el de proveer una serie de abstracciones que faciliten el modelado de redes para su posterior simulación. En lo que respecta a la metodología de análisis, se optó por una simulación basada en fotografías, o *snapshots*, del sistema, en contrapartida a otros modelos basados en eventos. De este modo, cada fotografía representa un instante discreto de tiempo, en el que se aplican los modelos implementados sobre los elementos de red pertinentes y en un orden establecido. Además, el estado resultante de una fotografía es usado para alimentar la siguiente, de modo que se puede capturar la memoria del sistema, lo que es especialmente importante para analizar la evolución de servicios.

El entorno de simulación GWNSyM ha sido implementado como un conjunto de librerías en `C++`, teniendo como uno de sus objetivos principales la re-utilización de código, que puede darse en dos sentidos: (1) código generado en el entorno GWNSyM se pueda re-usar en otros entornos, y (2) código existente pudiera integrarse dentro del simulador. Para el primer caso, se decidió no imponer restricciones de herencias a las clases `C++` que implementan las diferentes entidades y modelos del simulador, sino que se estableció un mínimo interfaz que asegura la interacción de elementos GWNSyM. De forma resumida, por medio de técnicas de meta-programación se asegura que la compatibilidad entre los modelos implementados (por ejemplo, propagación o selección de acceso) y los elementos de red a los que se aplican, sin requerir ningún tipo de implementación (jerarquía, *namespaces*, etc.) específica, fomentando, de este modo, la separación de los modelos del sistema en el que se ejecutan. Del mismo modo, se ha añadido una funcionalidad de *wrapper*, que permite la integración de código existente dentro del entorno de simulación, dotándolo del interfaz requerido.

A continuación se explicará con más detalle los aspectos fundamentales de la simulación con GWNSyM, tanto en lo que se refiere a los elementos de simulación como al flujo de experimentación.

A. Elementos de simulación

Con el objetivo de no restringir el comportamiento de la herramienta de simulación a ninguna tecnología ni sistema en particular, se han definido dos elementos básicos que constituyen los escenarios de simulación GWNSyM: los *Tipos* representan elementos de red, mientras que las *Acciones* implementan modelos que se aplican sobre los *Tipos*.

Los *Tipos* definen la estructura de un elemento de red de forma general, junto con una configuración concreta. En este sentido, un elemento puede abarcar desde dispositivos

Tabla I: Análisis de características de alternativas de simulación para redes inalámbricas (LTE). Una clasificación subjetiva se otorga a los parámetros, de forma que en círculos rellanos significa bueno, mientras que los vacíos indican pobre rendimiento en ese parámetros

	Parámetro *Descripción de la característica de simulación que se necesita soportar*	**Simulación a nivel de enlace** *Modelado detallado de capas inferiores, lo que dificulta analizar escenarios con mas de un par fuente/destino*	**Simulación a nivel de sistema** *La mayoría de la literatura usa Matlab para realizar análisis. Vienna LTE Simulator es uno de los ejemplos más significativos*	**Simulación de red** *ns-3, junto con la extensión LENA es una de las alternativas más relevantes*
CARACTERÍSTICAS DEL ESCENARIO	*Complejidad del escenario: # de usuarios y estaciones base*	Debido al gran nivel de detalle de estas herramientas, el número de elementos el bastante bajo, normalmente un elemento de acceso y un conjunto de usuarios[9]	Se suelen asumir alsgunas simplificaciones, de modo que el número de elementos suele ser mayor	El tiempo de simulación requerido para analizar escenarios grandes es normalmente inaceptable [10], posibles alternativas usando paralelización [11]
	Dimensión temporal: tiempo que puede ser simulado y posibilidad de estudiar la evolución de servicios	Debido a la carga computacional [12], el tiempo simulado es bastante reducido, sin necesidad de mantener evolución de servicios	El uso de entornos de desarrollo pesado (Matlab) normalmente impide tiempos largos de simulación	Normalmente se considera la evolución de servicios, sin embargo, el tiempo de cómputo para simulaciones largas es muy elevado
	Precisión: grado de precisión de los modelos usados	El modelado detallado de las capas inferiores es su principal objetivo, por lo que la precisión es muy alta	Se asumen algunas simplificaciones aunque implementaciones disponibles siguen las especificaciones del 3GPP	Aunque los modelos pueden ser simplificados, la implementación de los protocolos es bastante precisa
MARCO TECNOLÓGICO	*Cambio de arquitectura: posibilidad de añadir y soportar nuevos paradigmas de red: SDN and NFV*	Como solución a nidel de enlace, no se consideran problemáticas de arquitectura	Algunas de las posibilidades de las nuevas funcionalidades de red (tighter cooperation schemes) normalmente se pueden modelar	Aunque la implementación puede ser costosa, la integración de nuevas opciones de arquitectura son normalmente posibles
	Soporte de diferentes tecnologías y soluciones/técnicas	Se encuentran bastante limitadas a las funcionalidades iniciales. La integración de diferentes tecnologías es normalmente compleja	Normalmente tienen flexibilidad para incorporar nuevas técnicas debido a las simplificaciones de capas inferiores	Las plataformas de simulación de redes son bastante flexibles, y permiten la integración de diferentes tecnologías y técnicas nuevas
	Modelado de servicios. Si se asume condiciones de saturación o carga constante	No se presta mucha atención al modelado de servicios, normalmente se centra en cómo los paquetes llegan a su destino en las capas inferiores	Presentan caracterización básica de servicios, aunque normalmente se asume carga constante o *full-buffer* [13]	Modelado de servicios relativamente avanzado. Permite incluso el uso de aplicaciones y servicios reales
OTROS ASPECTOS	*Propósito específico Vs. genérico y curva de aprendizaje*	Como su ámbito de aplicación está bastante delimitado, la curva de aprendizaje es relativamente corta	Aunque más específicos que los simuladores de red, no todos sus componentes son siempre de interés	Normalmente son grandes entornos de propósito general, por lo que requiere bastante tiempo de aprendizaje antes de poder realizar análisis
	Uso de metodologías complementarias. Técnicas de optimización	Normalmente se centran en analizar el rendimiento de una técnica concreta, y típicamente no buscan el rendimiento óptimo	Aunque no está entre sus objetivos principales, las técnicas de optimización se pueden integrar	La arquitectura del simulador ofrece una visión de conjunto, lo que permitiría aplicar estrategias de optimización global

de usuario a operadores, pasando por servicios o elementos virtuales. De acuerdo a su configuración, un *Tipo* puede agregar elementos de otro *Tipo*, de forma que se pueda definir la composición de cada elemento de red como una combinación de *Tipos*. La instanciación de elementos de un determinado *Tipo* define el conjunto de elementos correspondiente. A modo de ejemplo, la Figura 1 ilustra la creación de *Tipos* en GWNSyM para un caso genérico. Como se puede ver, el sustrato de los *Tipos* consiste en clases `C++` que, junto a una configuración concreta, da lugar a un *Tipo*. Así, la Figura 1 muestra cómo una misma clase `C++` $C1$ da lugar a dos *Tipos* ($T1$ y $T2$), en función de la configuración que se le aplica. Además, de acuerdo a la configuración, las instancias de cada *Tipo* pueden dar lugar a diferentes agregaciones.

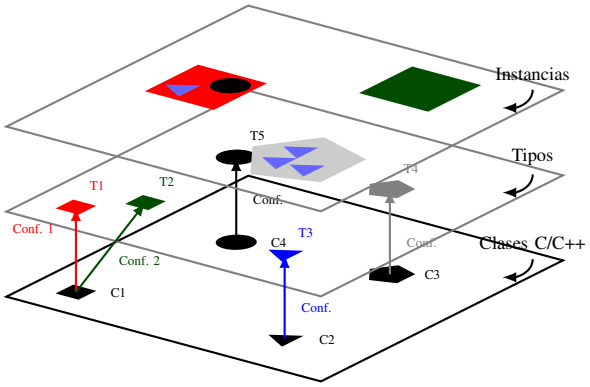


Fig. 1: Modelo de instanciación de GWNSyM

Las *Acciones* representan un modelo particular que se va a aplicar a uno o más conjuntos de *Tipos*. En general, las *Acciones* representan comportamientos del sistema en sentido amplio, pudiendo abarcar desde fenómenos físicos tales como modelos de propagación a políticas concretas, como selección de acceso. Cada *Acción* toma como parámetros uno o más conjuntos de *Tipos*, y se ejecutan de forma secuencial en cada fotografía del escenario. Por otro lado, hay ocasiones en las que una determinada *Acción* únicamente tiene sentido al inicio o final del experimento, como el despliegue de elementos estáticos, por ejemplo. Estos supuestos se han tenido en cuenta definiendo dos categorías de acciones, *Pre-Acción* y *Post-Acción*, que se ejecutan al principio y final del experimento, respectivamente.

B. Metodología de simulación

Una instancia de simulación, o experimento, representa el análisis de un escenario concreto, sujeto a una configuración específica. Cada experimento contiene dos bucles: uno exterior y otro interior. El primero realiza interacciones fotografía a fotografía, actualizando el estado de la red de acuerdo al resultado de la iteración (fotografía) anterior. Por su parte, el segundo bucle se encarga de aplicar los modelos correspondientes a los elementos de red dentro de una fotografía.

El Algoritmo 1 ilustra la metodología general. Como primer paso, se definen los *Tipos* y se crean sus instancias, dando lugar a los conjuntos previamente mencionados, de acuerdo a su configuración. Seguidamente, una vez que todos los elementos del sistema han sido instanciados, se ejecutan las acciones definidas como *Pre-Acciones*. Llegado este punto, se inicia el primer bucle, línea 13, que itera sobre cada fotografía y el bucle interior, línea 14, que se encarga de ejecutar las acciones de forma secuencial. Tras finalizar ambos bucles, se aplican las *Post-Acciones*, normalmente encargadas de extraer resultados y generar trazas.

Algoritmo 1 Flujo general de simulación

1: Definición de *Tipos*
2: Configuración
3: Instanciación y agregación
4: $T \leftarrow$ *Conjuntos*
5: $A_{pre} \leftarrow$ *Pre-Acciones*
6: $A \leftarrow$ *Acciones*
7: $A_{post} \leftarrow$ *Post-Acciones*
8: $i = 0$
9: $n \leftarrow$ *# Fotografías*
10: **for** $b \in A_{pre}$ **do**
11: Ejecutar pre-Acción $b(M_b \subseteq T)$
12: **end for**
13: **while** $i < n$ **do**
14: **for** $a \in A$ **do**
15: Ejecutar Acciónn $s(M_s \subseteq T)$
16: **end for**
17: **end while**
18: **for** $e \in A_{post}$ **do**
19: Ejecutar post-Acción $e(M_e \subseteq T)$
20: **end for**

C. Ejemplo de definición de un escenario

A modo de ejemplo, la Figura 2 muestra los principales pasos de la definición de un escenario por medio de código. Como se puede observar en la Figura 2a, el primer paso consiste en el registro de tipos dentro del sistema. Los tipos consisten en dos objectos `C++`: el elemento de red y su configuración. Por ejemplo, el tipo *USER* se define en base a los objetos `User` y `UserConf`.

Antes de instanciar los elementos de cada tipo, el sistema comprueba aquellos que debe agregar, a través del nombre del tipo correspondiente. Como se muestra en la Figura2a, la configuración del tipo `USER` indica que

```
gnsm::System system;
...
system.AddType<User, UserConf>("USER");
system.AddType<LteUe, LteUeConf>("LTE_UE", {Params});

system.AddType<LteCell, LteCellConf>("CELL", {Params});
system.AddType<LteEnb, LteEnbConf>("MACRO", {Params});
...
```

(a) Defninición de *Tipos*

```
UserConf::ReadInnerConf(void) const
{
   return{{"LTE_UE", 1}}; // read from configuration
}
```

(b) Agregación

```
system.PreAction<MacroDeploymentr>({"MACRO"}, {Params.});
...
system.Action<LteScan>({"USER", "MACRO"}, {Params.});
...
system.PostAction<MacroLoad>({"MACRO::*::CELL"}, {Params.});
...
system.Run();
```

(c) Acciones

Fig. 2: Ejemplo de definición de escenario

agrega una instancia del tipo `LTE_UE`, de modo que se instancia un elemento del tipo `LTE_UE` por cada elemento del tipo `USER`.

Una vez establecidos los tipos, se registran las acciones que definen el comportamiento del sistema, tal y como se muestra en la Figura 2c. Es importante destacar que se ha dotado al entorno de simulación de un módulo que permite la búsqueda de los elementos instanciados, de forma que el paso de conjuntos de elementos a las acciones se lleva a cabo de una manera muy flexible. Por ejemplo, `MACRO::*::CELL` indicaría el conjunto de elementos `CELL` agregados en todas los elementos `MACRO`.

III. ESCENARIO Y EVALUACIÓN

A fin de validar el funcionamiento del entorno de simulación, esta sección presenta los principales resultados del estudio de diferentes técnicas de selección de acceso en un escenario LTE heterogéneo. En particular, se centra en la mejora en términos de potencia, y por tanto recursos, en el enlace ascendente, al aplicar diferentes políticas. Si bien la mayoría de los estudios se han centrado tradicionalmente en la gestión de recursos en el enlace descendente, la evolución de los servicios hace cada vez más necesario el prestar más atención a los recursos usados en la comunicación desde los usuarios a las estaciones base.

En las redes celulares un usuario se conecta habitualmente, tanto en el enlace ascendente como en el descendente, a aquella estación base de la que recibe una mejor calidad de la señal: en el caso concreto de redes LTE a la estación de la que recibe mayor *Reference Signal Received Power* (RSRP). Sin embargo, en entornos heterogéneos en los que co-existen estaciones base con un alto número de recursos y potencia de transmisión (*macro-cells*) con otras de capacidad más reducida (*small cells*), la selección de acceso basada en *RSRP* tiende a asociar a la mayor parte de usuarios con las que tienen una mayor potencia de transmisión, no permitiendo aprovechar el aumento de

capacidad que aportan las *small cells*. Además, dado que el parámetro *RSRP* únicamente tiene en cuenta la calidad del enlace descendente, la potencia necesaria para transmitir desde los usuarios a las estaciones podría estar lejos de ser óptimo. Se debe tener en cuenta que la potencia recibida de las *macro-cells* puede ser mucho mayor que la de las *small cells*, incluso si la primera está situada notablemente más lejos del usuario.

A fin de establecer políticas de selección de acceso que aprovechen de forma más eficiente los recursos de las redes heterogéneas, se propuso el uso de técnicas *Cell Range Extension* (CRE) [14], que se basan en incrementar sintéticamente la potencia de la señal de referencia de las *small-cells* en un valor fijado (o bias), favoreciendo así las asociaciones a estos elementos de acceso. Además, teniendo en cuenta el incremento de los recursos consumidos en el enlace ascendente como consecuencia de la aparición de nuevos servicios, surge la necesidad de realizar una gestión más óptima de los recursos en el enlace ascendente. En este sentido, se ha propuesto recientemente una estrategia de selección de acceso que independiza los enlaces ascendente y descendente, conocida como *Downlink Uplink Decouplig* (DUDe) [15], de forma que cada usuario se asocie a aquel elemento que minimice los recursos necesarios para la transmisión. A continuación se presentarán los principales resultados del análisis de estas soluciones ante diferentes configuraciones.

El escenario que se ha utilizado durante el análisis está descrito en la Tabla II, que indica la topología del despliegue y los modelos de propagación implementados, que son los definidos en las recomendaciones del 3GPP. Se contemplan 7 celdas, desplegadas siguiendo un patrón hexagonal, y un número variable de *small-cells* en la zona de de las *macro-cells* central. Sobre este escenario se han desplegado varios usuarios, y se han aplicado las diferentes técnicas de selección: (1) RSRP, (2) técnicas CRE con diferentes valores de bias, y (3) DUDE.

Como se ha comentado anteriormente, este análisis se centra en el efecto que tienen las diferentes soluciones de selección de acceso sobre la potencia de transmisión necesaria en el enlace ascendente. Para ello se ha asumido que todos los usuarios requieren un valor de *Signal to Noise plus Interference Ratio* (SINR) fijo de 5 dB, y se ha considerado que el sistema se encuentra en condiciones de saturación. Bajo estas premisas, se calcula la potencia de transmisión necesaria para que cada usuario alcance la SINR objetivo, de acuerdo al modelo de control de potencia en lazo abierto definido en la Ecuación 1 [16]:

$$P_{tx}[dBm] = \min P_{max}, P_0 + 10\log_{10}(N_{RB}) + \alpha L \quad (1)$$

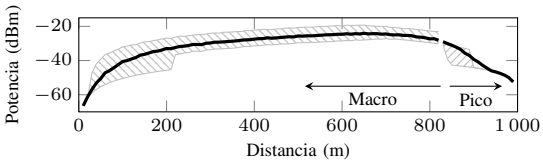
donde P_{tx} y P_{max} representan la potencia transmitida, y su valor máximo, respectivamente; P_0 indica la potencia que se transmite por unidad de recurso, N_{RB} el número de recursos necesarios para transmitir, L las pérdidas de propagación y α es el factor de compensación de propagación, que se ha fijado en 0.4. Los parámetros P_{max} y P_0 se han fijado a 24 y -80 dBm, respectivamente.

Tabla II: Configuración de la simulación

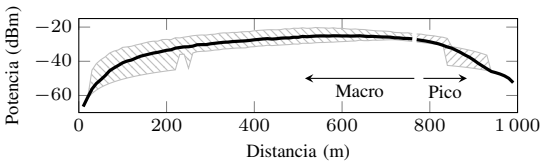
despliegue LTE	**FDD 2x20 MHz @2.1 GHz**
Capa Macro	ISD 500 m, 7 tri-sector sites Max. tx. power 46 dBm Ganancia de antena 15dBi, 15 down-tilt
Capa Small	Despliegue aleatorio Max. potencia transmisión 37dBm Omni-antenna
UE	DL NF 7dB Rx. Gain 7dB Max. Potencia transmisión
Despliegue LTE	L (dB) función de la distancia d $[m]$
$\text{Macro}_{\textbf{NLOS}}$	$139.1033 + 39.0864 * (log_{10}(d) - 3)$
$\text{Macro}_{\text{LOS}}$	$36.2995 + 22 * log_{10}(d)$ if $d < 328.42$ $40 * \log_{10}(d) - 10.7953$ if $d > 328.42$
$\text{Small}_{\textbf{NLOS}}$	$145.48 + 37.5 * (\log_{10}(d) - 3)$
$\text{Small}_{\textbf{LOS}}$	$103.8 + 20.9 * (\log_{10}(d) - 3)$
	Probabilidad LOS función de la distancia d $[m]$
Macro	$P_{LOS} = \min(\frac{18}{d}, 1) \cdot (1 - e^{\frac{-d}{36}}) + e^{\frac{-d}{36}}$
Small	$P_{LOS} = 0.5 - \min(0.5, 5 \cdot e^{\frac{-156}{d}}) + \min(0.5, 5 \cdot e^{(\frac{-d}{30})})$

Antes de analizar el comportamiento global de las diferentes estrategias de acceso, se ha estudiado la mejora potencial que se puede obtener en un escenario sencillo, en el que se sitúa una estación base de cada tipo, *macro* y *small*, con una separación de 1000 m entre ellas. Se asume que un usuario se irá desplazando en la línea recta que une ambos elementos de acceso, para caracterizar la potencia necesaria en el *uplink* en función de la posición en la que se encuentre. La Figura 3 muestra la potencia de transmisión en función de la distancia a la estación base *macro*. Cada punto representa el promedio de 1000 experimentos independientes, de manera que pueden haberse dado situaciones de conexión con cualquiera de las dos estaciones base, así como los percentiles 10 y 90 de dicha potencia de transmisión. Además, se indica el punto en el con mayor probabilidad se cambia de estación base. Como muestra la Figura 3, el uso de técnicas *CRE* supone una notable ventaja en lo que se refiere a la utilización de la *small-cell*, de forma que cuanto mayor es el valor del *bias*, el traspaso de la *macro* a la *small-cell* ocurre antes. Siguiendo esta tendencia, se puede observar que el uso de *DUDe* aporta una mejora sustancial respecto a las otras alternativas, comprobándose que la variación de la potencia transmitida se reduce. De alguna manera, la técnica *DUDe* equivale a usar *CRE* con un valor de *bias* optimizado en cada momento.

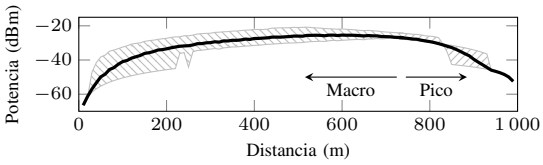
A modo de resumen, la Figura 4 muestra los valores medio y máximo de potencia, utilizando las diferentes soluciones de selección de acceso. La figura muestra además de los valores globales (Max y Mean), los valores relativos a los casos en que se conecta a la estación base *macro* (Max_M y Mean_M) y *small* (Max_S y Mean_S). En general, se puede observar que a medida que aumenta el *bias* de *CRE*, y especialmente con el uso de *DUDe*, la potencia media necesaria al conectarse con la *small-cell* se incrementa, como consecuencia de que la conexión se establece con anterioridad. Sin embargo, también se observa que el hecho de potenciar el uso de las *small-cells* no conlleva una mejora sustancial en la potencia media



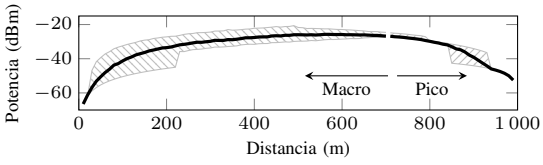
(a) RSRP



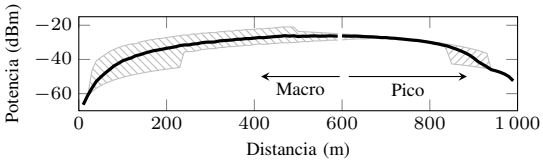
(b) CRE $6dB$



(c) CRE $9dB$



(d) CRE $12dB$



(e) DUDe

Fig. 3: Potencia media de transmisión en el enlace descendente a diferentes distancias de las estaciones base. La superficie sombreada delimita los percentiles del 10 y 90%

transmitida, que se mantiene bastante constante para las diferentes estrategias de selección de acceso.

Finalmente se ha estudiado el rendimiento global de las diferentes técnicas en el escenario descrito en la Tabla II. Bajo la zona de cobertura de cada una de las *macro-cell* se ha desplegado un número creciente de *small-cells*, y se ha evaluado el comportamiento global de 6000 conexiones. La Figura 5 muestra la potencia media transmitida por cada una de las asociaciones usando las diferentes soluciones de selección de acceso, y para diferentes densidades de *small-cells*. Como se puede observar, al incrementar la presencia de *small-cells* se reduce notablemente la potencia necesaria para transmitir, siendo mucho menos relevante el impacto de las diferentes soluciones de selección de acceso. Además, la Figura 6 presenta la probabilidad de realizar una asociación con una *macro-cell* ante diferentes densidades de *small-cells*.

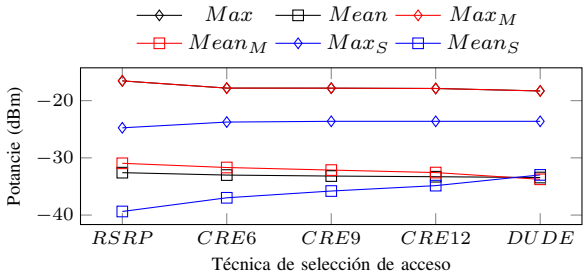


Fig. 4: Resumen de la potencia transmitida en las diferentes configuraciones. El eje de abscisas indica la técnica de acceso

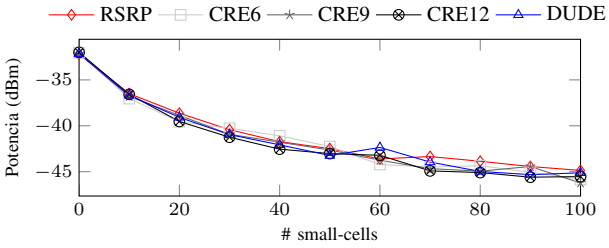


Fig. 5: Potencia media de transmisión ante diferente número de *small-cells* desplegadas

En este caso, sí se puede observar una diferencia notable para las diferentes técnicas de acceso utilizadas, de forma que *DUDe* consigue incrementar considerablemente las conexiones con las *small-cells*, que incluso se llegan a duplicar, si se comparan con *RSRP*. Como es de esperar, esta diferencia se mitiga a medida que la densidad de *small-cells* aumenta.

IV. CONCLUSIONES

Aunque la eclosión de las nuevas tecnologías ha ampliado de manera considerable las posibilidades de las redes celulares, su estudio se ha vuelto notablemente más complejo. En este sentido, la comunidad investigadora no se limita a analizar nuevas técnicas y soluciones que mejoren la calidad de los servicios, sino que es necesario evaluar las nuevas tendencias que afectan a las redes desde un punto de vista de su arquitectura. Como consecuencia, en este trabajo se ha argumentado que no existe una metodología o entorno de análisis de-facto. Así, aunque

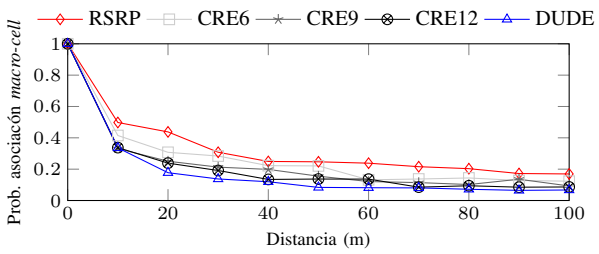


Fig. 6: Probabilidad media de asociación con la *macro-cell*. El eje de abscisas indica el número de *small-cells* desplegadas

existen diversas alternativas, cada una de ellas presenta limitaciones, especialmente desde el punto de vista del impacto de los servicios (y los patrones de tráfico asociados) y de la cada vez mayor complejidad de las topologías.

Con el objetivo de dar respuesta a estas limitaciones, en este trabajo se ha presentado un entorno de simulación alternativo, *Generic Wireless Network System Modeler-GWNSyM*. Su flexbilidad permite la definición de diferentes tipos de escenarios, que pueden analizarse de manera relativamente rápida y sencilla. Esta herramienta ha sido validada mediante el estudio de diferentes técnicas de selección de acceso, sobre un escenario heterogéneo. El análisis se ha centrado en cómo las diferentes técnicas explotan las capacidades adicionales de las *small-cells* y cómo esto afecta a la potencia transmitida en el enlace ascendente, desde el usuario a la estación base. De acuerdo a los resultados obtenidos, las nuevas técnicas, principalmente *DUDe*, son capaces de fomentar las conexiones con las estaciones base de menor capacidad. Sin embargo, mientras que la densificación de la red permite conseguir reducciones notables de la potencia transmitida, no se ha puesto de manifiesto un gran impacto de las diferentes técnicas.

Dada la flexibilidad del entorno de simulación presentado, se pretende aprovecharlo en diferentes ámbitos. En concreto, y en relación al análisis inicial presentado en este trabajo, en el futuro se pretende mejorar el modelado del escenario, teniendo en cuenta la evolución de los servicios, haciendo uso de un patrón de tráfico más acorde con los utilizados en la realidad. Además, se hará uso de herramientas que permitan establecer límites teóricos al rendimiento de las diferentes técnicas de selección, para lo que se integrarán algoritmos de optimización, tal y como ya se ha hecho previamente [17].

AGRADECIMIENTOS

Los autores agradecen la financiación del Gobierno de España (Ministerio de Economía y Competitividad, Fondo Europeo de Desarrollo Regional, FEDER) de este trabajo a través del proyecto ADVICE, Dynamic provisioning of connectivity in high density 5G wireless scenarios (TEC2015- 71329-C2-1-R).

REFERENCIAS

[1] C. and/or its affiliates, "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 white paper," Feb. 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[3] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[4] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5g heterogeneous cloud radio access networks," *IEEE Network*, vol. 29, no. 2, pp. 6–14, March 2015.

[5] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5g wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[6] C. Mehlführer, J. Colom Ikuno, M. Šimko, S. Schwarz, M. Wrulich, and M. Rupp, "The vienna lte simulators - enabling reproducibility in wireless communications research," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–14, 2011. [Online]. Available: http://dx.doi.org/10.1186/1687-6180-2011-29

[7] "The ns-3 network simulator," http://www.nsnam.org/.

[8] G. Piro, N. Baldo, and M. Miozzo, "An lte module for the ns-3 network simulator," in *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques*, ser. SIMUTools '11. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 415–422. [Online]. Available: http://dl.acm.org/citation.cfm?id=2151054.2151129

[9] C. Schneider and R. S. Thomä, "Evaluation of lte link-level performance with closed loop spatial multiplexing in a realistic urban macro environment," in *Antennas and Propagation (EUCAP), 2012 6th European Conference on*, March 2012, pp. 2725–2729.

[10] R. M. Fujimoto, K. Perumalla, A. Park, H. Wu, M. H. Ammar, and G. F. Riley, "Large-scale network simulation: how big? how fast?" in *Modeling, Analysis and Simulation of Computer Telecommunications Systems, 2003. MASCOTS 2003. 11th IEEE/ACM International Symposium on*, Oct 2003, pp. 116–123.

[11] J. Pelkey and G. Riley, "Distributed simulation with mpi in ns-3," in *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques*, ser. SIMUTools '11. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 410–414. [Online]. Available: http://dl.acm.org/citation.cfm?id=2151054.2151128

[12] J. C. Ikuno, *LTE Link- and System-Level Simulation*. John Wiley & Sons, Ltd, 2011, pp. 243–270. [Online]. Available: http://dx.doi.org/10.1002/9781119954705.ch11

[13] M. Taranetz, T. Blazek, T. Kropfreiter, M. K. Müller, S. Schwarz, and M. Rupp, "Runtime precoding: Enabling multipoint transmission in lte-advanced system-level simulations," *IEEE Access*, vol. 3, pp. 725–736, 2015.

[14] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.

[15] F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski, and S. Singh, "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 110–117, March 2016.

[16] S. Berger, B. Almeroth, V. Suryaprakash, P. Zanier, I. Viering, and G. Fettweis, "Dynamic range-aware uplink transmit power control in lte networks: Establishing an operational range for lte's open-loop transmit power control parameters $(alpha, p_0)$," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 521–524, Oct 2014.

[17] L. Diez, G. P. Popescu, and R. Agüero, "A geometric programming solution for the mutual-interference model in hetnets," *IEEE Communications Letters*, vol. 20, no. 9, pp. 1876–1879, Sept 2016.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6584

Multimedia communications in vehicular adhoc networks for several applications in the smart cities

Cristhian Iza Paredes, José Antonio Uribe Ramírez, Nely P. López Márquez, Leticia Lemus, Ahmad M. Mezher, Mónica Aguilar Igartua

Departamento de Ingeniería Telemática,

Universidad Politécnica de Catalunya (UPC)

Campus Nord UPC. Despacho C3-301. C/ Jordi Girona 1-3. 08034 Barcelona.

{cristhian.iza, antonio.uribe, leticia.lemux, nely.lopez, ahmad.mezher, monica.aguilar}@entel.upc.edu

***Resumen*- Road safety applications envisaged for vehicular ad hoc networks (VANETs) depend largely on the exchange of messages to deliver information to concerned vehicles. Safety applications as well as inherent VANET characteristics make data dissemination an essential service and a challenging task. We are developing a decentralized efficient solution for broadcast data dissemination through two game-theoretical mechanisms. Besides, VANETs can also include autonomous vehicles (AVs). AVs might represent a revolutionary new paradigm that can be a reality in our cities in the next few years. AVs do not need a driver to work; instead, they should copy a proper human behavior to adapt the driving according to the current circumstances, such as speed limit, pedestrian crossing street or wheather conditions. We will develop an AV software module including artificial intelligence (AI) techniques so that AVs can interact with the dynamic scenario throughout time. Finally, we also will include electrical vehicles (EV) in the VANET, so that special services such as finding and reserving an EV charging station place will be welcome. In addition, we are developing a multimetric geographic routing protocol for VANETs to transmit H.265 video (traffic accident, traffic state, commercial….) over VANETs.**

***Palabras Clave*- Vehicular Ad hoc NETworks, video dissemination, privacy in VANETs, electrical vehicle (EV), autonomous vehicle (AV), artificial intelligence (AI).**

I. INTRODUCCIÓN

Vehicular ad hoc networks (VANETs) are foreseen as an essential component of the future intelligent transportation systems (ITS) to support safety, traffic management, and user infotainment applications. VANETs are a type of mobile ad-hoc networks (MANETs) forming self-organized networks without the requirement of permanent infrastructure, in which nodes are vehicles. Thus, the network topology might change rapidly due to the high mobility of nodes (vehicles), which makes communications a challenge.

VANET vehicles can stablish vehicle-to-vehicle (V2V) communications to share information, as well as vehicle-to-infrastructure (V2I) communications for other kind of services. VANETs are intented for safe-applications, traffic management, and enhanced navigation among other services.

Today, road traffic is the ninth biggest cause of death worldwide. By 2030, the increase in the number of vehicles will see road traffic become the fifth largest cause of death. One of the most important uses of VANETs are safety applications, which usually rely on broadcast-based algorithms. Flooding-based algorithms can be used to disseminate emergency messages rapidly and efficiently through the VANET to warn vehicles around *e.g.* an accident. Thus, a key research problem is the design of a scalable dissemination scheme, which is efficient, reliable and incurres short delay under different VANET conditions. To tackle this issue, we have developed an Adaptive Distributed Dissemination (ADD) [1] protocol, based on game-theoretical algorithms to compute the forwarding probability of nodes based on several design parameters.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

In the considered VANETs, we can find autonomous vehicles (AVs) and electrical vehicles (EV), which have special features and necessities that we will consider in our simulated scenarios. For instance, AVs must communicate not only with other vehicles through the VANET, but also with the infrastructure in the city (*e.g.* traffic lights), with pedestrian crossing the streets and with any possible obstacle around. In addition, AVs must impersonate the human behaviour to take proper driving decisions (*e.g.* brake, accelerate, stop…). To do so, we will include artificial intelligence (AI) algorithms in the AV module. On the other hand, EVs have special needs. For instance they require electric vehicle charging stations (EVCS) to recharge their batteries during their trip if necessary. Thus, it would be basic to develop a smart service so that EVs could reserve an electrical recharging point in special EVCS deployed in the city.

Another goal of our research team is the development of a multimetric geographical routing protocol for VANETs to transmit H.265 video. Possible applications include: (a) the transmission of a short video clip about an accident to alert quickly the smart emergencies services (*e.g.* 112 or 911) in the city; (b) a light video to report the state of the traffic to the traffic management unit in the city; (c) the transmission of a short light video advertisement for the passengers about restaurants or shops discounts around, among other examples. Furthermore, sometimes, it is not necessary to send the exact geographic location (*e.g.* GPS location) of the vehicle in order to keep the privacy of the user (e.g. not to be tracked). To do so, we will apply microagregation techniques [8].

II. An Adaptive Game-theoretical Distributed Dissemination Protocol for VANETs

Road safety applications for VANETs rely on the dissemination of warning messages to deliver information to vehicles near *e.g.* an accident. We have proposed an Adaptive Distributed Dissemination (ADD) protocol [1] to perform data dissemination in VANETs. ADD is designed to operate without any roadside infrastructure in urban scenarios under diverse road traffic conditions. ADD uses a decentralized solution for the broadcast data dissemination problem through two game-theoretical mechanisms. Using game theory (GT), we have designed a mechanism to predict behavior in situations where a state is the result of a series of interactions between different nodes (players), who act according to their preferences regarding future performance and existing incentives. In first place, the Asymmetric Volunteers Dilemma Game is evaluated as a mechanism to tackle the broadcast storm problem. The probability that a node forwards a broadcast message is calculated using the number of candidate vehicles to forward the message, *i.e.*, the number of vehicles that are listening to the transmission. The cost/benefit relation to forward the message by a vehicle is obtained from metrics like distance (to an intersection, from the incident, to an access point…), average packet delay and link quality. Also, the Forwarding Game is designed as a second mechanism to mitigate the broadcast storm problem. In this case, the strategy of the players consists on selecting a forwarding probability that maximizes the pay-off using a utility function, which depends on the player's availability and the forwarding probability of other players. Availability of a player is a normalized factor based on several metrics. Our proposal ADD considers the position of the vehicle in the network (distance between receiver to next junction, distance between transmitter and receiver), an estimation of the link quality (derived from signal quality, channel quality and collision probability), time elapsed since the generation of the message and estimated available bandwidth. Nodes themselves calculate those metrics and broadcast them in periodic beacons to their neighbours.

Extensive simulations reveal that our proposal excels in terms of number of transmissions, lower end-to-end delay and reduced network load while maintaining high delivery ratio, compared to other proposals.

III. Multimetric QoS-aware Geographic Routing Protocol for VANETs

We start from the well-known GPSR (Greedy Perimeter Stateless Routing) [7], which uses the position of the nodes to take packet forwarding decisions. GPSR uses greedy forwarding to forward packets to nodes that are always progressively closer to destination. In regions of the network where such a greedy path does not exist (*i.e.*, the only path requires that one move temporarily farther away from the destination), GPSR recovers by forwarding in perimeter mode, in which a packet traverses successively closer faces of a planar subgraph of the full radio network connectivity graph, until reaching a node closer to destination.

Our multimetric geographical routing protocol for VANETs, takes into account several metrics to take the best hop-by-hop forwarding decisions under the current circumstances. We consider distance to destination, vehicle's velocity, vehicles' trajectory, vehicles' density, available bandwidth and packet losses. The design of the metrics is detailed in [9] and here we will just summarize one of them, the vehicles' density.

Vehicles' density: It is computed as the number of vehicles (N_V) in the neighbors' list of each node at the moment of sending the current hello message, divided by the area within the transmission range ($\pi.TR^2$) of that vehicle, being *TR* the transmission range. Each node computes its density of nodes ρ_{Ngh} using Eq. (1) and includes it in the next hello message.

$$\rho_{Ngh} = \frac{N_V}{\pi.TR^2} \quad (1)$$

$$u_{dns,Ngh} = \begin{cases} \frac{-1}{\rho_{max}{}^2}\rho_{Ngh}{}^2 + \frac{2}{\rho_{max}}\rho_{Ngh} & \text{if } \rho_{Ngh} \le 2\rho_{max} \\ 0, & \text{if } \rho_{Ngh} > 2\rho_{max} \end{cases} \quad (2)$$

Neighbors' lists are updated upon reception of new hello messages, which are sent by nodes (usually one per second) within their neighbourhood. The algorithm gives a higher score when the neighbor node *Ngh* has a higher value of ρ_{Ngh}. Nodes with a denser area in the transmission range will have more possibilities to forward the packet to a next node. This is true until a maximum nodes' density ρ_{max}, above which there are too many vehicles and the collisions' probability increases. We set ρ_{max} to 200 vehicles/km^2.

We have designed a concave function for the density metric shown in Fig. 1. This function has its maximum at $\rho_{Ngh} = \rho_{max}$ and above ρ_{max} it decreases till $2\rho_{max}$ where again it reaches zero and keeps on zero for $\rho_{Ngh} > 2\rho_{max}$.

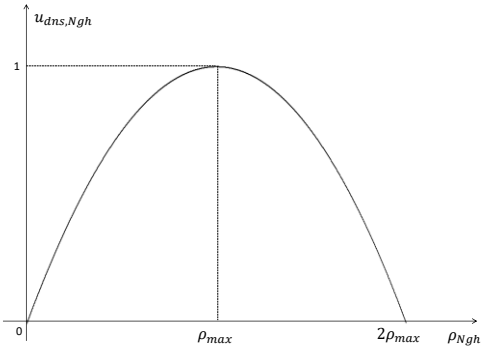


Fig. 1. Designed metric for the vehicles' density, $u_{dst,Ngh}$.

Eq. (2) describes how we calculate the vehicles' density metric $u_{dns,Ngh}$ ($0 \leq u_{dns,Ngh} \leq 1$) as shown in Fig. 8. This way, we penalize those nodes whose number of neighbors in their transmission range is above a threshold ($\rho_{Ngh} > \rho_{max}$).

Furthermore, we will include machine learning (ML) techniques to take best forwarding decisions at each moment by weighting all considered metrics with proper dynamic weights. The idea is to test machine learning models (*e.g.*, trees, logistic regression, KNN) to see which one predicts better (higher accuracy). This way, the performance evaluation of our proposal will notably improve the quality of service (QoS) offered by other similar routing protocols.

IV. Machine Learning and Privacy in VANETs

Machine learning (ML) is a scientific discipline in the field of artificial intelligence (AI) that uses systems that learn automatically. Learning in this context means identifying complex patterns in tones of data. The machine that learns is an algorithm that checks data and is able to predict future behavior. ML systems improve autonomously over time without human intervention

The use of machine learning in VANETs is an idea to improve the performance of this kind of networks. Specifically, we plan to use principal component analysis (PCA), which is a technique used to emphasize variations and extract patterns from a dataset. The first step is to apply a ML technique offline. Then, use PCA online to make a self-learning to choose the best routes to transfer the information, trying to decrease the packet losses in the network. PCA will provide in an offline analysis the distribution of energies that each metric represents. Thus, we could estimate the correct weights of each one of the considered metrics. After that, and by recalculating the correct weights instead of giving equal weights, best forwarding routes will be chosen. To improve the results, we will test different ML algorithms to find the one that performs better in VANETs.

On the other hand, privacy is a right to be free from secret surveillance and to determine whether, when, how, and to whom, one's personal information is to be revealed. ML techniques allow us to hide important information, although keeping the proper functioning of the network. We claim that vehicles have the right to choose what information want to share or want to make public, so in this scenario ML is a good tool that can help VANETs to keep information secure.

V. Realistic Simulation Platform

In order to attain realistic simulations to trust the obtained results from our proposals, it has paramount importance to prepare a realistic simulation platform. In our case, we use OMNeT++ [2] to perform the simulations and SUMO [3] to generate the vehicular movement traces. OMNeT++ provides a baseline to develop different type of projects which implement the actual simulators. Two of these projects can be used together to provide a vehicular network simulator, INET [4] and VEINS [5]. For a more realistic mobility behavior, we defined a scenario including different type of vehicles (car, bus and truck) with an associated probability of occurrence. Besides, seeking to dispose a scenario prepared as much realistic as possible, we use real maps extracted from OpenStreetMap [6]. This simulation platform allows us to include realistic features of real world maps and to interact with realistic vehicular mobility models.

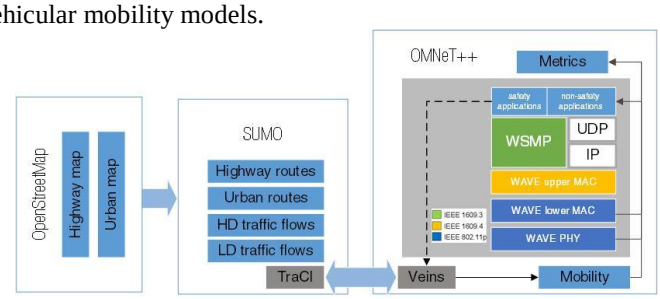


Fig. 2. Simulation framework

The structure of the simulation platform is shown in Fig. 2. VEINS provides a comprehensive set of simulation models of IEEE 802.11p and IEEE 1609.4 DSRC/WAVE network layers. Bidirectional coupling to a road traffic simulator is achieved by a node mobility model that interfaces with a running instance of a TraCI server such as SUMO. These models, together with the OMNeT++ simulation environment, provide the platform in which custom simulation models of protocols can be integrated.

VI. VANETs and Autonomous Vehicles

Autonomous vehicles (AV) are a promising driverless type of vehicle that can be part of the VANETs in the near future. AV need to communicate with others cars (V2V) and with the infrastructure around (V2I), in order to move in the roads interacting with the environment without problems, copying the human behaviour, adapting the driving according to the circumstances such as speed limit, pedestrian crossing the street or water in the road. Both kind of communications are necessary to detect pedestrians or obstacles (localization), movements of other cars (planning) to take a decision of what to do (execution).

Fig. 3. Autonomous Driverless cars and people walking on the street. *https://www.123rf.com*

One major challenge most cities share is to find efficient ways to manage mobility. According to a study by Texas A&M Transportation Institute in 2015 the time US commuters are stuck in traffic has risen by 133% since 1982 which equals a total of 42 hours. This means that on average drivers spend almost two full days in their vehicles per year. On average drivers loose an extra of 72 litres of fuel per year during traffic jams. In total, fuel emissions have gone up by 520% since the 1980s, which is a huge strain on the environment.

On the other hand, as more and more people move to city outskirts, traffic congestion during rush hours is likely to become cities primer challenge. While most commuters drive into the city centre by car to get to work, most cities try to manage this problem by introducing traffic management systems and restrictive policies to regulate cars accessing the centre.

As traffic density increases, managing traffic and congestion will become more complex. In this context, AV and VANETs could alleviate drastically many issues related to mobility in cities, improving driving safety, decreasing pollution and reducing traffic congestion.

VII. VANETs, Electrical Vehicle and Smart Road Ligting

Finally, we are working on a global urban environment involving drivers, pedestrians and smart lighting of the roads. Looking for energy savings and reducing light pollution in the sky, we plan to link vehicle mobility behavior and road lighting. By modeling the mobility behavior of vehicles in the roads (either highways and city streets), road lighting could adapt to the actual lighting necessities throughout time, along the day and the week.

Besides, we also consider the presence of electrical vehicles (EVs) in the VANET. EVs have special needs, such as the reservation of an EVCS point during the required time to recharge their battery. An updated accurate knowledge of the mobility state in the city and the available EVCS points can help to manage better those reservations.

Fig. 4. Electrical vehicle charging station in a city. *https://cdn.autocentre.ua/images/stories/2014/10/08/b/general-motors-planiruet-rasshirit-lineyku-eko-modeley-4.jpg*

VIII. Conclusions

In this article, we have presented several research tasks that we are developing in our research team. The common topic is the design of algorithms, protocols and a realistic simulation platform to contribute in the development of vehicular adhoc networks (VANETs) in smart urban scenarios, focusing on video-streaming services with QoS provision.

Acknowledgements

This work was partly supported by the Spanish Government through the project TEC2014-54335-C4-1-R INcident monitoRing In Smart COmmunities, QoS and Privacy (INRISCO). Cristian Iza is recipient of a grant from Secretaria Nacional de Educación Superior, Ciencia y Tecnología SENESCYT. Ahmad Mohamad Mezher is a postdoctoral researcher with the Information Security Group (ISG) at the Universitat Politècnica de Catalunya (UPC).

References

[1] Cristhian Iza, Ahmad M. Mezher, Mónica Aguilar, "An Adaptive Game-theoretical Distributed Dissemination Protocol for VANETs", Sensor Networks, *In preparation*.
[2] Omnet++, discrete event simulator. http://www.omnetpp.org
[3] Sumo – simulation of urban mobility. http://goo.gl/uvvD4N
[4] Inet framework. https://inet.omnetpp.org/
[5] Veins, vehicular network simulations. http://veins.car2x.org
[6] OpenStreetMap. http://www.openstreetmap.org
[7] Karp, B. and Kung, H.T. "GPSR: Greedy Perimeter Stateless Routing for wireless networks," MobiCom 2000.
[8] Agusti Solanas, Úrsula Gonzalez-Nicohis, Antoni Martinez-Ballesté, "A Variable-MDAV-Based Partitioning Strategy to Continuous Multivariate Microaggregation with Genetic Algorithms", IJCNN 2010.
[9] Ahmad Mohamad Mezher and Mónica Aguilar Igartua, "Multimedia Multimetric Map-aware Routing protocol to send video-reporting messages over VANETs in smart cities", IEEE Transactions on Vehicular Technology, June 2017, DOI: 10.1109/TVT.2017.2715719

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6585

Implementación de mecanismos de mitigación de tormentas de broadcast en redes de área local mediante Redes Definidas por Software

Bárbara Valera Muros, Jonathan Prados Garzón, Juan José Ramos Muñoz, Jorge Navarro Ortiz
Departamento de Teoría de la Señal, Telemática y Comunicaciones,
Universidad de Granada
Calle Periodista Daniel Saucedo Aranda s/n, E-18071 (Granada)
bvaleramu@gmail.com, jpg@ugr.es, jjramos@ugr.es, jorgenavarro@ugr.es

***Resumen*—El uso de Ethernet como tecnología de red para redes corporativas se justifica por su bajo coste y facilidad de configuración y mantenimiento. Sin embargo, estas redes no son muy escalables, debido en parte a las inundaciones o tormentas de broadcasts, que afectan al rendimiento tanto de los dispositivos de red como finales. Para mitigar el impacto de las inundaciones por broadcast, se ha previsto utilizar técnicas de filtrado y caché en distintos nodos de la red. Sin embargo, el paradigma de Redes Definidas por Software permite definir nuevas aproximaciones, gracias a la capacidad de reprogramar la red de forma centralizada y flexible que proporciona. En este trabajo se aborda la implementación de una red de área local con soporte para filtrar algunos paquetes broadcast mediante la utilización de Redes Definidas por Software. Esta solución permitiría desplegar redes de área local más amplias, adecuadas para los requisitos de redes corporativas. Para ello, se describe el desarrollo de filtros para varios protocolos de red, su implementación en el controlador OpendayLight, y la evaluación del rendimiento obtenido.**

***Palabras Clave*—Address Resolution Protocol, ARP, Broadcast, Controlador, Filtrado, Internet Control Message Protocol version 6, OpenDayLight, Redes Defindas por Software, SDN**

I. INTRODUCCIÓN

Los servicios de datos móviles se han convertido poco a poco en imprescindibles para la mayoría de usuarios. Esta tendencia supone un incremento del tráfico en las redes inalámbricas. Dicho incremento representa uno de los mayores retos a los que cualquier red de comunicación tendrá que enfrentarse en el futuro [1]. A esto se le sumaría la reducción de la latencia y los costes asociados, de forma que se plantean nuevos diseños y arquitecturas de red con el fin de satisfacer las crecientes exigencias de futuras aplicaciones. Este nuevo paradigma surge como una de las posibilidades para suplir esa creciente demanda, considerándose una opción dinámica, gestionable, económica y adaptable. Además, reduce los costes asociados a las redes, conocidos como CAPital EXpenditures (CAPEX) y OPerating EXpense (OPEX), lo que se suma a las ventajas de utilizar esta arquitectura a la hora de renovar las redes de comunicación [2]. El aprovechamiento de las redes es un sector en el que se ha invertido gran cantidad de recursos; pero los requerimientos de las telecomunicaciones son cada vez mayores, por lo que un salto de generación móvil implicaría un cambio completo de la red que suponga una solución definitiva y no temporal, como se ha hecho hasta ahora.

Este artículo presenta el desarrollo de procedimientos para redes SDN (Software Defined Networks) que permiten reducir el problema de los broadcasts para mejorar la escalabilidad de las redes. La sistemática seguida en el proyecto se inicia con el estudio bibliográfico de las tecnologías implicadas, principalmente SDN, y una familiarización con las herramientas necesarias para trabajar. Entre ellas destacan: el protocolo OpenFlow [3], el controlador OpenDayLight (ODL) [4] y el emulador de redes Mininet [5]. Una vez definidos los escenarios sobre los que aplicar la solución, se diseña un algoritmo para la detección e identificación, filtrado y reenvío de paquetes en el controlador SDN. Posteriormente, se realiza la programación del código a partir de un conmutador inteligente. Por último, se evalúa la solución implementada teniendo en cuenta las mejoras en el rendimiento del sistema, así como las posibles vías de aplicación futuras. El artículo se estructura en seis secciones. La Sección II describe el problema abordado y la visión general de la solución. La Sección III presenta la revisión del estado del arte de las tecnologías implicadas. El grueso del artículo se incluye en la Sección IV, que contiene las fases de diseño e implementación de la solución. La sección V plantea diferentes entornos experimentales en los que comprobar el funcionamiento de la solución, describiendo las pruebas realizadas y los resultados obtenidos. Por último, la sección VI presenta

las conclusiones y la proyección de futuro.

II. MOTIVACIÓN

A la hora de definir la arquitectura de las redes 5G, existen propuestas en las que se presentan propuestas de arquitecturas basadas en Ethernet, para aprovechar la capacidad de autoconfiguración, la existencia de hardware que lo implementa, y la sencillez en la gestión de este protocolo [6][7]. Concretamente, la arquitectura [6] está optimizada para el protocolo IPv6, y donde SDN es la clave para resolver los retos previstos. Con esta visión se pretende eliminar parte de la complejidad del Evolved Packet Core (EPC) de la red 5G, de forma que sea más escalable y eficiente. Sin embargo, la escalabilidad de las redes Ethernet está limitada por las inundaciones de red (o tormentas de broadcast) [8], causadas por los protocolos de arranque que utilizan los usuarios finales, como Address Resolution Protocol (ARP) [9] o Dynamic Host Configuration Protocol (DHCP) [10]. Para superar estas limitaciones, las nuevas arquitecturas basadas en Ethernet tratan de reducir las inundaciones de las tramas broadcast a la vez que ofrecen los servicios Ethernet esperados. Las tormentas de broadcasts o difusión se producen cuando varios dispositivos envían paquetes a la dirección de difusión de la red. Este fenómeno no sólo consume recursos de red, sino que que afecta al rendimiento de los dispositivos. Por otra parte, StateLess Address AutoConfiguration (SLAAC) [11] permite la autoconfiguración de los hosts IPv6, lo que a su vez supone multitud de solicitudes multicast.

Así, el principal problema tratado es la disminución de la eficiencia de estos sistemas al producirse una inundación. Como solución, se propone el filtrado de mensajes, de forma que el controlador disponga de unas tablas identificativas para cada nodo y sea el encargado de reenviar los mensajes, dirigiéndolos a un destinatario limitado directamente y evitando las inundaciones de red siempre que sea posible. El controlador actuará como un conmutador inteligente capaz de aprender no solo las direcciones de los nodos implicados en el intercambio de mensajes, sino también su papel en dicho intercambio, identificando los diferentes agentes que intervienen para posteriormente dirigir los mensajes del protocolo en cuestión y evitar la sobrecarga de la red, mejorando en definitiva la eficiencia del sistema.

III. ESTADO DEL ARTE

A. Tormentas de Broadcast

Como ya se ha mencionado, el principal problema en cuanto a escalabilidad de las redes viene dado por las tormentas de broadcast. En términos generales, esta difusión amplia es una forma de distribución de información en la que un nodo envía un mensaje a todos los nodos de la red de manera simultánea. Es utilizado principalmente por los protocolos de arranque y de configuración y disminuye la eficiencia de los sistemas aumentando el tráfico de la red. Además, ya que no solo aumenta el tráfico sino el número de paquetes recibidos por cada terminal, reduce también el rendimiento de los equipos. En [8] se estudia el efecto de una tormenta de broadcast en hosts de redes IP. Se trata de un experimento realizado por Cisco para medir el efecto de estas tormentas en una estación SPARC con una tarjeta estándar de Ethernet. Se demostró que una estación de trabajo puede dejar de funcionar debido a las inundaciones broadcast de la red. Además, se observaron puntualmente picos de miles de broadcast por segundo en las tormentas de broadcast, lo que supone una disminución del rendimiento del sistema de hasta el 25%. Se plantean nuevos diseños para mitigar estos problemas, entre los que cabe destacar Ethane [12] y SEATTLE [13]. En el caso de Ethane, se presenta la posibilidad de que las tormentas de broadcast sean gestionadas por un controlador, mientras que SEATTLE propone una transformación de los mensajes broadcast en unicast, aunque esto requeriría unos conmutadores específicos muy costosos. Se concluye que SDN es una opción viable frente a estos diseños para implementar una solución que supla la demanda de los usuarios.

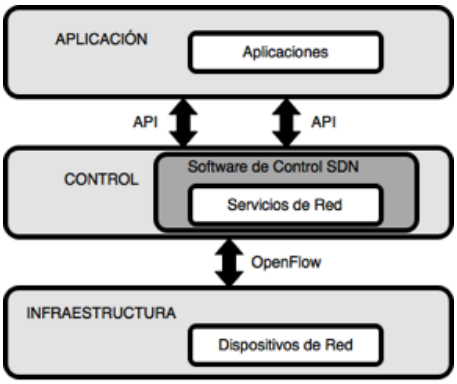


Fig. 1. Arquitectura simplificada de SDN.

Respecto a los protocolos que mayor cantidad de tráfico generan, destacan ARP e Internet Control Message Protocol version 6 (ICMPv6) [14], por lo que el diseño de la solución aborda los procedimientos de autoconfiguración de IPv6 y los procedimientos de resolución de direcciones IP y físicas en IPv4 [15] e IPv6 [16].

B. Software Defined Networking

Las redes SDN permiten atender las necesidades de las aplicaciones y servicios de la red de forma dinámica y escalable. En SDN, la red se programa de forma centralizada, con un controlador lógico que gobierna el funcinamiento de los distintos conmutadores SDN. Esto permite que la red se adapte al entorno con la posibilidad de utilizar el conocimiento de la red completa. En estas redes, el controlador actúa como "cerebro" encargado de comunicar a los conmutadores de la red qué deben hacer con cada flujo de paquetes nuevo. Se plantea el concepto de separación del plano de control de red (software) y del plano de datos (hardware que conmuta los paquetes de datos en la red), tal y como muestra Fig. 1 descrita en [17].

En las redes SDN, las aplicaciones de red usarán la

interfaz de programación (Application Programming Interface, API) NorthBound sobre el plano de control para reforzar sus principios en el plano de datos sin interactuar con el mismo directamente. La interfaz entre el plano de control y el de datos se apoya en SouthBound APIs, que permiten al controlador SDN comunicarse con los equipos de la red en el plano de datos [18]. Dichos equipos deberán soportar las APIs estandarizadas en este nivel. SDN posibilita así la administración de la red al completo a través de sistemas inteligentes que permitan la asignación de recursos según la demanda, redes virtualizadas o servicios cloud seguros. Por tanto, la red estática evoluciona en una plataforma de servicio independiente capaz de responder rápidamente a las necesidades de mercado de los usuarios finales, lo que simplifica en gran medida el diseño y las operaciones de red.

C. Protocolo OpenFlow

OpenFlow es un protocolo de comunicaciones diseñado para dirigir el manejo y enrutamiento del tráfico en una red conmutada, en términos de flujos. Un flujo es un grupo de paquetes definido. Dado que se trata de un estándar abierto, se ha convertido en el modelo estándar de implementación de SDN para la gestión de la red. Se ha traducido así como la primera interfaz de comunicaciones definida entre las capas de control y de transporte en esta arquitectura [19]. El diseño del protocolo se apoya sobre tres bases: los conmutadores con soporte para OpenFlow (que encaminan los paquetes), las tablas de flujos instaladas en dichos conmutadores para la gestión del tráfico y el controlador encargado de comunicar a los conmutadores la información necesaria para administrar el tráfico de la red (añadiendo y eliminando flujos) [20]. Así, aunque el conmutador OpenFlow es responsable del reenvío de paquetes, las decisiones de enrutamiento son tomadas por el controlador. Ambos se comunican a través de OpenFlow, que define los mensajes que hacen referencia a los paquetes enviados, recibidos, la identificación de estados y la modificación de las tablas de flujos para el encaminamiento. De esta forma, cuando el conmutador recibe un paquete para el que no tiene entradas en la tabla de flujo, se lo reenvía al controlador, que es el encargado de decidir si el paquete es descartado o si se agrega una entrada en las tablas. Una vez agregado, el conmutador podrá gestionarlo por sí mismo, en caso de volver a recibir un paquete similar. El proceso que determina qué hacer con cada paquete se denomina Pipeline. Básicamente, cada conmutador dispone de varias tablas, con multitud de flujos cada una. Cuando llega un paquete, se busca hacer el llamado emparejamiento o "matching" en el que se compara cada uno de los valores seleccionados como criterio de emparejamiento del paquete recibido, con los de los flujos de la tabla inicial. En caso de no encontrar ningún flujo coincidente, se pasa a la siguiente tabla. En otro caso, se ejecutaría la acción determinada para ese flujo, entre las que se encuentra la de enviar ese paquete a otra tabla de orden superior. En caso de no haber matching con ninguna tabla, se envía el paquete a una tabla "missing", que decide si debe inundar la red con el paquete, mandarlo al controlador o comenzar de nuevo con un matching más flexible.

D. Controlador OpenDayLight

Como controlador, existen diferentes opciones para implementar la solución con soporte para OpenFlow, a destacar NOX (basado en C++), POX y Ryu (basados en Python), que sin embargo suponen una lenta ejecución de la red. OpenDayLight (ODL) es una alternativa de código abierto y robusto que presenta buen rendimiento de ejecución y soporte de producción, ya que se encuentra respaldado, entre otros fabricantes de dispositivos de red, por Cisco [4]. Al desarrollarse sobre Java, su mayor limitación es la complejidad en la creación de aplicaciones, aunque esto lo hace compatible con la mayoría de sistemas operativos. ODL dispone de una capa de abstracción que separa el controlador de los elementos de red, y esta capa puede basarse en APIs o en modelos. Esta última unifica las APIs, proporcionando una mayor abstracción. Además, se utiliza el lenguaje de modelado YANG (Yet Another Next Generation) [21] para la descripción de las estructuras basadas en modelos, lo que simplifica el desarrollo de aplicaciones en el controlador [22][21].

IV. Propuesta

Para afrontar el problema planteado, se realiza un diseño que permite filtrar las inundaciones de paquetes en la red, aumentando su escalabilidad y el rendimiento del sistema. Para ello, se analizan qué procedimientos requieren de envíos broadcast para distintos protocolos de red. Se categorizan dos casos principales: IPv4 e IPv6. En IPv4 se analiza el protocolo ARP, identificando los mensajes "Request" y "Reply" para resolver la asociación de las direcciones IP y físicas de los dispositivos de la red cada vez que inician una conexión. Sin embargo, en IPv6 se tiene en cuenta el procedimento de autoconfiguación de direcciones (SLAAC) de los hosts, lo que supone multitud de peticiones multicast, ya que en esta versión del protocolo no existe broadcast propiamente dicho. Al conectarse un nodo a una red IPv6, se inicia el procedimiento de descubrimiento de vecino (NDP, Neighbor Discovery Protocol) para descubrir la presencia de otros nodos en el mismo enlace [16]. Se envía una solicitud de router (Router Solicitation) de enlace local mediante multicast, para conocer los parámetros de configuración de red. El router responde con un anuncio de router (Router Advertisement). Se realiza también la solicitud (Neighbor Solicitation) y anuncio (Neighbor Advertisement) de nodos correspondiente para comunicarse con el resto de elementos de la red.

El algoritmo de filtrado se detalla en las siguientes subsecciones.

A. Funcionamiento general

El diseño consta de tres fases. En primer lugar se realiza la clasificación e identificación de paquetes para su posterior filtrado, de acuerdo al tipo de protocolo al que pertenece. Posteriormente, según el tipo de mensaje

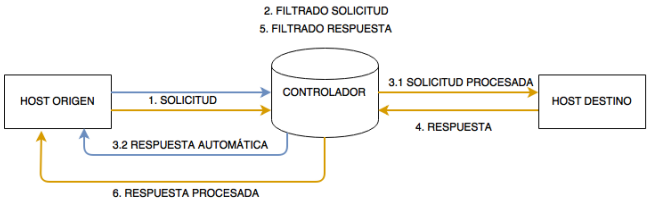


Fig. 2. Esquema básico de la comunicación establecida entre los hosts y el controlador.

que contenga el paquete, se realiza un filtrado en el que el controlador almacena información de los nodos y decide si reenviar una solicitud al resto de la red, o si es capaz de devolver una respuesta apropiada automáticamente. En la fase de reenvío, el controlador realiza el envío, bien de la respuesta directamente, o del paquete a los nodos de la red que sean necesarios.

Se consideran por tanto dos posibles casos: uno, cuando el controlador filtra la solicitud y responde automáticamente. Y otro, cuando se pone en contacto con el nodo destino y se encarga de procesar la respuesta del mismo posteriormente, tal y como muestra la Fig. 2.

B. *Clasificación de paquetes de broadcast*

En la fase de clasificación, se estudia el tipo de trama Ethernet recibida, para determinar qué pasos de filtrado seguir. Además, se extraen las direcciones físicas e IP de origen y destino para almacenarlas en la tabla del controlador en la que se encuentra la información para generar respuestas automáticas para resoluciones de direcciones. La fase de clasificación sigue el diagrama de flujo de Fig. 3.

Posteriormente, si es la trama es un paquete ARP, se comprueba si la entrada está en la tabla y está actualizada. Si se trata de una respuesta, se almacena la carga del paquete para utilizarla posteriormente en las respuestas automáticas. Si se trata de una solicitud, se estudia si el controlador puede o no responder con la información de la que dispone. El diagrama de flujo de esta fase se muestra en Fig. 4.

El caso de datagramas IPv6, el proceso es más complejo. En primer lugar, se comprueba si de trata de un mensaje ICMP y, en ese caso, de qué tipo. Si es una *solicitud de router*, se comprueba si se dispone de una respuesta almacenada para responder automáticamente al nodo solicitante. Si es un *anuncio de router*, se almacena el prefijo de la red con la información necesaria para la configuración del resto de nodos y, en caso de disponer de una entrada previa y actualizada del router, se bloquea el reenvío de este paquete al resto de la red. En la solicitud de vecino se estudia si el nodo origen está en la tabla y se crea una entrada si no dispone ya de una, y posteriormente realiza el mismo estudio para el nodo destino. De esta manera, se comprueba si se puede responder automáticamente dicha solicitud. Los anuncios de vecino crean o actualizan las entradas del controlador y, en caso de ser una entrada actualizada, son bloqueados por el controlador para que no se realice el reenvío al resto de la red. Este proceso se muestra en el diagrama de flujo de Fig. 5.

Una vez se ha filtrado el paquete, se comprueba si la variable de reenvío ha sido modificada por el controlador. De no ser así se realiza un reenvío por defecto al resto de la red, mientras que si se ha editado se realiza el bloqueo de este paquete y se envía la respuesta generada por el controlador directamente al nodo solicitante. La fase de reenvío aparece representada en Fig. 6.

V. Evaluación de la solución

Por último, se realiza la evaluación de los resultados obtenidos mediante una serie de experimentos sobre una plataforma de SDN emulada. Concretamente, se comprueba el funcionamiento de la solución en IPv4, en IPv6, el funcionamiento conjunto de ambos ejecutándose de forma simultánea y concurrente y la mejora de rendimiento que supone la implementación frente a un sistema que no realice el filtrado de mensajes.

Para ello, se utiliza el controlador OpenDayLight, montado en una red Mininet [23]. El emulador de redes virtuales Mininet es la herramienta básica para trabajar con SDN. Permite crear redes virtuales junto con todos sus elementos en una única máquina, facilitando la posterior interacción con dichas redes mediante líneas de comandos. Al ser emulador, en lugar de simular el funcionamiento de la red introduce errores aleatorios para que los resultados obtenidos sean los más parecidos a la realidad posible. Como ventajas, destacar que permite el desarrollo de redes diseñadas en hardware, además de la ejecución en tiempo real, lo que permite evaluar condiciones de errores en la red. Dispone de multitud de topologías para la emulación de diferentes escenarios y permite crear nuevas topologías mediante la programación de entornos con Python, lo que facilita el estudio de las redes SDN.

1) Descripción del experimento con protocolo IPv4: Para evaluar el funcionamiento del algoritmo de filtrado para el caso de tráfico IPv4, se diseña esta prueba donde se evalúa el funcionamiento del controlador cuando se produce broadcast de tramas ARP. Así, tras iniciar la red, disponiendo de 3 hosts sin entradas en la tabla del controlador, se realiza un “ping” del host h1 al host h2. Básicamente, al realizar un “ping” se envía un mensaje ICMP desde el nodo origen de tipo *echo request*, que el nodo destino debe responder con un *echo reply*. En ese momento, se produce una petición ARP *request* con origen en h1 y destino en h2. Como respuesta, h2 genera un mensaje ARP *reply* con destino h1. De esta forma, ambos hosts han quedado reconocidos y registrados en el controlador junto con la información relevante para que éste sea capaz de generar respuestas automáticas para atender futuras peticiones. De igual forma, se realiza otro “ping” de h1 hacia h3. Llegados a este punto, si h2 realizase un ping con destino h3, habría tres posibilidades:

a. De no funcionar correctamente el filtrado, no sería posible completar la petición, ya que el controlador podría reconocer las entradas pero no generar una respuesta apropiada.

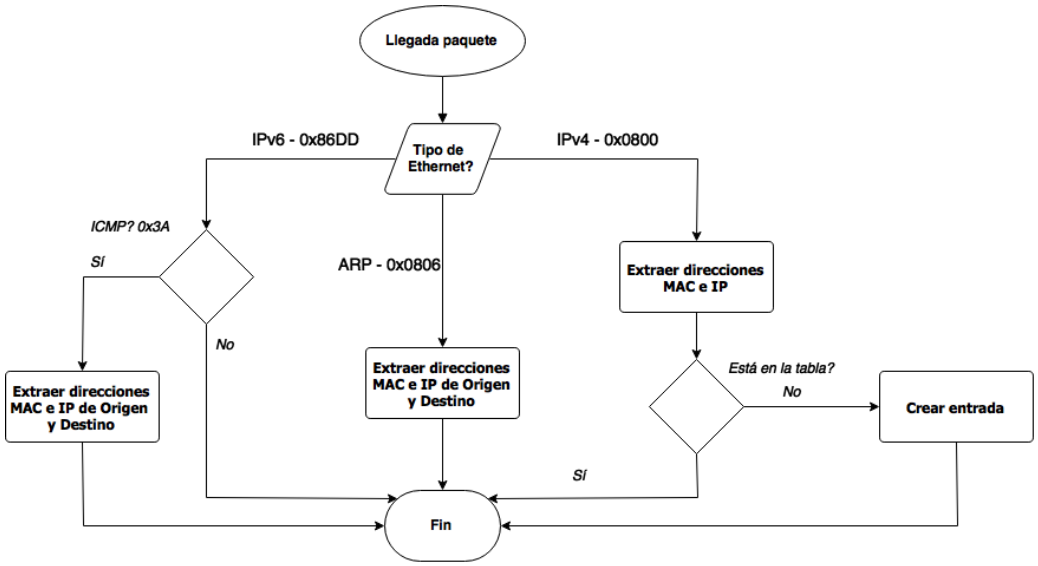


Fig. 3. Diagrama de flujo de la fase de clasificación.

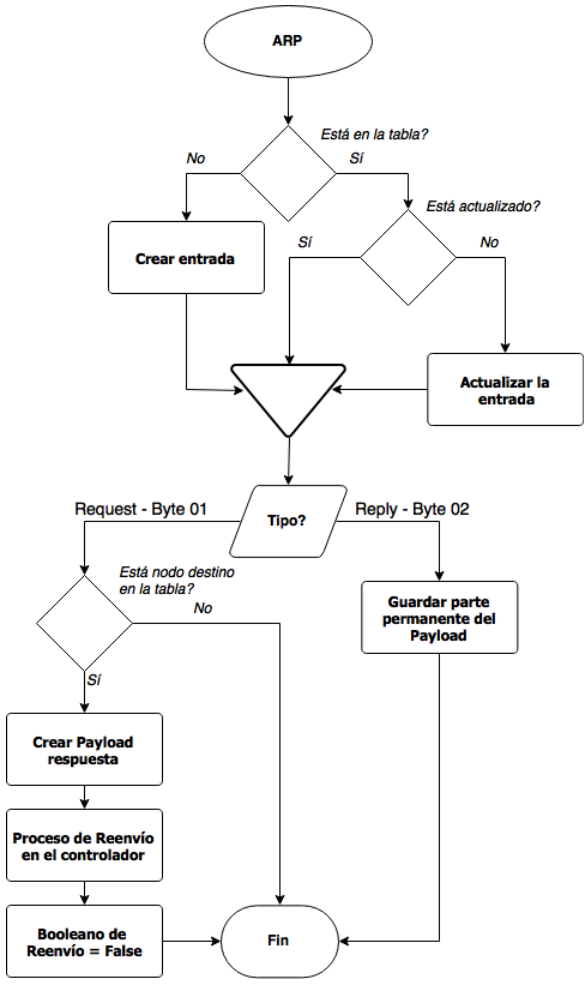


Fig. 4. Diagrama de flujo de la fase de filtrado para ARP.

b. De no realizar el filtrado, h3 respondería con un ARP *reply* a h2. El ping original *request* se habría enviado como broadcast al resto de la red, generando dicha respuesta por parte de h3.
c. De haber filtrado correctamente los mensajes, el controlador bloquea el broadcast de h2 y responde automáticamente con la dirección física asociada a h3. Por tanto, la comunicación entre nodos sería posible sin necesidad tampoco un mensaje ARP *reply* desde h3. Este es el resultado obtenido en este experimento.

2) Descripción del experimento con protocolo IPv6: Para evaluar el funcionamiento del algoritmo de filtrado para el caso de tráfico IPv6, se filtran los mensajes ICMPv6. Estos mensajes se envían al iniciarse una red o añadirse un nuevo nodo, no solo cuando los nodos se van a comunicar, por lo que en este caso se realiza un "ping" para que ambos nodos tengan la dirección del vecino y posteriormente se deshabilita una interfaz, simulando que un nodo desaparece de la red. Al volver a activarse, son necesarios mensajes de solicitud y anuncio de router para realizar la configuración del nodo, y solicitud y anuncio de vecino para ponerse en contacto con el resto de la red. Sin embargo, en caso de realizarse un filtrado apropiado, ambos nodos podrían comunicarse sin necesidad de estos mensajes, tal y como ocurre en el experimento. Si se realiza el envío de mensajes "ping" y "ping6" simultáneamente, se comprueba que no solo es posible sino que se realiza el filtrado de los mensajes de manera conjunta.

3) Descripción del experimento con protocolos IPv4 e IPv6: En este escenario se desea evaluar si la implementación realizada soporta la coexistencia de paquetes de IPv4 e IPv6.

A. Entorno experimental

Para la realización de los experimentos se utilizó un equipo personal con procesador Intel Core i7 a 2.7GHz, memoria RAM de 4GB y unidad de estado sólido, Solid-State Drive (SSD) de 500GB de capacidad. Sobre este equipo se ejecutaba una máquina virtual Oracle VirtualBox, Sistema Operativo Ubuntu 14.04 (64 bits), con la

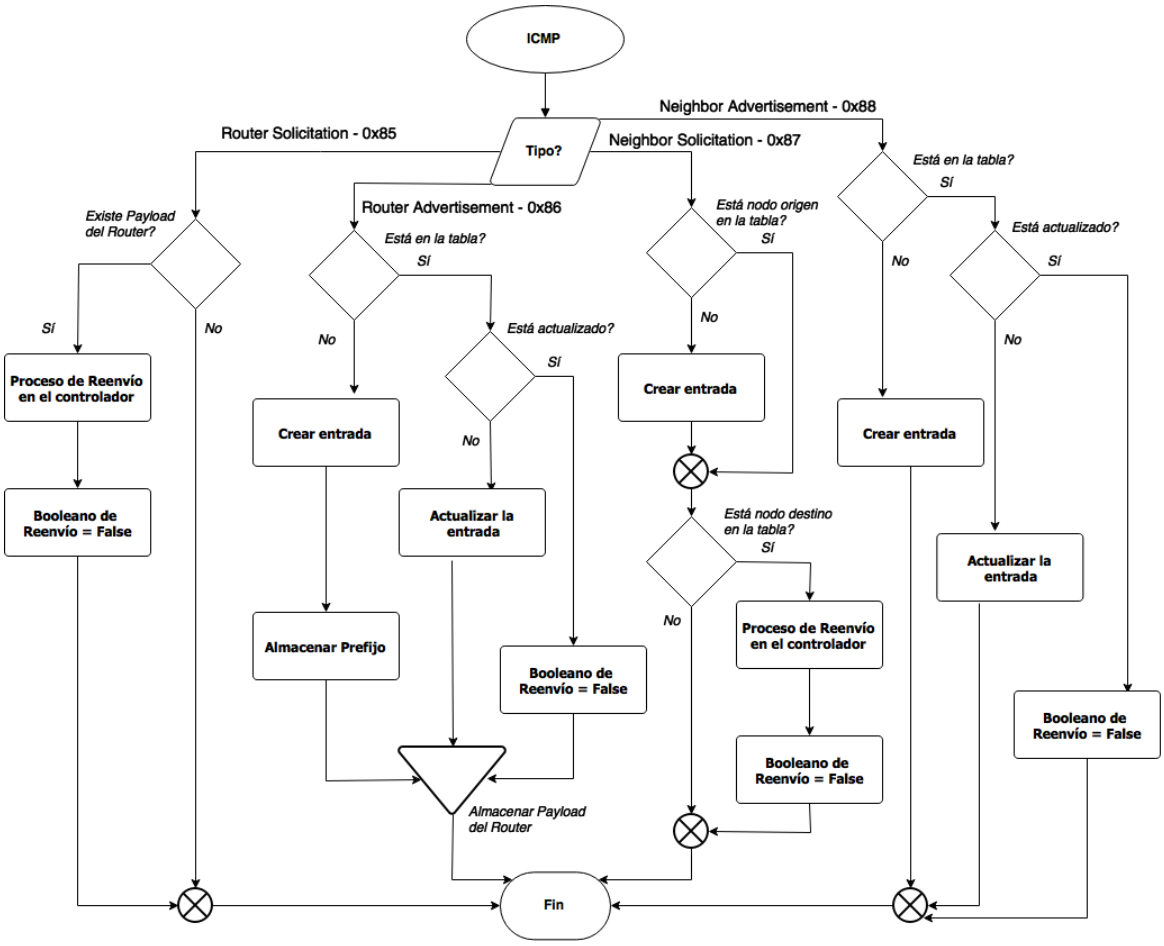


Fig. 5. Diagrama de flujo de la fase de filtrado para ICMP.

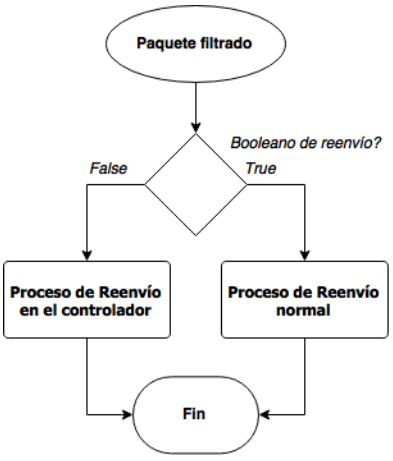


Fig. 6. Diagrama de flujo de la fase de reenvío.

herramienta Mininet, controlador OpenDayLighty el analizador de protocolos Wireshark.

Básicamente, se estudió el rendimiento del sistema para una red con 5, 10 y 20 nodos que generan tráfico de los paquetes IPv4 ó IPv6 identificados en las secciones previas, unas veces con, y otras sin el algoritmo de filtrado implementado. De esta forma se pretende obtener resultados comparables para ambos casos, y analizar si realmente el filtrado supone o no una mejora para la red. Para que la diferencia del tráfico generado no sea excesiva, la opción sin filtrado tiene un controlador funcionando como conmutador con aprendizaje de direcciones de enlace, en lugar de como concentrador. De esta manera, aprende las direcciones de los nodos y las añade a una tabla para su posterior enrutamiento, en lugar de únicamente dejar pasar todo el tráfico. Se utiliza la topología básica de IPv6 para realizar este experimento, mostrada en Fig. 8, variando en este caso el número de nodos.

1) Escenario con protocolo IPv4: Se plantea un escenario con ODL como controlador remoto, un conmutador OpenFlow y tres hosts. Para ello, se lanza el controlador, se instala la aplicación, se crea la topología en Mininet y se añade una regla para que el flujo del conmutador se dirija al controlador.

2) Escenario con protocolo IPv6: En este caso se dispone de dos hosts y un router, además del conmutador OpenFlow y del controlador, tal y como muestra Fig. 8.

Para realizar la configuración de la red en IPv6 es necesario instalar y configurar el demonio de anuncios de router (Router ADVertisement Daemon, RADVD) [24]. Con el archivo de configuración, se establece el prefijo de red que utilizan los dispositivos en la autoconfiguración de direcciones, así como la interfaz del router destinada al envío de dicho prefijo, que es la interfaz del nodo que

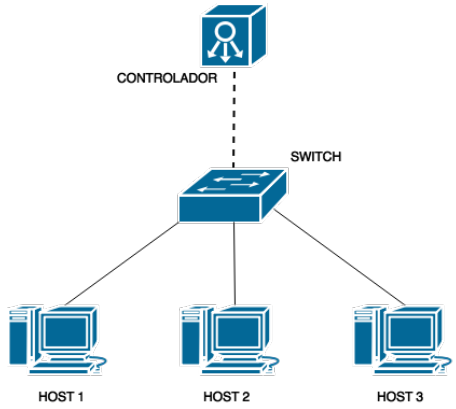


Fig. 7. Entorno experimental con IPv4 y conmutador OpenFlow.

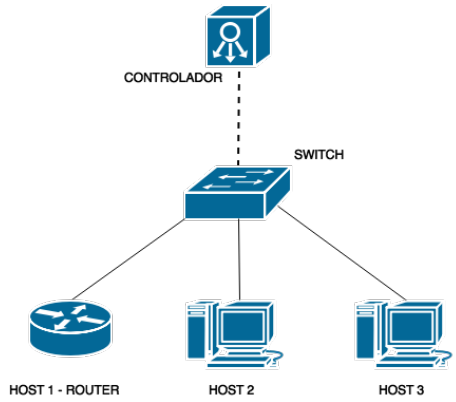


Fig. 8. Entorno experimental con IPv6 y conmutador OpenFlow.

actúa como router en la red. Tras lanzar la topología en Mininet, se configura el router y el resto de nodos, se habilita IPv6 en Mininet y se inicia el servicio RADVD. Así, cada nodo dispone de una dirección de enlace IPv6 y, en el caso de los nodos que no actúan como router, de una dirección global asignada por el router a partir del prefijo creado.

Con este entorno no sólo se evalúa la implementación correcta del algoritmo en el controlador, sino también la configuración de IPv6, ya que en este escenario hay un host encargado de funcionar como router, asignando al resto de elementos de la red sus direcciones globales.

3) Escenario con protocolos IPv4 e IPv6: Se crean archivos para la configuración automática de IPv6 y la generación de paquetes "ping" y "ping6". Se realiza la transmisión de 20 paquetes de cada protocolo por nodo, con un intervalo de 5 segundos entre paquetes. Posteriormente, se deshabilitan las interfaces, se vuelven a habilitar y se vuelve a lanzar el script de envío de paquetes.

B. Resultados experimentales

En el caso de las pruebas de funcionamiento correcto del filtrado en el controlador SDN, los resultados muestran que para ambos protocolos, tanto independiente como simultáneamente, la solución propuesta es viable.

Respecto al rendimiento, se muestra en tabla I el resumen de las estadísticas para cada caso estudiado,

Tabla I
NÚMERO DE PAQUETES TRANSMITIDOS PARA CADA CASO EN LA PRUEBA DE RENDIMIENTO.

Escenario	ARP	ARP filtrado	ICMPv6	ICMPv6 filtrado
5 Nodos	94	30	1064	654
10 Nodos	1585	811	9348	7947
20 Nodos	–	63458	–	583506

presentando el número de paquetes transmitidos por protocolo en cada caso, para una ejecución de cada escenario. Debido a las limitaciones en cuanto a equipamiento para la realización del proyecto, los experimentos se realizan para redes con un máximo de 20 nodos. En el caso del sistema con un controlador sin filtrado, la red Mininet con 20 nodos deja de funcionar debido a la sobrecarga de la misma, por lo que no se puede comparar este resultado con el que se obtiene para la red con la solución implementada. Destaca sin embargo la mejora de alrededor del 70% en cuanto a disminución de paquetes ARP para el caso con 5 nodos y del 50% para el caso de 10 nodos; y del 40% y 15% para 5 y 10 nodos en el caso de IPv6, respectivamente.

Por tanto, los resultados muestran que la utilización de un controlador centralizado con la solución implementada supone una notable disminución del tráfico de la red.

VI. CONCLUSIONES

En este trabajo se aborda la implementación de un algoritmo de filtrado basado en SDN para mitigar la degradación de rendimiento de los sistemas por las inundaciones de red por broadcasts debidas a protocolos de arranque. El objetivo es filtrar dichos mensajes para aumentar la escalabilidad de estas redes Ethernet. Se diseña para varios procedimientos de de ARP e ICMPv6, con un diseño escalable a futuros protocolos. Posteriormente, se realizan pruebas de rendimiento en un entorno emulado y se obtienen unos resultados satisfactorios en cuanto a la disminución del tráfico enviado en estas redes para ambos protocolos.

Como vías de investigación futuras, se propone la integración de la solución implementada en una red real de telecomunicaciones; la escalabilidad del código implementado a otros protocolos que generan tráfico broadcast o multicast como DHCP, Bonjour, etc; y el diseño de la arquitectura de la quinta generación móvil como una red Ethernet implementada sobre SDN, en la que la nube de acceso dispondría de un controlador SDN encargado de gestionar el tráfico de la red.

VII. AGRADECIMIENTOS

Este trabajo está parcialmente financiado por el Ministerio de Economía, Industria y Competitividad y el Fondo Europeo de Desarrollo Regional FEDER (proyectos TEC2016-76795-C6-4-R y TIN2013-46223-P).

REFERENCIAS

[1] Cisco, "Visual networking index: Global mobile data traffic forecast update, 2015–2020," *Tech. Rep., Cisco*, 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/

visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] O. N. Foundation, "Software defined networking: The new norm for network," Tech. Rep., April 2012. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf

[3] ——, "Openflow switch specification," Open Networking Foundation, Tech. Rep., 2012. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.1.pdf

[4] L. Foundation, *OpenDaylight Developer Guide*, 2015. [Online]. Available: http://go.linuxfoundation.org/l/6342/2015-06-28/2l76qr/6342/128124/bk_developers_guide_20150629.pdf

[5] B. Lantz, N. Handigol, B. Heller, and V. Jeyakumar, *Introduction to Mininet*, December 2015. [Online]. Available: https://github.com/mininet/mininet/wiki/Introduction-to-Mininet

[6] A. F. Cattoni, P. E. Mogensen, S. Vesterinen, M. Laitila, L. Schumacher, P. Ameigeiras, and J. J. Ramos-Munoz, "Ethernet-based mobility architecture for 5g," in *Cloud Networking (CloudNet), 2014 IEEE 3rd International Conference on*, Oct 2014.

[7] P. Ameigeiras, J. J. Ramos-Muñoz, L. Schumacher, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Link-level access cloud architecture design based on sdn for 5g networks," *IEEE Network*, vol. 29, no. 2, March 2015.

[8] Cisco, "Internetwork design guide - broadcasts in switched lan internetworks." [Online]. Available: http://docwiki.cisco.com/wiki/Internetwork_Design_Guide_--_Broadcasts_in_Switched_LAN_Internetworks#Table:_Average_Number_of_Broadcasts_and_Multicasts_for_Novell_Networks

[9] D. C. Plummer, *RFC 826 An Ethernet Address Resolution Protocol*, Std., November 1982.

[10] R. Droms, *RFC 2131 Dynamic Host Configuration Protocol*, Std., March 1997. [Online]. Available: https://www.ietf.org/rfc/rfc2131.txt

[11] S. Thomson, T. Narten, and T. Jinmei, *RFC 4862 IPv6 Stateless Address Autoconfiguration*, Std., September 2007. [Online]. Available: https://tools.ietf.org/pdf/rfc4862.pdf

[12] M. Casado, Ed., *Ethane: Taking Control of the Enterprise*, vol. 37, no. 4. NY, USA: Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications, October 2007.

[13] C. Kim, M. Caesar, and J. Rexford, "Floodless in seattle: A scalable ethernet architecture for large enterprises," vol. 29, no. 1. NY, USA: ACM Trans. Computer Systems (TOCS), February 2011.

[14] A. Conta and S. Deering, *RFC 2463 ICMP for the Internet Protocol Version 6 (IPv6)*, Std., December 1998. [Online]. Available: http://tools.ietf.org/pdf/rfc2463.pdf

[15] P. L. Weiguo, *Address Resolution Protocol*, Std., September 2015. [Online]. Available: http://icourse.cuc.edu.cn/networkprogramming/lectures/Unit2_ARP.pdf

[16] S. Deering and R. Hinden, *RFC 2460 Internet Protocol, Version 6 (IPv6) Specification*, Std., December 1998. [Online]. Available: http://tools.ietf.org/html/rfc2460

[17] M. Jammal, T. Singh, A. Shami, R. Asal, and Y. Li, "Software defined networking: State of the art and research challenges," *Computer Networks*, vol. 72, pp. 74 – 98, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614002588

[18] F. Longo, S. Distefano, D. Bruneo, and M. Scarpa, "Dependability modeling of software defined networking," *Computer Networks*, vol. 83, pp. 280 – 296, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128615001139

[19] I. F. Akyildiz, A. Lee, P. Wang, M. Luo, and W. Chou, "A roadmap for traffic engineering in sdn-openflow networks," *Computer Networks*, vol. 71, pp. 1 – 30, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614002254

[20] T. N. D. and K. Gray, *SDN: Software Defined Networks*, 1st ed. O'Reilly Media, Inc., 2013.

[21] J. Medved, R. Varga, A. Tkacik, and K. Gray, "Opendaylight: Towards a model-driven sdn controller architecture," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on a*, June 2014, pp. 1–6.

[22] A. L. Stancu, S. Halunga, A. Vulpe, G. Suciu, O. Fratu, and E. C. Popovici, "A comparison between several software defined networking controllers," in *Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), 2015 12th International Conference on*, Oct 2015, pp. 223–226.

[23] "Mininet: An instant virtual network on your laptop (or other pc)." [Online]. Available: http://mininet.org

[24] L. S. Design. (2016) Linux ipv6 router advertisement daemon (radvd). (último acceso el 2/5/2017). [Online]. Available: http://www.litech.org/radvd/

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6612

Evaluación de un sistema DASH para el streaming de vídeo 3D

Paola Guzmán Castillo, Pau Arce Vila, Juan Carlos Guerri.
Grupo Comunicaciones Multimedia, iTEAM (Instituto de Telecomunicaciones y Aplicaciones Multimedia)
Universitat Politècnica de València
Camino de Vera, s/n.
paoguzc1,paarvi @iteam.upv.es, jcguerri@dcom.upv.es

***Resumen*- La distribución de contenidos multimedia, y en particular el streaming de vídeo, domina actualmente el tráfico global de Internet y su importancia será incluso mayor en el futuro. Miles de títulos se agregan mensualmente a los principales proveedores de servicios, como Netflix, YouTube y Amazon. Y de la mano del consumo de contenidos de alta definición que se convierte en la principal tendencia, se puede observar nuevamente un incremento en el consumo de contenidos 3D. Esto ha hecho que las temáticas relacionadas con la producción de contenidos, codificación, transmisión, calidad de servicio (QoS) y calidad de experiencia (QoE) percibidas por los usuarios de los sistemas de distribución de vídeo 3D sean un tema de investigación con numerosas contribuciones en los últimos años. Es importante tener en cuenta que en un sistema de distribución de vídeo las degradaciones debidas a la producción y la codificación, así como los errores de transmisión, puede degradar la calidad del vídeo recibida y percibida por el usuario. Por tanto, como parte de este trabajo se ha realizado en primer lugar una comparación del rendimiento de los estándares de codificación de vídeo más populares H.264, H.265 y sus correspondientes extensiones para vídeo 3D. Por otra parte, se ha realizado una evaluación experimental de la calidad del vídeo recibida en un escenario HTTP de streaming adaptativo (DASH) de vídeo 3D.**

***Palabras Clave*- DASH, vídeo 3D, HEVC, AVC, streaming adaptativo, QoE, QoS.**

I. INTRODUCCIÓN

El servicio de streaming de vídeo crece y evoluciona a una velocidad increíble. Según las previsiones y estadísticas disponibles, el tráfico de vídeo representará el 82% de todo el tráfico de Internet para 2021, frente al 73% en 2016[1]. Por su parte, las recientes mejoras en la tecnología de vídeo 3D han suscitado nuevamente un creciente interés hacia el consumo de dichos contenidos, como una alternativa para expandir la experiencia del usuario. Nuevos contenidos demandan nuevos esquemas de representación y codificación, que se ajusten a las condiciones de transporte y restricciones de ancho de banda, y permitan maximizar la calidad de servicio (QoS, Quality of Service) y la calidad de experiencia (QoE, Quality of Experience) del usuario. En este sentido, tanto las perdidas asociadas a los procesos de codificación y compresión, como los errores y pérdidas durante la transmisión pueden afectar a la calidad percibida por el usuario.

Para poder estudiar el impacto que tiene cada uno de estos aspectos, nuestro primer objetivo ha sido realizar una comparación en términos de evaluación objetiva usando los parámetros PSNR (Peak Signal-to-Noise Ratio) y SSIM (Structural Similarity), empleando los estándares más populares de codificación de vídeo, como H.264/AVC (Advanced Video Coding) y H.265/HEVC (High Efficiency Video Coding) con sus respectivas extensiones para formatos multivista utilizando el estándar MVC (Multiview Video Coding).

En diversos estudios previos se han hecho comparativas de varias generaciones de estándares de codificación en términos de PSNR y pruebas subjetivas, e incluso comparando los nuevos codificadores de alta eficiencia H.265 y VP9. En un escenario de distribución de vídeo 3D, se utilizaban formatos estereoscópicos compatibles con 2D (Frame-compatible o Full-resolution Frame-compatible), pero la aparición de sistemas basados en pantallas auto-estereoscópicas o aplicaciones de telepresencia inmersiva ha motivado el desarrollo de formatos multivista como MVC o MVD (Multiview Video Plus Depth).

Otros factores que pueden afectar a la calidad de vídeo son las pérdidas y retardos ocasionados durante la transmisión. En la actualidad, el transporte de flujos de

vídeo en Internet se realiza cada vez más utilizando HTTP adaptativo, del cual el estándar DASH (Dynamic Adaptive Streaming over HTTP) es el protocolo más representativo.

Así, en [2] se evalúa el impacto del protocolo de transporte en la calidad percibida por el usuario cuando el vídeo se transmite sobre un enlace de ancho de banda limitado, utilizando YouTube como ejemplo. Por otro lado, en [3] se presenta un análisis comparativo de las diferentes estrategias de adaptación de bitrate en un escenario de streaming adaptativo tanto monoscópico como estereoscópico.

Asimismo, en publicaciones recientes [4] se revisan de forma exhaustiva los métodos de evaluación tanto objetivos como subjetivos relacionados con el servicio de videostreaming. Por su parte en [5] se aborda la evaluación de la QoE en escenarios de vídeo 3D comparando diferentes técnicas de codificación y utilizando como referencia la recomendación BT.500.13 ITU-R.

En este sentido, se plantea evaluar el efecto que produce la adaptación de la tasa de vídeo 3D al ancho de banda disponible y cuál es su influencia en la QoE, utilizando el escenario de la Fig. 1.

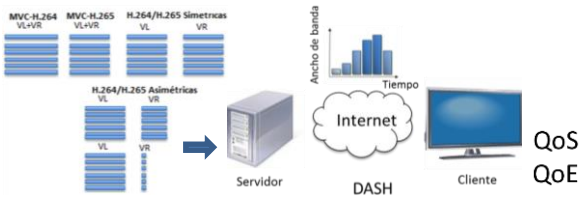


Fig. 1. Diagrama general.

Por tanto, en la sección II se hará una descripción del procedimiento seguido para la selección de secuencias y los resultados de la comparación de los estándares H.264/AVC y H.265/HEVC. En la Sección III se mostrarán los resultados de la evaluación experimental de un sistema de DASH para vídeo 3D. Finalmente, las conclusiones se exponen en la Sección IV.

II. Selección de Sequencias y Comparación de Codificadores

A. Selección de secuencias

Las secuencias utilizadas en este trabajo han sido tomadas de las bases de datos de vídeos 3D HD estereoscópico, Nantes-Madrid-3D-Stereoscopic-V1, NAMA3DS1 y RMIT3DV, que se encuentran disponibles abiertamente en sus sitios web. Dichas bases de datos están compuestas por secuencias estereoscópicas (vistas por separado) con resolución 1920x1080 a 25 fps, diseñadas para representar una amplia gama de contenidos y condiciones visuales. Asimismo, se ha empleado la ya popular secuencia de animación Big Buck Bunny en 3D producida por Blender Foundation. Se evaluarán un total de 13 secuencias con duración entre 16 s y 634 s la más larga.

Con el fin de cuantificar y evaluar la variedad de los contenidos seleccionados y la dificultad de codificación, tal como se describe en la Recomendación ITU-T P.910, se ha calculado el índice de información espacial (SI) y el índice de información temporal (TI) de cada una da las secuencias bajo estudio. En la Fig. 2 se muestran los índices SI y TI calculados sobre la componente de luminancia de cada secuencia y una miniatura de cada una de ellas.

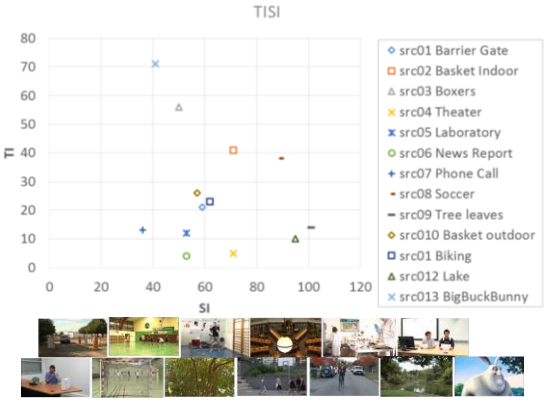


Fig. 2. Indices SI –TI

Dados los tiempos de codificación de algunas de las herramientas que serán empleadas para realizar la comparación entre los estándares H.264/AVC y H.265/HEVC, para esta primera fase, se seleccionaron 10 s de cuatro de las secuencias mencionadas anteriormente. Intentando cubrir un espectro amplio en cuanto a tipo de contenidos, se ha elegido una de tipo outdoor, una indoor, una de deportes y otra de animación.

B. Comparación de codificadores

Para la comparación de los estándares H.264/AVC y H.265/HEVC en el ámbito de la codificación de secuencias estereoscópicas, además del software libre ffmpeg-libx264 y ffmpeg-libx265, se emplearon los codificadores de referencia. Por un lado, HEVC test Model (HM 16.7) para las codificaciones HEVC Simulcast y MVC-HEVC, y por otro lado, los codificadores JM19.0 y JMVC 8.5 para las codificaciones AVC simulcast y MVC-AVC, respectivamente.

Para generar variaciones de calidad, se emplearon los parámetros de cuantificación (QP) 24, 28, 32, 36 y 40, pero debido a la diferencia significativa en las capacidades de los codificadores, se buscó que los parámetros de configuración fuesen equivalentes en todos los codificadores, eligiendo los mismos valores de GoP (Group of Pictures) y separación entre frames Intra.

La Fig. 3 muestra las curvas RD (Rate-Distortion) en función del PSNR, para dos de las secuencias bajo estudio. Para ésta y las figuras siguientes usaremos PSNR como medida de calidad. PSNR es la métrica más utilizada en la compresión de vídeo ya que, aunque no siempre refleja la calidad perceptual, es una manera simple de medir la fidelidad a la fuente (40 dB o

superior es muy buena calidad, por debajo de 35 dB mostrará artefactos de codificación).

Fig. 3. Curva RD (PSNR) comparación de codificadores.

Como se puede ver en la Fig. 3, en todos los casos evaluados los codificadores basados en HEVC presentan una mejor relación entre la calidad del vídeo obtenida y el bitrate. También podemos observar que los codificadores de referencia, en particular las versiones MVC, tienen mejores prestaciones comparados con los resultados obtenidos al codificar cada vista por separado. Esto se debe a que además de considerar las similitudes entre frames dentro de cada vista, explotan también la similitud intervista, lo que permite suprimir frames de tipo Intra al codificar una de las vistas en función de la otra.

Sin embargo, enfocados en nuestro siguiente objetivo que se centra en la transmisión de contenidos de vídeo 3D en un entorno adaptativo, además de la eficiencia de codificación, se ha valorado la opción de poder generar codificaciones asimétricas y los tiempos de codificación. La Tabla II muestra la amplia diferencia entre los tiempos de codificación usando codificadores de referencia respecto al ffmpeg. Las codificaciones asimétricas se valen de las características del sistema de visión humano, el cual al visualizar un vídeo 3D donde cada una de las vistas ha sido codificada con un factor de calidad, se ha demostrado que prevalece la sensación de la vista con mejor calidad [6] y resultan una buena alternativa para la reducción de la tasa de bits en función de la calidad en un escenario de streaming adaptativo.

Tabla II
TIEMPOS DE CODIFICACIÓN- COMPARACIÓN DE CODIFICADORES DE REFERENCIA Y FFMPEG (MVC VS VISTAS POR SEPARADO)

Codificador	Tiempo de codificación (s)	Segundos por frame
FFMPEG H.264	121	0,4
JM H.264	23743	79,1
JMVC MVC-H.264	23778	79,3
FFMPEG H.265	3179	10,6
HM H.265	31382	104,6
HM MVC-H.265	1595	5,3

Por tal motivo, en adelante se usará como herramienta para la codificación ffmpeg con las librerías de codificación libx264 y libx265. La Tabla II muestra el promedio del tiempo de codificación con cada aplicación. El ordenador donde se realizaron las codificaciones es un Pentiun Dual-Core a 2.5GHz con una versión Ubuntu 16.10 de 64 bits.

De los resultados obtenidos en un trabajo anterior [6], donde se emplearon métodos de evaluación tanto objetivos como subjetivos de calidad de vídeo, se pudo demostrar que el estándar H.265/HEVC proporciona un ahorro significativo en el bitrate con respecto a H.264/AVC. En línea con estos resultados, para las secuencias tomadas en el presente trabajo, tal como se muestra en la Tabla III, la reducción en la tasa de bits para una misma calidad varía entre un 66,8% a un factor de calidad bajo (QP40) para una secuencia de tipo indoor con bajo nivel de movimiento (src06 – informativo noticias), y un 10,9% para un factor de calidad alto QP24 y una secuencia de deporte (src010 – Basket outdoor). En la secuencia de tipo animación (src013 – BigBuckBunny) la reducción de bitrate presenta una variación menor, oscilando entre un 36,9% con QP24 y un 59,3% con QP40.

Tabla III
REDUCCIÓN EN LA TASA DE BITS DE H.265/HEVC (FFMPEG) RESPECTO A H.264/AVC (FFMPEG)

Secuencia	QP24	QP28	QP32	QP36	QP40
Src01	23,4%	42,8%	48,8%	52,6%	55,7%
Src06	36,0%	56,2%	62,6%	66,0%	66,8%
Src010	10,9%	36,2%	41,7%	45,0%	49,0%
Src013	36,9%	48,1%	53,2%	56,1%	59,3%

III. EVALUACIÓN DE LA TRANSMISIÓN CON DASH

A. Producción de contenidos DASH

Como paso previo a la generación de los contenidos DASH se debe realizar un proceso de codificación, que como se mencionó anteriormente en nuestro caso se ha realizado empleando la aplicación ffmpeg y las librerías libx264 y libx265. Como parámetro de codificación se ha empleado la tasa de bits máxima. Como paso previo a la generación de las secuencias en formato SBS (Side by Side) Frame-compatible, las secuencias correspondientes a cada una de las vistas (izquierda y derecha) son codificadas con diferentes valores de bitrate (2000kbps, 1600kbp, 1000kbps, 700kbps, 500kbps). Teniendo en cuenta que no todas las secuencias se comportan igual frente a determinados parámetros de codificación, la selección de la tasa de bits óptima es un tema a tener en cuenta. Al realizar un gráfico comparativo que incluye las curvas RD de las 13 secuencias codificadas con 5 valores diferentes de QP, se observa como mientras algunas secuencias de vídeo alcanzan un PSNR muy alto (45 dB o más) a bitrates de 2000 kbps o menos, por otra parte, algunos vídeos a estas tasas de bits sólo alcanzan un valor aceptable de PSNR de 38 dB.

Cuando se utiliza el estándar DASH, el vídeo se segmenta de forma que el cliente pueda pedir los

segmentos uno a uno, y en función del ancho de banda que mida en cada momento, pueda escoger descargar los siguientes segmentos de vídeo de mayor o menor calidad. De entre las secuencias evaluadas en la Sección II, se han seleccionado las codificaciones óptimas a partir de la curva de PSNR, maximizando la relación de PSNR y bitrate.

Una vez codificadas las secuencias, se utiliza MP4Box para convertir el vídeo en segmentos DASH de 5 segundos y generar un archivo de índice MPD (Media Presentation Description), que contiene toda la información sobre las diferentes calidades de vídeo usadas y los anchos de banda de cada una. Este es el archivo que el cliente utilizará para saber qué segmentos descargarse en función del ancho de banda medido. Finalmente, como servidor de contenidos hemos usado un equipo Apache 2.4.18 para Linux Ubuntu. De entre todos los clientes disponibles para reproducir DASH, se utiliza una versión modificada por los autores del Shaka Player V2.0.0, del que se puede obtener información relativa al throughput, tiempo de descarga y nivel del buffer de reproducción.

B. Evaluación prestaciones DASH

Para evaluar el comportamiento de adaptación de DASH y cómo afecta a la calidad de vídeo 3D, se emulan diferentes canales de transmisión con ancho de banda variable. Para ello se utiliza la herramienta NetEm, que es capaz de modificar y restringir el ancho de banda de salida del servidor (también el retardo o la pérdida de paquetes) de forma que el cliente perciba cambios en la tasa de descarga instantánea y adapte la calidad de vídeo en consecuencia.

Para los experimentos se han definido diferentes canales de transmisión, variaciones rápidas y lentas de ancho de banda, teniendo en cuenta el tamaño de segmento utilizado (5 s). Por cuestiones de espacio, en este trabajo se presenta únicamente el escenario con variaciones de ancho de banda cada 40 s.

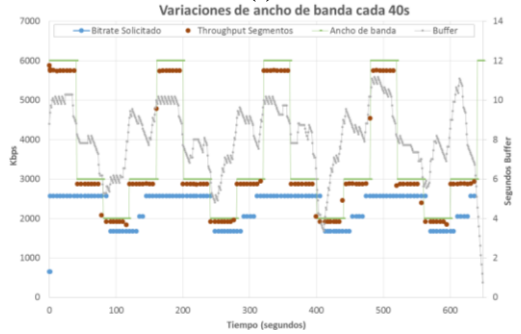
La Fig. 4(a) muestra el throughput por segmento alcanzado en un escenario sin restricciones en el que el cliente puede descargar las representaciones de más alta calidad. El archivo MPD ofrece 9 calidades entre 0,65 Mbps y 2,5 Mbps. Por su parte en la Fig. 4(b) se muestra el comportamiento del Shaka Player frente a un escenario de variaciones persistentes de ancho de banda.

IV. CONCLUSIONES

El uso de codificaciones asimétricas para la representación de las secuencias de vídeo estereoscópico resulta de especial interés en un escenario HTTP de streaming adaptativo (DASH) ya que, aprovechando las características del sistema audiovisual humano donde prevalece la sensación de la vista con mejor calidad, permite reducir el ancho de banda de las transmisiones y aumentar la granularidad en los cambios de calidad, mejorando así la calidad subjetiva percibida por el usuario.

El mecanismo de adaptación de Shaka Player, además de evitar el congelamiento del vídeo frente a una caída del ancho de banda, hace un buen uso del buffer en situaciones de reducción de ancho de banda, lo cual le permite minimizar el número de cambios de calidad efectuados en un escenario de variaciones de ancho de banda con alta frecuencia, al mismo tiempo que reacciona con gran rapidez en situaciones de aumento de ancho de banda.

(a)



(b)

Fig. 4. (a) Throughput y el buffer por segmentos sin restricciones de ancho de banda. (b) Ancho de banda, Throughput por segmento, bitrate solicitado, ocupación de buffer de reproducción

AGRADECIMIENTOS

Este artículo se enmarca en el Proyecto PROMETEOII/2014/003 financiado por la Generalitat Valenciana y el Proyecto "Desarrollo de Nueva Plataforma de Entretenimiento Multimedia para Entornos Náuticos" (CDTI IDI -20170348).

REFERENCIAS

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016-2021,", 2017.

[2] T. Hoßfeld, R. Schatz, and U. R. Krieger, "QoE of YouTube Video Streaming for Current Internet Transport Protocols," in *Measurement, Modelling, and Eval. of Computing Systems and Dependability and Fault Tolerance*, 2014, pp. 136–150.

[3] S. Tavakoli, J. Gutierrez, and N. Garcia, "Subjective quality study of adaptive streaming of monoscopic and stereoscopic video," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 684–692, 2014.

[4] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2015.

[5] T. Tian, X. Jiang, and X. Du, "Subjective quality assessment of compressed 3D video," in *2014 7th International Congress on Image and Signal Processing*, Dalian, 2014,

pp. 606–611.

[6] P. Arce, I. De Fez, F. Fraile, S. González, P. Guzmán, and J. C. Guerri, "QoE en redes adhoc, descarga adaptativa de contenidos y vídeo 3D,"*Proc. of Jornadas de Ingeniería Telemática (JITEL)*, Mallorca, 2015.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6620

Diseño y evaluación de un servicio OpenFlow de provisión de Calidad de Experiencia sobre Mininet

Cristian Alfonso Prieto Sánchez, Pilar Andres-Maldonado, Jonathan Prados-Garzon, Juan José Ramos-Munoz,
Departamento de Teoría de la Señal, Telemática y Comunicaciones,
Universidad de Granada
Calle Periodista Daniel Saucedo Aranda s/n, E-18071 (Granada)
rps@ugr.es, pam91@correo.ugr.es, jpg@ugr.es, jjramos@ugr.es

***Resumen*—Las redes definidas por software suponen un nuevo paradigma de red que potencialmente pueden proporcionar nuevas soluciones para proporcionar calidad de experiencia.**

En este trabajo se propone un diseño de aplicación de redes definidas por software para proveer calidad de experiencia a flujos con distintos requisitos de red. Esta aplicación se implementa en OpenDayLight, una implementación de referencia, y se ejecuta en el emulador Mininet.

En este artículo se detallan los algoritmos implementados para proporcionar la aplicación para OpenDayLight.

***Palabras Clave*—Controlador, Dijkstra, Java, Mininet, OpenDayLight, OpenFlow, QoE, QoS, Redes, Routing, SDN.**

I. INTRODUCCIÓN

El crecimiento del tráfico multimedia experimentado en los últimos años, unido al crecimiento exponencial esperado (como las previsiones realizadas por Cisco en [1]), ahondan en la necesidad de diseñar una arquitectura de red capaz de soportar y cumplir los requisitos de calidad de experiencia (QoE) para soportar el tráfico multimedia y los nuevos tipos de aplicaciones.

Por otro lado, es recurrente la búsqueda de técnicas o arquitecturas de red capaces de reducir costes y aumentar capacidades sobre las que ofrecer nuevos servicios a los usuarios, por parte de operadores de red. Estas dos necesidades básicas en la industria han impulsado la adopción de redes definidas por software (Software Defined Networking, SDN) como solución con mayor aceptación, dadas las ventajas en las se hará hincapié a continuación.

La solución diseñada en este proyecto se basa en aprovechar las ventajas que ofrece disponer de un elemento de toma de decisiones centralizado, el controlador SDN, que tiene una visión global de toda la red, para poder desplegar rutas que satisfagan los requisitos de calidad de experiencia (QoE) de varias clases de flujos multimedia.

Concretamente, se ha adaptado un protocolo de encaminamiento de *estado de enlace*, para generar las rutas por cada flujo multimedia que llegue a la red SDN. Para ello, el protocolo de encaminamiento toma como costes distintas métricas, que dependen del tipo de flujo para el que se crea la ruta, y sus requisitos. Así, tras identificar el tipo de un nuevo flujo, se calculará y configurará una ruta a su destino, que minimice los parámetros de red que degradan la calidad que percibirá el destinatario.

Además, el servicio diseñado monitoriza el estado de los enlaces para recalcular nuevas rutas, por tipo de aplicación, si las condiciones de red empeoran.

II. DISEÑO DE LA SOLUCIÓN

La propuesta actual se basa en el funcionamiento de las aplicaciones SDN, en las que un controlador lógicamente centralizado monitoriza y programa los conmutadores SDN. De esta manera, las decisiones del controlador se pueden realizar con información completa de la red y, consecuentemente, se podrían tomar decisiones óptimas.

La solución propuesta se compone de varios elementos, que se especifican en las siguientes subsecciones:

A. *Clasificación de flujos paquetes*

Para poder seleccionar la ruta que mejor se ajuste a los requisitos de QoE de un flujo de paquetes, es necesario identificar a qué tipo de aplicación corresponde dicho flujo. La clasificación de paquetes se hace de acuerdo a la figura 1, atendiendo a las características bien conocidas de cada tipo de paquete.

Inicialmente, en este trabajo se distingue entre tráfico TCP, ICMP, y RTP con perfil de audio y vídeo.

B. Algoritmo de encaminamiento

Una vez reconocido el tipo de flujo, es necesario ejecutar un algoritmo que proporcione una ruta hacia el destino del flujo. Los parámetros de QoS de los enlaces que componen dicha ruta debería maximizar la calidad que el usuario final al que va destinado el flujo percibirá. Para ello hace falta estimar previamente cuál es el coste de cada enlace para el algoritmo. A partir de la estimación de los pesos para cada enlace, se aplica el algoritmo Dijkstra [2] que nos dará la "mejor" ruta en el momento de consulta.

C. Definición de la matriz de costes por enlace

El algoritmo seleccionado para construir las rutas para cada flujo necesita una representación de la red (qué conmutadores SDN y qué enlaces existen). Esta representación será la matriz de costes. Para calcular dicha matriz, es necesario definir una métrica de calidad que se base en los parámetros objetivos de cada enlace, y que tenga en cuenta el tipo de flujo que debe seguir dicha ruta. La estimación de la matriz de costes se realiza con cuatro parámetros de calidad, que, para un enlace $Edge$, dan el peso total $C_{[}Edge]$ calculado en 2:

$$\begin{aligned} C_{[}Edge] = \alpha \cdot C_{latency}[Edge] \\ +\beta \cdot C_{jitter}[Edge] \\ +\gamma \cdot C_{load}[Edge] \\ +\omega \cdot C_{loss}[Edge] \end{aligned} \quad (1)$$

donde $C_{latency}[Edge]$ representa la componente del coste debido a retardo, $C_{jitter}[Edge]$ la componente del coste debido a la variación de retardo entre paquetes (*jitter*), y $C_{load}[Edge]$ la componente del coste debido a la carga en el enlace.

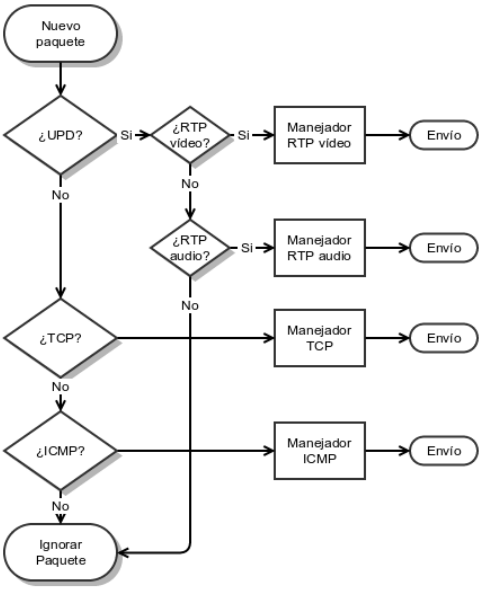


Fig. 1. Procedimiento de clasificación de los flujos de paquetes.

Cada componente está ponderado con un factor que representa la influencia de dicho parámetro sobre el coste final, y dependen del tipo de tráfico que demanda la nueva ruta.

D. Estimación de los parámetros de calidad de servicio por enlace

Para extraer los parámetros objetivos de calidad de cada enlace, es necesario definir cómo calcularlos a partir de la información que proporciona cada conmutador. El cálculo de los parámetros de calidad dará la información necesaria para el cálculo final del coste de cada enlace, y por tanto, de la estimación de la matriz de tráfico.

III. Función de evaluación de costes

Para la evaluación de costes de cada enlace E se propone usar 4 funciones diferentes, cada una relacionada con el tipo de evaluación que realiza: función de evaluación de latencias ($C_{latency}[E]$), función de evaluación de *jitter* ($C_{jitter}[E]$), función de evaluación de pérdidas de paquetes ($C_{loss}[E]$) y función de evaluación de carga ($C_{load}[E]$). Estas cuatro funciones se normalizan para que devuelvan valores de coste entre 0 y 100. Una vez definidas estas funciones de coste, es necesario ajustar los pesos α, β, γ y ω de cada uno de dichos componentes para obtener el coste total estimado para cada tipo de tráfico.

- *Función de evaluación de latencias*. Para este cálculo se empleará la latencia media estimada por un módulo de recolección de estadísticas. La latencia mínima encontrada en la red tendrá asociado un coste de 1. También se tendrá en cuenta la latencia máxima encontrada en la red pues será necesario acotar el coste resultante entre los límites establecidos (1 y 100), siendo en este caso 100 el coste máximo.
- *Función de evaluación de jitter*. La detección de *jitter* se lleva a cabo usando los valores instantáneos y medias de latencia obtenidas. Esta detección tiene ciertas complicaciones existen variaciones que no dependen únicamente del enlace en las medidas, sino que también dependen del tiempo de procesamiento en el controlador, como se comprobó experimentalmente en pruebas preliminares. Otro inconveniente de estas medidas está provocada por la relación entre *jitter* y las latencias extremo a extremo. Por ejemplo, un *jitter* de un milisegundo en una red con latencias promedio de 10 milisegundos tendría una relevancia muy importante. Sin embargo, este mismo *jitter* de un milisegundo en redes con latencias de 500 milisegundos tendrá una importancia mucho menor. No obstante, si solo se consideran *jitter* mínimo y máximo para el cálculo, se obtendrá el mismo coste para el enlace en ambos casos.
  Para solucionar estos dos problemas se establecen las siguientes medidas:
  1) Limitar una diferencia mínima entre los valores máximos y mínimos de *jitter*, a partir de la cual comenzar a calcular costes prefijados (si la diferencia fuera menor, los enlaces tendrían un coste asociado al *jitter* igual para todos).

2) Escalar el coste del *jitter* en función de las latencias máximas y mínimas encontradas, de modo que un *jitter* muy grande en una red de latencias pequeñas (y cercanas), será mucho más significativo en cuanto a coste que un *jitter* grande en una red lenta y de grandes diferencias.

Con estas consideraciones se ha decidido calcular el *jitter* como la ecuación 2:

$$jitter = |latencia_{instantanea} - latencia_{promedio}| \quad (2)$$

- Función de evaluación de pérdidas.
  Para poder evaluar las pérdidas en la red, se usará el porcentaje de pérdidas experimentados en el enlace. Este porcentaje equivaldrá al coste de forma directa, con lo que estará limitado entre 0 y 100. Para obtener el valor asociado al porcentaje de pérdidas se hará uso de las estadísticas mencionadas hasta ahora, y se usará la ecuación 3.

$$C_{loss}(E) = \frac{sentBytes(E) - receivedBytes(E)}{sentBytes(E)} \cdot 100 \quad (3)$$

IV. Recolección de estadísticas

Para recolectar las estadísticas, se hará uso de la API (*Application Programming Interface*) que proporciona el controlador OpenDayLight [3].

El controlador *OpenDayLight* cuenta en su núcleo con un recolector de estadísticas que almacena información sobre todos los elementos de red susceptibles de contar con estadísticas de red (paquetes enviados por un puerto, bytes recibidos por un conmutador, paquetes que pertenezcan a un flujo instalado en un conmutador, etc.). Haciendo uso de este almacén de estadísticas conseguiremos obtener datos que se usarán para la construcción de los costes asociados a cada enlace.

V. Detección y actualización de cambios en la topología

En esta sección se describe el proceso que consigue responder ante cambios en la red tales como caídas de enlaces o nodos. Es uno de los pilares en los cuales se apoya todo el proceso para proveer de QoE y QoS a las aplicaciones que usan los usuarios. Para conseguir el objetivo, el procedimiento diseñado se divide en dos fases: detección y actualización de rutas.

A. *Detección de cambios*

Mediante el procedimiento de detección de cambios de la topología de red, la solución propuesta es capaz de lanzar el procedimiento que actualice la información sobre la red, y recalcular las rutas que sean necesarias. Estos cambios comprenden la modificación de las características de los enlaces, la caída de nodos, etc. Si bien el propio controlador *OpenDayLight* puede detectar los cambios en la topología (existe un módulo capaz de avisar a las aplicaciones de red de cambios en la topología), se decidió prescindir de esta posibilidad, ya que nuestro diseño contempla el tratamiento de los flujos en el mismo módulo donde se reciben los paquetes.

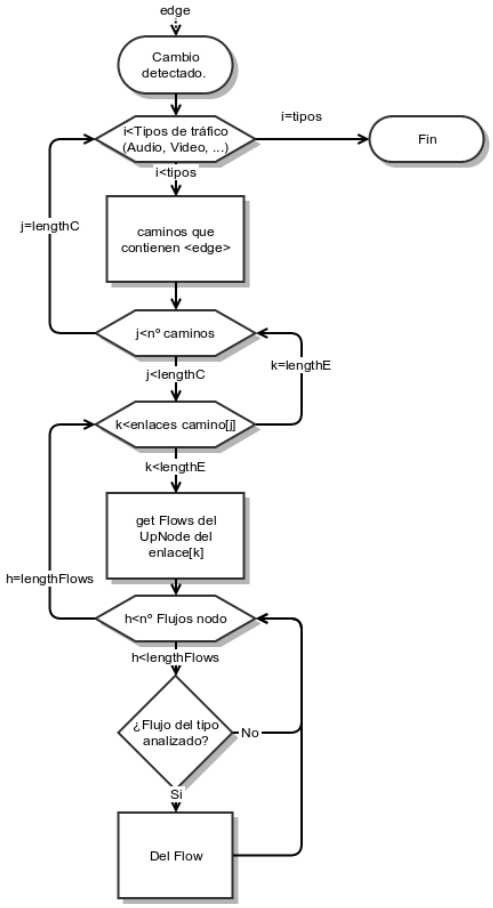


Fig. 2. Diagrama de flujo del proceso de actualización de topología

La detección se lleva a cabo mediante consultas sucesivas al estado de la red. El tiempo entre consultas es el mismo tiempo que el de actualización de estadísticas, t_{update}. El valor de este tiempo de actualización se propone como 100 ms, pudiendo ser modificado según las necesidades de red o usuario. La propuesta está basada en:

- Se establece un período de 100 ms, que en la experimentación preliminar se observó que era suficiente para que se completara la consulta sobre cambios en la topología que proporcionaba ODL.
- No se pretende sobrecargar el controlador eligiendo un tiempo menor, puesto que podría darse una situación en la cual se están actualizando estadísticas, mientras se detectan cambios en la topología, provocando dos accesos a un dato al mismo tiempo. Suponemos que con esta elección de periodos de tiempo se garantiza que el tiempo de procesamiento es mucho menor a la actualización, evitando posibles problemas.

B. *Actualización de rutas*

El proceso de actualización para las rutas sigue el flujo presentado en la figura 2.

1) Se recibe el enlace que ha cambiado según la detección (cuando un nodo cae, todos sus enlaces son detectados y pasan por este proceso de forma individual).
2) Por orden se buscan caminos afectados por cada uno de los 4 tipos de tráfico afectados.
3) Por cada tipo de tráfico:
   a) Se comprueba si alguno de los $lengthC$ caminos que contienen el enlace, están afectados por el cambio.
   b) En caso positivo se recorren los nodos del camino, compuesto de $lenghtE$ enlaces. Suponemos estos caminos ordenados, al haber sido almacenados tras una reordenación.
   c) En cada nodo se comprueban los $lenghtFlows$ flujos instalados buscando flujos coincidentes con los que se deben eliminar. Por cada flujo se comprueba si su acción de envío coincide con el enlace que corresponda del camino. Se podría intentar eliminar solo el enlace afectado, pero en ese caso surgen problemas con el cálculo de caminos extremo a extremo al haber flujos instalados ya.
4) Finalmente se devuelve el control al programa de detección que actualizará mapas y estadísticas de de los manejadores.

VI. Entorno experimental

Para evaluar el funcionamiento del servicio propuesto, se elige el emulador de redes basadas en software Mininet [4], y el controlador OpenDayLight.

Además, se configura mediante guiones programados en Python, la topología representada en la Figura 3. El objetivo es tener distintas rutas posibles a la hora de ejecutar los algoritmos de reencaminamiento, a la hora de adaptarse cuando caiga o se modifique uno de los enlaces.

Además, para comprobar el funcionamiento, se generan tráfico de audio y vídeo mediante la herramienta VLC.

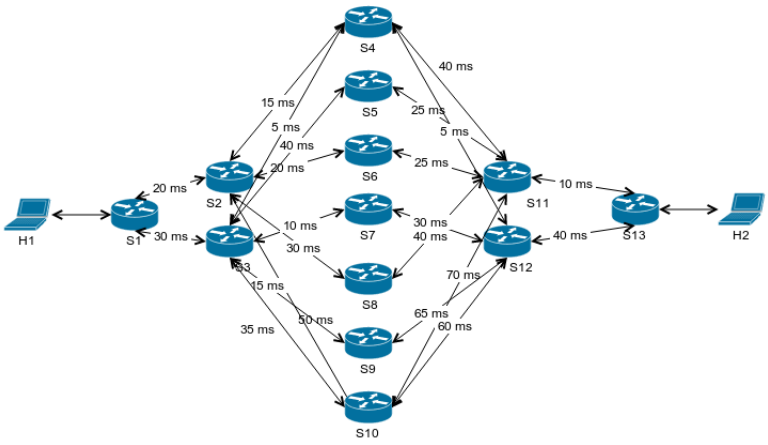


Fig. 3. Topología para realizar evaluación sobre el sistema.

VII. Conclusiones

En este trabajo se presenta el diseño de una aplicación SDN que proporcione calidad de experiencia a flujos con distintos requistos. Para ello, se diseña un módulo para OpenDayLight que monitoriza los enlaces de la red, y reprograman los encaminadores para crear nuevas rutas que cumplan los requisitos de los flujos multimedia.

Resultados preliminares muestra que la implementación permite la atomatización de este servicio en la red, pues se consiga encaminar en tiempo real y reaccionar ante caídas de enlaces, de manera automática. No obstante, aún es necesario ejecutar una batería de experimentos que permitan obtener resultados sólidos.

VIII. Agradecimientos

Este trabajo está parcialmente financiado por el Ministerio de Economía, Industria y Competitividad y el Fondo Europeo de Desarrollo Regional FEDER (proyectos TEC2016-76795-C6-4-R y TIN2013-46223-P).

Referencias

[1] CISCO, "Cisco visual networking index: Forecast and methodology, 2014-2019," May 2015.
[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Dijkstra's algorithm," in *Introduction to Algorithms 2nd edition*. MIT Press, ch. 24, pp. 595–599.
[3] "Página oficial de opendaylight." [Online]. Available: https://wiki.opendaylight.org/view/Main_Page
[4] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: Rapid prototyping for software-defined networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, ser. Hotnets-IX. New York, NY, USA: ACM, 2010, pp. 19:1–19:6. [Online]. Available: http://doi.acm.org/10.1145/1868447.1868466

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6611

Evaluación de la calidad de experiencia de YouTube Live en redes inalámbricas

Luis Jiménez, Marta Solera, Matías Toril, Pablo Oliver.
Dpto. de Ingeniería de Comunicaciones,
ETSI Telecomunicación, Universidad de Málaga,
Campus Universitario de Teatinos, s/n E-29071 Málaga (España).
lrjp@ic.uma.es, msolera@ic.uma.es,mtoril@ic.uma.es,pob@ic.uma.es

***Resumen*—YouTube Live is one of the most popular services on the Internet, enabling an easy streaming of a live video with acceptable video quality. Thus, understanding user´s perception of this service is of the utmost importance for network operators. As in other videostreaming services, YouTube Live traffic is sometimes affected by delays due to unfavourable network conditions, which translate into unacceptable initial reproduction times or image freezes as a result of client´s buffer underrun. Detecting these events is key to ensure an adequate Quality of Experience (QoE). Unfortunately, data encryption makes it very difficult for operators to monitor QoE from packet-level data collected in network interfaces. In this paper, an analytical model to estimate the QoE for encrypted YouTube Live service from packet-level data collected in the interfaces of a wireless network is presented. The inputs to the model are Transport Control Protocol (TCP)/Internet Protocol (IP) metrics, from which three Service Key Performance Indicators (S-KPIs) are estimated, namely initial video play start time, video interruption duration and video interruption. The model is developed with an experimental platform, consisting of a user terminal agent, a WiFi wireless network, a network-level emulator and a probe software. Model assessment is carried out by comparing S-KPI estimates with measurements from the terminal agent under different network conditions introduced by the network emulator.**

***Palabras Clave*—YouTube Live, Streaming, QoE, S-KPIs, Modeling, Pocket, Netem.**

I. INTRODUCCIÓN

En la última década, el incremento exponencial de usuarios y la aparición de nuevos servicios ha traído consigo una completa revolución en las redes de comunicaciones móviles. Se estima que, para el año 2021, existirán 31.750 billones de *smartphones* activos, conectados a diferentes redes [1].

En la actualidad, los operadores se han visto obligados a cambiar sus métodos de gestión de la red, pasando de utilizar indicadores objetivos enfocados en el rendimiento de la red y la calidad de servicio *(Quality of Service,* QoS), a indicadores más modernos y centrados en la opinión del usuario y la calidad de experiencia *(Quality of Experience,* QoE). La gestión de calidad de experiencia tomará, si cabe, una mayor importancia con la incorporación de la quinta generación de tecnologías de telefonía móvil (5G), cuyo lanzamiento se prevé para 2020, y que estará claramente dominada por los servicios de vídeo que representarán el 70% de la demanda de tráfico total [1][2]. En un entorno en el que la oferta de redes y servicios es similar en todos los operadores, la calidad de experiencia se convierte en uno de los principales factores que diferencia a unos operadores de otros y que permitirá fidelizar a los usuarios.

Entre todos los servicios, la reproducción de vídeos a través de la descarga progresiva (*video streaming*) es la aplicación que más tráfico genera en Internet en la actualidad [3]. De las diferentes técnicas, la más popular y la que proporciona una mayor accesibilidad desde cualquier punto de la red es la que utiliza los protocolos HTTP/HTTPS (*HyperText Transfer Protocol/ Secure*). En ella, los vídeos se almacenan en segmentos de diferentes longitudes (de 2 a 10 segundos normalmente), que se codifican con distintas resoluciones (regímenes binarios). En este tipo de *streaming*, el usuario es quien solicita a través de un mensaje HTTP el contenido multimedia, y la descarga del vídeo comienza como respuesta a esta solicitud. Durante la descarga, se cede el control de la descarga al cliente, que va solicitando los segmentos de vídeo mediante mensajes HTTP adaptándose a las fluctuaciones de las condiciones de la red. Las principales plataformas de servicio, tales como *YouTube, Hulu* y *Netflix*, emplean este tipo de *streaming*, siendo el primero el líder indiscutible del mercado.

YouTube ha añadido recientemente una nueva funcionalidad consistente en ofrecer secuencias de vídeo de alta calidad en directo (*Live video streaming*), causando un aumento exponencial en la generación de contenido por parte de los usuarios. Por esta razón, es necesario comprender las características del tráfico generado por este nuevo servicio, para poder monitorizar

y controlar la QoE percibida por sus usuarios [4].

Tradicionalmente, la QoE se ha evaluado mediante pruebas subjetivas realizadas con observadores reales, que reflejan su grado de satisfacción mediante un indicador de puntuación medio de opinión (*Mean Opinion Score,* MOS) [5]. Este tipo de enfoque es complejo, requiere tiempo, y no es válido para monitorización a gran escala. Por ello, en los últimos años se han estudiado nuevos métodos para estimar la calidad de experiencia a partir de indicadores de rendimiento construidos con datos recolectados en los equipos e interfaces de la red. Especialmente prometedores son los métodos basados en la información recolectada a nivel de paquete por sondas de nivel de red (habitualmente Internet Protocol, IP), ubicadas en las interfaces de la red. Dichos métodos son posibles gracias a que, en los servicios que utilizan como protocolo de transporte a TCP (*Transport Control Protocol*), los mecanismos de control de congestión y flujo hacen que los indicadores de rendimiento de nivel de red (p. ej., caudal, tasa de pérdidas, retardo medio) sean un reflejo de la calidad de servicio extremo a extremo. La principal dificultad estriba en identificar la relación existente entre los indicadores de rendimiento específicos del servicio (*Service Key Performance*, S-KPI), (en el caso del *videostreaming*, p. ej., tiempo inicial de espera para la reproducción, número y duración de las interrupciones de la reproducción, ..) y las métricas TCP/IP.

Hasta la fecha, los S-KPIs para el servicio de *videostreaming* se han venido obteniendo mediante la identificación de las distintas fases de la sesión a partir del análisis de los mensajes del protocolo HTTP. Sin embargo, dicho análisis ya no es posible, ya que, desde 2016, el 97% por ciento del tráfico de *YouTube* (video convencional y *Live streaming*) se cifra mediante conexiones HTTPS con *Transport Layer Security* (TLS) y *Secure Sockets Layer* (SSL). La situación se ha complicado aún más con la inclusión de técnicas de *streaming adaptativo* (*Dynamic Adaptive Streaming over HTTP*, DASH). Ambos procesos dificultan enormemente la estimación de los S-KPIs de *YouTube*. En [6], se presenta un modelo de QoE para el servicio de *YouTube* convencional, basado en la estimación del nivel de buffer del cliente. Sin embargo, hasta donde se sabe, ningún trabajo ha propuesto un modelo analítico sencillo para estimar los S-KPIs del servicio de *YouTube Live* a partir de métricas TCP/IP recolectadas en las interfaces de una red inalámbrica.

En este artículo, se presenta un modelo de regresión para estimar los principales indicadores de rendimiento del servicio de *YouTube Live* a partir de medidas obtenidas del análisis de paquetes capturados en las interfaces de una red inalámbrica. Las entradas del modelo son métricas TCP/IP comunes (p. ej., caudal, tasa de pérdida de paquetes o tiempo de ida y vuelta), a partir de las que se estiman tres de los S-KPIs más importantes del servicio, como son el retardo inicial de reproducción, el número total de interrupciones y el tiempo total de interrupción. El modelo se desarrolla utilizando técnicas de regresión sobre datos tomados con una plataforma experimental, consistente en un terminal de usuario, una red de acceso inalámbrica WiFi y un emulador de red. Esta plataforma permite: a) automatizar la creación y emisión de un *Live video streaming*, b) emular la interacción del usuario con el *Live video streaming* a través de un *smartphone*, y c) modificar las condiciones de la red mediante el emulador de red. La validación del modelo se lleva a cabo comparando las estimas de los S-KPIs con las medidas de los mismos realizadas por el agente de usuario bajo diferentes condiciones de red establecidas con el emulador de red. El resto del artículo se estructura de la siguiente manera. La sección II describe la plataforma experimental. La sección III explica el modelo de rendimiento del servicio de *YouTube Live*. La sección IV muestra las curvas de regresión con las que estimar los S-KPIs del servicio con métricas TCP-IP, que es la principal contribución. Por último, la sección V, expone las conclusiones del trabajo.

II. PLATAFORMA EXPERIMENTAL DE PRUEBAS

La Fig.1 muestra un esquema de la plataforma utilizada para automatizar la toma de medidas necesarias para la construcción de los modelos de QoE. La plataforma consta de 2 módulos: un primer módulo (izquierda de la figura) encargado de la emisión en directo de una secuencia de vídeo (*Live video streaming*) de *YouTube*, y b) otro módulo (derecha de la figura) encargado de la recopilación y análisis de las medidas. El módulo de emisión consta únicamente de un PC ejecutando la aplicación *Wirecast*. Por su parte, el módulo de medidas se compone de un terminal móvil conectado por WiFi a un PC con salida directa a Internet. En el terminal móvil, se ejecuta la aplicación que funciona como agente de usuario (*TEMS Pocket*), que modela las interacciones del usuario durante la sesión de *videostreaming*. Esta aplicación permite además recoger las medidas reales de los S-KPIs, al tener acceso a uno de los extremos de la comunicación. En el PC, se ejecuta un emulador de red (*NetEm*) para modificar de forma controlada las condiciones de red (p. ej., ancho de banda disponible, retardo medio y/o tasa de pérdidas de paquetes). Las medidas de red a nivel de paquetes se toman de la interfaz del emulador de red hacia Internet. Esta información se procesa en tiempo diferido con una aplicación de análisis y monitorización de tráfico (*Network Probe*), con la que se obtienen las estimas de los S-KPIs. Dichas estimas construidas con el modelo propuesto se contrastan con las medidas tomadas por el agente de usuario. A continuación, se describen cada uno de estos elementos de la plataforma, detallando el proceso de generación, modificación y captura del tráfico en la plataforma.

A. *Emisión de vídeo Live Streaming*

El proceso de emisión permite la creación de un *Live video streaming* utilizando la plataforma de *YouTube*. Como punto de partida, se requiere un equipo que genere el contenido multimedia a distribuir en tiempo real. En

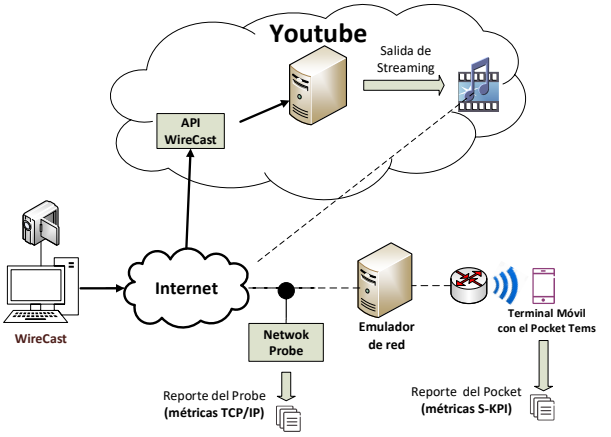


Fig. 1. Plataforma experimental de pruebas.

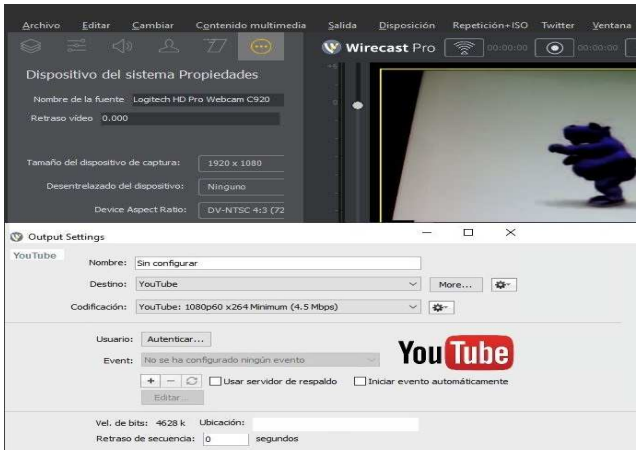


Fig. 2. Configuración *Wirecast Pro* 7.3

la plataforma, se utiliza un servidor de *Live Streaming* consistente en un PC *DELL PowerEdge* T430 con dos procesadores Intel(R) Xeon (R) E5-2630 V3 @ 2.40GHz, cada uno con 8 núcleos. Posee un disco duro STD de 1.4 TB, sistema operativo *Microsoft Windows Server 2016 Datacenter* ver. 10.0.14933 con 64GB de memoria RAM y tarjeta multimedia Matrox® G200 integrada con iDRAC8. El servidor está conectado a un punto de red aislado para evitar variaciones indeseadas del ancho de banda disponible (velocidad de subida de 94 Mbps) para el envío del flujo multimedia a *YouTube*. En el servidor se instala el software *Telestream Wirecast* y la cámara web. *Wirecast* es un software empleado para la emisión del *videostreaming*. Permite producir y transmitir eventos multimedia en tiempo real hacia *YouTube* desde una estación de trabajo [7]. La versión de *Wirecast* utilizada en la plataforma experimenta es *Wirecast* Pro 7.3 de 64 bits, con capacidad de emisión de vídeo en alta definición (HD). Para la captura del vídeo en directo, *Wirecast* emplea un dispositivo externo, que en la plataforma utilizada es una *webcam Logitech HD Pro* C920, que permite realizar grabaciones de alta definición (HD 1080p). Para este dispositivo, debe configurarse el formato de vídeo, el compresor de vídeo y el régimen binario de salida. En este trabajo, se selecciona la resolución 1920x1080, el formato progresivo, el codificador H.264 AVC y el régimen binario de 3-6 Mbps (perfil 1080p). Tras configurar el flujo de vídeo de subida, se debe crear un evento de emisión en la web de *YouTube* iniciando sesión en el servicio *Creator Studio de YouTube*. Este servicio permite organizar el canal donde se efectúa la transmisión, la creación del evento que proporciona el identificador del vídeo (ID) y la gestión de algunas características del flujo de vídeo en tiempo real (p. ej., codificador de audio y vídeo) [8]. Al crear el evento, se deben definir diversos parámetros, como su nombre, la hora de inicio de emisión, las características del flujo de vídeo de bajada y las características tipo de evento (Público, Privado, Oculto). El evento Público distribuye el ID del *Livestreaming* en las listas en directo de *YouTube* a nivel global permitiendo que cualquier usuario de la red pueda acceder al *Live streaming*, en el Privado y el Oculto se necesita obligatoriamente el ID del evento para acceder al *Live streaming*. En las pruebas realizadas, utilizamos un evento oculto dado que nos proporciona un menor nivel de encriptación manteniendo reservado el *Live streaming*.

La Fig.2 muestra el software *Wirecast* con la configuración seleccionada para el flujo de vídeo de bajada, igual al de subida (1920x1080p, H.264, 3 Mbps mínimo). Dicho evento se enlaza a la captura en directo. Posteriormente, *YouTube* transcodifica el contenido, recibido por el enlace de subida a una tasa de bits determinada, creando un flujo principal con el régimen binario establecido al seleccionar el codificador en la configuración del evento. No obstante, *YouTube* hace varias réplicas del flujo con diferentes regímenes binarios para que la emisión del *Live video streaming* esté disponible para todo tipo de usuarios, regulados por la capacidad de sus terminales y las condiciones de la red que utilicen para el acceso al contenido multimedia del *Live Streaming* [8].

B. Recepción de vídeo Live Streaming

El tráfico de *Live video streaming* se genera creando peticiones de visualización al evento público de emisión creado anteriormente. Para ello, se utiliza un *smartphone* con la aplicación *Tems Pocket* ver. 16.3, que se encarga de emular la interacción del usuario con el *Live video streaming*. Dicha aplicación se emplea también para la toma de medidas de los S-KPIs.

El terminal se conecta vía WiFi a Internet. La red inalámbrica está formada por un punto de acceso inalámbrico, configurado en modo puente, para interconectar los dispositivos WiFi a la subred de medición. La conexión entre el punto de acceso inalámbrico y la subred se realiza mediante un cable de par trenzado, conectado desde el punto de acceso hacia una de las tarjetas eth0 del PC que contiene el emulador de red.

C. Modificación de las condiciones de red

Para modificar las condiciones de red, se utiliza el emulador de red *Netem* [9], incluido en el kernel de

Linux desde la versión 2.6. Con él, se pueden introducir efectos controlados sobre la subred, tales como retardo, pérdida, duplicación y reordenamiento de paquetes. En *Netem*, el retardo de paquetes y el *jitter* se describen por el valor medio, la desviación estándar y el coeficiente de correlación. Por defecto, se utiliza una distribución uniforme para el retardo, que puede ser sustituida por otras funciones, tales como Pareto, Pareto-normal, normal o distribuciones personalizadas creados a partir de datos experimentales o de simulación. En la plataforma, se instala *Netem* sobre un PC con un procesador i5-750 a 3 GHz, 8 GB de RAM y sistema operativo Ubuntu 16.04 LTS 64 bits. Dicho PC incluye dos tarjetas de red enlazadas mediante una tabla de enrutamiento, para proporcionar el acceso a Internet a los dispositivos conectados a la subred inalámbrica.

Para recrear las distintas condiciones de red, se configuran los parámetros de *Netem* con el comando "tc" de acuerdo a la siguiente sintaxis [10][11]:

tc qdisc ... dev DEVICE] add netem OPTIONS

OPTIONS := [LIMIT] [DELAY] [LOSS] [RATE]
LIMIT := limit packets
DELAY := delay TIME [JITTER [CORRELATION]]]
LOSS := loss PERCENT
RATE := rate RATE ,

donde *qdisc*(*queuing discipline*) es la abreviatura de la cola asociada al dispositivo de interfaz a través del cual se envían los paquetes. Con respecto a los parámetros OPTIONS, LIMIT acota el efecto de otras opciones seleccionadas al número indicado de paquetes siguientes, DELAY añade el retardo medio elegido en milisegundos a los paquetes que salen a la interfaz de red elegida, JITTER se utiliza para cuantificar la variación de retardo en milisegundos, CORRELATION es un porcentaje que controla cuánto depende el valor de retardo actual del anterior, LOSS es la probabilidad de pérdida de paquetes (expresada en porcentaje) y RATE limita la tasa binaria de transmisión mediante un proceso de *throttling* (en kbps) [12]. En las pruebas realizadas, sólo se ajustan los parámetros *delay, loss* y *throttling* por simplicidad.

D. Recopilación de medidas

Tal como se refleja en la Fig.1, en la plataforma se habilitan dos puntos principales de medida: terminal y salida de Internet. El primero se dedica a la recopilación de estadísticas de rendimiento de la transmisión de paquetes a nivel de red (métricas TCP/IP), mientras que el segundo se dedica las medidas de los S-KPIs.

Medidas de nivel de red

Para la obtención de las métricas TCP/IP, se utiliza la herramienta *Network Probe*, que es un software propietario destinado al análisis y monitorización del tráfico. Este proceso se realiza mediante el análisis en tiempo diferido de archivos ".pcap" con las trazas de tráfico capturadas en la red. Dichos archivos de trazas se generan con la herramienta de código abierto *Tcpdump* [13].

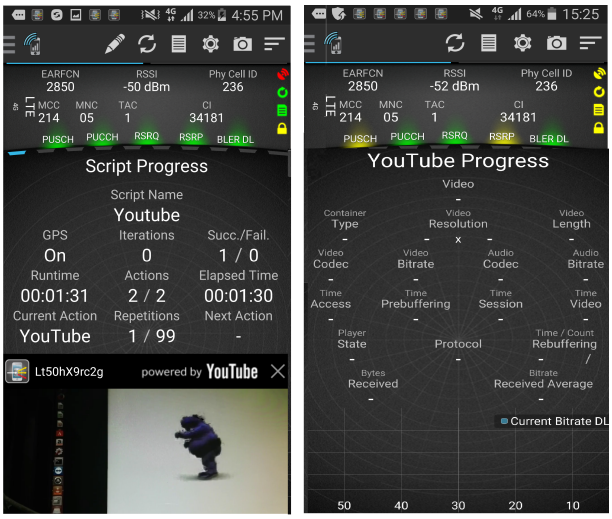


Fig. 3. Pantalla *Tems Pocket* mientras se ejecuta el *YouTube script.*

A partir de los ficheros de trazas, la herramienta *Network Probe* genera archivos de medidas en formato *.csv*, que incluyen las principales métricas de la transmisión de paquetes, desglosadas por usuario, conexión y ráfaga de paquetes.

Medidas de los S-KPIs

Para la toma de medidas de los S-KPIs, se utiliza el software *TEMS Pocket* versión 16.3. Dicha herramienta es un agente de usuario utilizado para la verificación, el mantenimiento y la resolución de problemas de redes móviles [14]. *TEMS Pocket* permite la creación de diferentes programas para automatizar pruebas de servicios tales como Facebook, Instagram, Twitter, WhatsApp, YouTube, etc. En el caso de la evaluación del servicio de *YouTube Live Streaming*, se utiliza la opción de *YouTube* convenientemente configurada para obtener un informe con las estadísticas de los diferentes S-KPIs. La Fig.3 muestra dos pantallas disponibles en el terminal cuando se efectúa una medición. En ella, se observa el tiempo de progreso del *script*, una pequeña pantalla que muestra el *Live video* a evaluar, los valores de los S-KPIs para cada segmento de *streaming* en curso.

E. Automatización

El proceso de automatización debe incluir la configuración automática del agente de usuario, el emulador de red y la captura de tráfico a nivel de paquetes. Dichos procesos deben realizarse de forma sincronizada para permitir posteriormente el procesado de las medidas.

Agente de usuario

La Fig.4 presenta las distintas pantallas de configuración del *TEMS Pocket*. La primera de ellas (*Script settings*) se utiliza para la creación del *script* encargado de descargar y reproducir el flujo multimedia de la sesión de *Live video streaming*. Sus principales opciones son: a) Video indica el identificador (ID) del canal de vídeo streaming a reproducir y analizar; b) *Streaming duration* (SD) indica la

duración de un periodo de medición, es decir, por cuánto tiempo se van a recoger estadísticas de los S-KPIs; c) *Preguard* (PG), *Postguard* (PTD) indican periodos de guarda que se insertan automáticamente antes y después de la medición, respectivamente; el propósito de los periodos de guarda es asegurar que el establecimiento y la liberación de la sesión de vídeo se grabe en el archivo de registro. Para *YouTube*, el valor recomendado para ambos es de 10 seg [14]; d) *Repeat action* (RA) indica el número total de veces que se ejecuta todas las opciones del *script*.

La segunda pantalla (*Actions*) indica las acciones que realiza un *script* determinado. En este trabajo, se seleccionan dos acciones: *YouTube script*, descrito anteriormente, y *Log file recordings*, que permite grabar en un solo archivo las medidas realizadas separadas por cada lazo de repetición, para su posterior post-procesamiento.

La tercera pantalla 3 (*YouTube settings*) es el *script* general conformado por *script setting* y *actions*. Además, posee la opción de establecer la cantidad total de veces que se ejecuta el *script* (*Max interation*s, MI). El total de mediciones obtenidas depende de las veces que se repita el *script setting* (RA) multiplicadas por las veces de ejecución del *script* general (MI).

$$E = MI \cdot RA \cdot (SD + PG + PTD) \qquad (1)$$

Emulador de red

La Fig.5 muestra el *script* utilizado para la configuración automática de *NetEm*. En el *script*, se definen como parámetros de entrada el retardo promedio y el *jitter* en milisegundos ($2, $3), la tasa de pérdida de paquetes expresada en % ($4) y la tasa de transmisión máxima del enlace descendente en kbps (*throughput*, $5). El comentario de la línea inicial indica el tipo de *shell* utilizado para interpretar el *script* (/bin/bash). La línea 3 se emplea para eliminar cualquier regla establecida anteriormente, ejecutando el comando *tc* con la opción "del". Posteriormente, la línea 6 establece los valores de retardo, *jitter* y tasa de pérdidas a los designados en los parámetros de entrada $2 - $4 para el próximo millón de paquetes. Las líneas 9-11 posteriores fuerzan la máxima tasa de transmisión (*throttling*) al valor del parámetro $5. Adicionalmente, el *script* contiene algunas líneas cuyo propósito es facilitar la monitorización, reflejando por pantalla cada cambio de las condiciones de red (Líneas 5 y 8) y almacenando en un fichero de salida la configuración de *NetEm* correspondiente junto a su fecha y tiempo de inicio (*timestamp*) (Línea 13). Conviene aclarar que la interfaz controlada por *NetEm*, denominada eth0, corresponde en la Fig.1 a la interfaz del emulador de red hacia el punto de acceso inalámbrico.

La Fig.6 muestra el *script* empleado para automatizar los cambios en la configuración de *NetEm* mientras se realiza la medición con el terminal móvil. Básicamente, consta de una matriz que contiene todas las combinaciones de los valores deseados para cada parámetro de entrada (THROUGHPUT, PACKET LOSS, DELAY). Para reducir el número de configuraciones simuladas, el parámetro

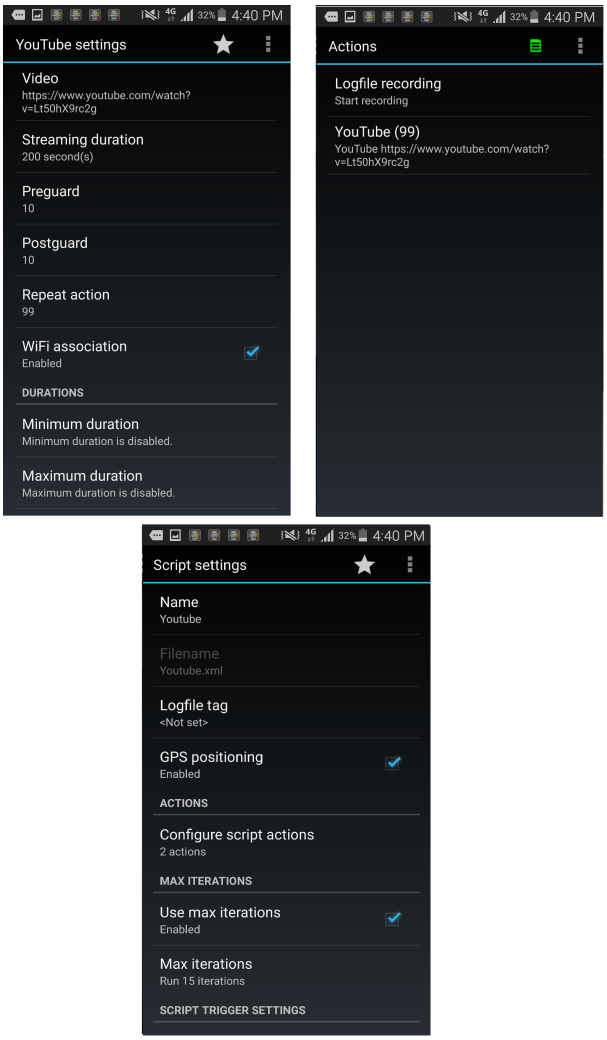


Fig. 4. Configuración del *Pocket script*.

JITTER se fija a 0. Así, cada efecto producido por el *script* de configuración de *Netem*, mostrado en la Fig.1, corresponde a una combinación de los tres parámetros indicados en la matriz (Líneas 7-12). A continuación, se establece el tiempo duración de cada efecto (1980 segundos, 33 minutos) y se realiza la captura de tráfico por este intervalo de tiempo usando *tcpdump* (línea 14). Con la sintaxis utilizada, los archivos de captura ".pcap" generados por cada efecto se almacenan con un nombre de acuerdo al siguiente formato: "Nombre dado al *script* (ej., 1.Csv)" _ "Combinación del efecto (DELAY, PACKET_LOSS, THROUGHPUT) "_ "%F (Fecha)"_ "%T (Hora)". La interfaz de captura, denominada eth1, corresponde a la interfaz entre el emulador de red e Internet. La opción **-G** se usa para indicar el tiempo de duración del proceso de captura, **-W** limitar el número de archivos creados. Al inicio y al final del *script* de automatización, se eliminan las reglas establecidas para garantizar que cada medida obtenida corresponde a un efecto deseado (líneas 3 y 20).

III. METODOLOGÍA EXPERIMENTAL

Esta sección se describe el proceso llevado a cabo para capturar las trazas de datos necesarias para estimar los S-KPIs de un servicio encriptado de *Live video streaming* de *YouTube Live* a partir de las métricas TCP/IP. En primer lugar, se describen los S-KPI más relevantes para este servicio. En segundo lugar, se explica cómo se llevó a cabo la batería de pruebas.

A. Definición de los S-KPI

La experiencia del usuario en *YouTube* (convencional y *Live streaming*) se caracteriza por tres problemas básicos: el retardo inicial de reproducción del video (*Initial buffer time*), la detección de la reproducción del video (*rebuffering /stalling event*) y la duración de este estancamiento [15][16]. Para analizar estos problemas se seleccionan tres S-KPI que nos ofrece la herramienta *TEMS Pocket*: *Streaming Video Play Start Time* (*Initial Buffering Time*), *Streaming Video Interruption Duration* y *Streaming Video Interruption Count* [17].

- Retardo inicial de reproducción (*Streaming Video Play Start Time* , SPT): Es el tiempo desde que el usuario envía la petición para iniciar el *streaming* (click en el ID) hasta que aparece la primera imagen de vídeo en la pantalla.
- Frecuencia de interrupción de reproducción (*Streaming* Video Interruption Frequency, IF): Es la frecuencia con la que se interrumpe la reproducción de vídeo en una sesión de *streaming* por razones de *rebuffering*, calculada a partir del número total de interrupciones en la sesión (*Interruption Count*, IC) y la duración del vídeo reproducido, SD (en minutos), como

$$IF = IC/SD \ [1/min]. \tag{2}$$

- Ratio de duración de interrupción de reproducción (*Streaming Video Interruption Duration Ratio*, IDR): Se define como el cociente entre el tiempo total de interrupciones en una sesión de *streaming* por razones de *rebuffering* (*Interruption Duration*, ID) respecto al tiempo total de la sesión, calculado a partir de la duración del vídeo reproducido (*Streaming Duration*, SD), como

$$IDR = ID/(ID + SD). \tag{3}$$

B. Batería de pruebas

Para identificar qué métricas impactan sobre el rendimiento de un servicio *Live video streaming* de *YouTube*, se realizaron baterías de pruebas durante el periodo comprendido del 7 al 15 de abril del 2017. Estas fechas coinciden con una época de vacaciones y se eligieron para tener las mayores prestaciones de red disponibles.

La batería de pruebas consiste en la emisión de un *live streaming* con una resolución de 1080p (la máxima permitida por la aplicación) desde el servidor local, que tiene instalado el software *Wirecast* y usa la plataforma de *YouTube* como pasarela para la masificación del *Live video*. La configuración exacta se detalla en la sección 2 . El servidor está conectado a Internet por un enlace de 94 Mbps. Eso garantiza un ancho de banda suficiente para transmitir sin interrupciones el *video live streaming*.

```
#!/bin/bash

sudo tc qdisc del dev eth0 root

echo "Changing parameters: Delay=$2ms +- $3ms, loss = $4%"
sudo tc qdisc change dev eth0 root netem delay "$2"ms "$3"ms
 loss "$4"% limit 1000000

echo Setting throughput limit to "$5"Kbps
sudo tc qdisc add dev eth0 handle 1: root htb default 11
sudo tc class add dev eth0 parent 1: classid 1:1 htb rate 1000Mbps
sudo tc class add dev eth0 parent 1:1 classid 1:11 htb rate "$5"Kbit

echo `date --utc +%s`",$2,$3,$4,$5" >> $1.csv
```

Fig. 5. *Script* de configuración de *Netem*.

```
#!/bin/bash

sudo tc qdisc del dev eth0 root

echo "TIMESTAMP,DELAY,VARIABILITY,LOSS,THROUGHPUT" > $1.csv

for THROUGHPUT in 250 500 1000 2000 4000
  do
  for PACKET_LOSS in  0 0.75 1.5 3
      do
      for DELAY in 0 50 100 200 400
        do
            ./netem_and_log_alldl.sh $1 $DELAY 0 $PACKET_LOSS $THROUGHPUT
           sudo tcpdump -w "$1"_"$DELAY"_"$PACKET_LOSS"_"$THROUGHPUT"_%F
                        _%T.pcap -i eth1 -G 1980 -W 1 'not port 22'

      done
  done
done

#Back to normal
sudo tc qdisc del dev eth0 root
```

Fig. 6. *Script* de automatización de *Netem*.

Una vez lanzado el *Live video streaming*, se ejecuta de forma sincronizada el emulador de red y el cliente. La configuración de red se modifica de la siguiente manera:

- Tasa de pérdida de paquetes [%]: 0, 0.75, 1.5, 3.
- Retardo medio de paquete [ms]: 0, 50, 100, 200, 400.
- Máximo ancho de banda disponible (*throughput*) [kbps]: 250, 500, 1000, 2000, 4000.

Dada la matriz anterior, se obtiene un total de 100 configuraciones de *Netem* (efectos). Cada configuración se mantiene durante 33 min (1980 s). El tiempo total de la batería de pruebas es de 55 horas (3.300 min). Durante todo este tiempo, se capturan los datos agrupados en dos clases de métricas:

- Reporte del *Pocket*: Es el archivo generado por el terminal móvil que contiene los S-KPIs medidos y organizados por efecto de *Netem* generado.
- Datos recogidos en el *Network Probe*: Es el archivo que contiene métricas TCP/IP para cada una de las configuraciones de red. Este archivo es generado por el procesado *offline* de los todo los ".pcaps" capturados en la interfaz eth1. El intervalo de tiempo de captura de los ".pcaps" se determina por el *script* de automatización de *Netem*.

IV. RESULTADOS

A continuación, se desglosan los resultados del análisis de regresión entre las medidas de *throughput* medio (THRU) de sesión obtenidas con la sonda de red (*Network Probe*) y las medidas de los distintos S-KPIs

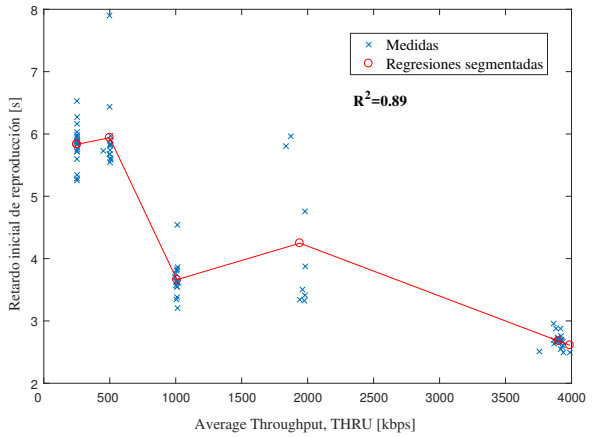


Fig. 7. Gráfica regresión SPT frente tasa de datos disponible.

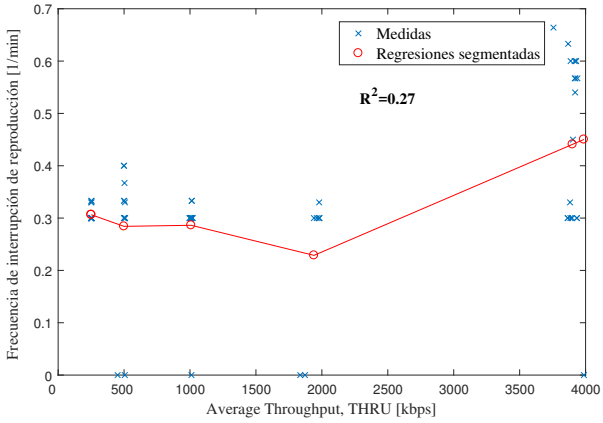


Fig. 8. Gráfica de regresión IF frente tasa de datos disponible.

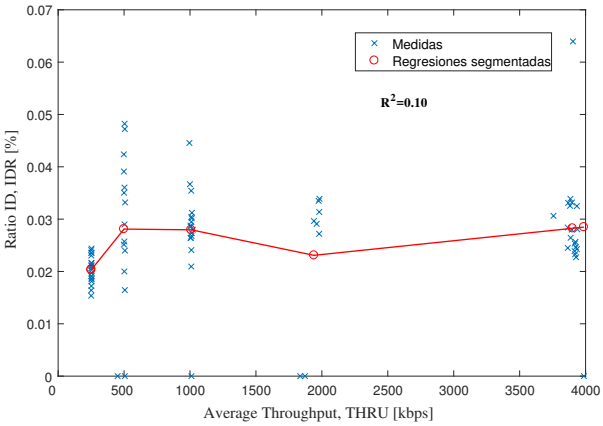


Fig. 9. Gráfica de regresión de IDR frente tasa de datos disponible.

recogidas por el agente de usuario (*Tems Pocket*). Las curvas de regresión que aquí se presentan son específicas del servicio de *Live Streaming*, del que no se han publicado resultados similares.

A. *Retardo inicial de reproducción*

En la Fig.7 se muestra un grafo de dispersión que relaciona el *throughput* medio con el retardo inicial de reproducción (*Streaming Video Play Start Time*, SPT). Al mismo tiempo, se superpone la curva de regresión segmentada que mejor ajusta los datos. En la nube de puntos, puede observarse cómo los puntos se agrupan por columnas, debido a la limitación de la velocidad disponible del canal establecido con el emulador de red (250, 500, 1000, 2000, 4000 kbps). La curva de regresión muestra cómo, en general, el SPT tiende a aumentar cuando se reduce el *throughput* medio. El principal incremento se produce al reducir el THRU por debajo de 1 Mbps. Aun así, llama la atención de que el SPT no se incremente de forma gradual, e incluso se decremente cuando pase el THRU de 2 a 1 Mbps. También llama la atención la gran dispersión de valores de SPT con THRU = 2 Mbps. Este resultado anómalo puede justificarse porque el *Tems Pocket* seleccione una calidad (resolución) del *Live video* al inicio de cada sesión de medida (efecto) en función del ancho de banda disponible para el usuario, lo que disminuiría el impacto de la reducción del THRU producida con el emulador de red. En cualquier caso, el modelo de regresión segmentado construido con las medidas de THRU de la sonda se ajusta razonablemente bien a las medidas del SPT realizadas con el TEMS, con un coeficiente de determinación de $R^2 = 0.89$.

B. *Frecuencia de interrupción de reproducción*

En la Fig.8 se muestra el grafo de dispersión entre el *throughput* medio y la frecuencia de interrupción (*Streaming Video Interruption Frequency*, IF). Este S-KPI contabiliza el número de veces que la reproducción del video se detiene. Siguiendo el mismo procedimiento anterior, se construye la curva de regresión a partir de la nube de puntos. A primera vista, se aprecia que, inesperadamente, cuanto mayor es el THRU, mayor es la frecuencia de interrupción. Este comportamiento anómalo se justifica por el mecanismo de adaptación de la resolución de vídeo introducido por DASH en *YouTube*. Cuando las condiciones de red son buenas, el agente de usuario selecciona una resolución de vídeo mejor, que conlleva una mayor tasa de transmisión. Con esta selección, si existe algún problema de congestión en la red, el *buffer* de reproducción se consume rápidamente, debido a la elevada tasa de codificación de vídeo, antes de que se descarguen los siguientes segmentos. Eso explica el por qué, aunque la capacidad del canal sea mayor, el número de interrupciones es mayor que en el caso de un ancho de banda más pequeño. Con peores condiciones de canal, los datos almacenados en el *buffer* se consumen más lentamente, disminuyendo la probabilidad de que el *buffer* de reproducción se vacíe. En cualquier caso, los valores de IF son bajos, inferiores a 0.7 interrupciones por minuto, lo que demuestra la capacidad de DASH para eliminar los problemas de interrupciones. Como se aprecia en la Fig.8, el modelo de regresión construido recoge la relación directa entre IF y THRU, aunque el coeficiente de correlación entre las estimas y las medidas de IF es sólo de $R^2 = 0.27$.

C. Ratio de duración de interrupción de reproducción

Por último, en la Fig.9 se muestra el grafo de dispersión que relaciona el *throughput* medio con el ratio de duración de interrupción (*Streaming Video Interruption Duration Ratio*, IDR). Este indicador refleja la duración total de las interrupciones debidas a que el *buffer* del reproductor se vacía. El análisis de regresión muestra una relación inversa entre las variables. Se observa cómo la tendencia es que a mayor THRU, menor es el tiempo de duración de las interrupciones. Este resultado, que de nuevo parece contraintuitivo, tiene también su explicación en los mecanismos de adaptación de la resolución del vídeo introducidos por DASH. Cuando empeoran las condiciones de canal, como resultado del aumento del retardo de transmisión, las pérdidas de paquetes o el *throttling*, el agente de usuario selecciona una resolución menor con un régimen binario que garantice la recepción de paquetes. Con ello, se reducen significativamente la duración de las interrupciones. Por el contrario, cuando el THRU es más alto, se elige una resolución de vídeo mayor, lo que aumenta la sensibilidad del reproductor a cambios bruscos de la tasa de transmisión. En cualquier caso, los valores de IDR son bajos, inferiores a 0.07 % del tiempo total de la sesión. Un análisis exhaustivo de las medidas (no mostrado aquí) refleja que la duración máxima de la interrupción es de 160 ms (equivalente a sólo 4 fotogramas con una frecuencia de cuadro de 25 Hz). El modelo de regresión construido recoge la relación directa entre IDR y THRU, aunque el coeficiente de correlación entre las estimas y las medidas de IF es sólo de $R^2 = 0.10$.

V. CONCLUSIONES

En este artículo se ha presentado un modelo de regresión para estimar los principales indicadores de rendimiento del servicio de *Live streaming* cifrado de *YouTube* en una red inalámbrica. El modelo se ha construido utilizando una plataforma experimental, compuesta por una estación de trabajo emisora, un terminal móvil reproductor, un emulador de nivel de red y una sonda de nivel de red. Con esta plataforma, se han capturado trazas de nivel de red y de usuario tras una batería de pruebas con más de 3000 minutos de reproducción de vídeo y 100 configuraciones diferentes del emulador de red. El análisis de los datos recogidos muestra la correlación existente entre los indicadores de rendimiento S-KPIs seleccionados y las métricas TCP/IP de la red. A partir de las curvas de regresión mostradas, pueden construirse modelos de QoE para el servicio de *Live Streaming* basados únicamente en métricas TCP/IP. Estos modelos de caja negra son la única opción de los operadores de red para monitorizar la QoE de servicios multimedia cifrados a gran escala.

Los resultados demuestran de manera clara el efecto beneficioso del DASH sobre el rendimiento del *Live streaming*, gracias a la disminución de las interrupciones. Estos resultados son coherentes con las estadísticas de QoE del servicio de *YouTube* convencional con DASH sobre móviles presentadas en [6]. De los resultados, también se deduce que se hacen necesarios nuevos indicadores de rendimiento del servicio que estimen el nivel medio de calidad de imagen ofrecida al usuario durante la sesión. En este sentido, conviene precisar que el agente de usuario utilizado en la campaña de medidas no recoge actualmente estadísticas del formato de vídeo reproducido. Actualmente, se trabaja en el descifrado y análisis del tráfico de aplicación para obtener dicha información.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad (Proyecto TEC2015-69982-R, UNMA13-1E-1864), y FEDER.

REFERENCIAS

[1] A. Ericsson, "Ericsson mobility report: On the pulse of the networked society", *Ericsson, Sweden, Tech. Rep. EAB-14*, vol. 61078, 2015.

[2] N. Alliance, "Next generation mobile networks recommendation on son and o&m requirements", *Req. Spec. v1*, vol. 23, 2008.

[3] I. Cisco, "Cisco visual networking index: Forecast and methodology, 2011–2016", *CISCO White paper*, pp. 2011–2016, 2012.

[4] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analysis and modelling of youtube traffic", *Transactions on Emerging Telecommunications Technologies*, vol. 23, no. 4, pp. 360–377, 2012.

[5] A. Diaz, P. Merino, and F. J. Rivas, "Customer-centric measurements on mobile phones", in : IEEE *Int. Symp. on Consumer Electronics*, 2008, pp. 1–4.

[6] F. Wamser, P. Casas, M. Seufert, C. Moldovan, P. Tran-Gia, and T. Hossfeld, "Modeling the youtube stack: From packets to quality of experience", *Computer Networks*, vol. 109, 2016.

[7] Telestream, "Wirecast user manual", 2017. [Online]. Available: www.telestream.net/application-content/wirecast/help/7-3/win/Wirecast-User-Guide-Windows.pdf

[8] Y. Corporation, "Creator studio", 2017. [Online]. Available: https://support.google.com/youtube/answer/6060318

[9] T. L. Foundation, "Netem", 2017. [Online]. Available: https://wiki.linuxfoundation.org/networking/netem

[10] S. Salsano, F. Ludovici, A. Ordine, and D. Giannuzzi, "Definition of a general and intuitive loss model for packet networks and its implementation in the netem module in the linux kernel", *University of Rome*, vol. 3, 2012.

[11] U. M. Repository, "Ubuntu 16.10", 2017. [Online]. Available: http://manpages.ubuntu.com/manpages/xenial/en/man8/tc-netem.8.html

[12] S. Hemminger *et al.*, "Network emulation with netem", in: *Linux conf au*, 2005, pp. 18–23.

[13] Tcpdump-workers, "Tcpdump", 2017. [Online]. Available: http://www.tcpdump.org/

[14] Ascom, "Tems pocket specifications", 2017. [Online]. Available: http://www.tems.com/products-for-radio-and-core-networks/radio-network-engineering/portable-testing-for-wireless-networks

[15] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of youtube QoE via crowdsourcing", in: IEEE *Multimedia Int. Symp. on Multimedia Quality of Experience*, 2011, pp. 494–499.

[16] P. Casas, A. Sackl, S. Egger, and R. Schatz, "Youtube & Facebook quality of experience in mobile broadband networks", in: IEEE *Globecom Workshops* (GC Wkshps), 2012, pp. 1269–1274.

[17] R. K. Mok, E. W. Chan, and R. K. Chang, "Measuring the Quality of Experience of HTTP Video Streaming", in: IFIP/IEEE *International Symposium on Integrated Network Management* (IM), 2011, pp. 485–492.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6489

Evaluating the Impact of Energy-Aware Routing on Software-Defined Networking Performance

Adriana Fernández-Fernández, Cristina Cervelló-Pastor, Leonardo Ochoa-Aday
Department of Network Engineering,
Universitat Politècnica de Catalunya
Esteve Terradas, 7, 08860, Castelldefels, Spain.
{adriana.fernandez, cristina, leonardo.ochoa}@entel.upc.edu

***Abstract*—Increasing power consumption and CO_2 emissions generated by large data networks have become a major concern over the last decade. For this problem, the emerging paradigm of Software-Defined Networks (SDN) can be seen as an attractive solution. In these networks an energy-aware routing model could be easily implemented leveraging the control and data plane separation. This paper investigates the impact of energy-aware routing on SDN performance. To that end, we propose a novel energy-aware mechanism that reduces the number of active links in SDN with multiple controllers, considering in-band control traffic, i.e. links are shared between data and control plane traffic. The proposed strategy exploits knowledge of the network topology combined with traffic engineering techniques to reduce the overall power consumption. Therefore, two heuristic algorithms are designed: a static network configuration and a dynamic energy-aware routing. Significant values of switched-off links are reached in the simulations using real topologies and demands data. Moreover, obtained results confirm that crucial network parameters such as control traffic delay, data path latency, link utilization and TCAM occupation are affected by the performance-agnostic energy-aware model.**

***Keywords*—Software-Defined Networking, energy-aware routing, in-band control traffic, heuristic algorithms.**

I. INTRODUCTION

Recently, the growing energy consumption of Information and Communication Technologies (ICT) has attracted the attention of the networking researchers. According to [1] in 2012, close to 4.7% of the world's electrical energy was consumed by ICT, releasing into the atmosphere roughly 1.7% of the total CO_2 emissions. Moreover, recent studies state that energy demand of ICT sector is growing faster than the overall one and the power consumption of the global Internet could rise to more than 10% of the world's electricity consumption by 2025 [2]. This implies that the reduction of power consumption in Internet Service Provider (ISP) backbone networks is crucial to accomplish significant energy savings in this sector. At the same time, increasing the energy efficiency in data networks can also substantially reduce the environmental impacts of other sectors.

Given that energy consumption of network equipment is only slightly influenced by their traffic load, an effective strategy to minimize the consumption of data networks is to reduce the number of active elements [3]. This feature can be implemented by putting into a low-power state (sleep mode) elements such as line cards or port interfaces that are not in use. Although turning off entire interconnection devices enables greater energy savings, in this work we do not consider this possibility because of resiliency concerns in case of network events. However, due to typical over-provisioning considered in the design and operation of backbone networks, considerable energy savings could be reached changing the status of network interfaces to sleep mode whenever a link is not transferring data.

Within this context, a promising solution for this problem is the use of Software-Defined Networking (SDN) [4]. The basic idea of SDN is to decouple control and data planes to make network environments more manageable. The logically centralized control plane in SDN has a global knowledge of network state information. Furthermore, it can manage network tasks and perform device programming without the need of additional software or hardware in each one of the switching elements. Meanwhile, network devices only forward traffic according to the rules set by the controller. This feature can be leveraged to perform an energy-aware routing that determines, in a coordinated and centralized way, the switch interfaces that should be put to sleep mode. Therefore, an energy-aware solution could be easily implemented in the control plane.

Despite consistent efforts to improve the network power efficiency, these techniques lead to performance degradations when QoS requirements are neglected. Inspired by this reality, this paper introduces a new energy-aware strategy and evaluate its impact on crucial performance metrics. Instead of restricting the path selection and po-

tential improvements in terms of energy efficiency to meet some specific metric bound, this work aims to quantify performance concerns of a fundamental research topic in recent communication networks.

Throughout this work we consider a SDN architecture with multiple controllers and, similar to our previous works, [5] and [6], in-band control traffic. This means that control messages are exchanged using the same links that data traffic without the need of additional edges. In this way, the energy-aware routing performance can be analyzed when, for physical and cost-related restrictions, implementing a dedicated control network is not feasible. Furthermore, this is a more realistic scenario for large backbone networks, where additional links dedicated to transfer the control messages between controllers and forwarding devices are impractical and cost-inefficient.

Specifically, the major contributions of this paper are as follows:

- We develop a novel energy-aware mechanism that reduces the number of active links in SDN with multiple controllers, considering that links are shared between data and control plane traffic.
- Two solution modules were conceived, exploiting knowledge of the network topology and traffic engineering techniques to reduce the overall power consumption.
- Using real topologies and traffic demands, we provide a performance comparison analysis of our proposal with another routing approaches.

The rest of this paper is structured as follows. In Section II we further discuss previous studies about different strategies to tackle the problem of power consumption. In Section III we explain the main characteristics of our energy-aware approach together with the description of its two comprised modules. The simulations strategies and the obtained results are presented and analysed in Section IV. Finally, in Section V we conclude our work and outline future research guidelines.

II. RELATED WORKS

Energy-aware techniques that reduce the number of active elements in the network can be divided into traffic-based and topology-based solutions, according to the elements considered in the model. In this section we analyze in more details works that deal with each one of these two approaches.

A. Traffic-Based Solutions

Under some assumption of expected traffic behaviour, traffic-based solutions are routing mechanisms that aggregate traffic over a subset of links and devices in over-provisioned networks, in order to switch off the unused network components.

For instance, Zhang et al. in [7] propose GreenTE, an intra-domain, centralized traffic engineering mechanism that finds a set of links that can be turned off under a given traffic load or matrix. The approach is based on a Mixed-Integer Linear Programming (MILP) formulation where the traffic demands are routed through a set of previously computed k-shortest paths. Performance requirements such as maximum link utilization (MLU) and network delay are considered as constraints in the problem. However, the implementation of such coordinated strategy is a difficult task given the distributed nature of network control in traditional networks.

More recently, in [8] the authors introduced a state-of-the-art study of energy efficiency strategies in SDN. This paper addresses the importance of implementing green routing methods in SDN, taking advantage of the flexibility given by dynamic configuration and centralized network view capabilities. A summary of some existing energy-aware techniques in SDN with their key properties (benefits and drawbacks) is presented.

The authors of [9] addressed the problem of saving energy in partially deployed SDN. They formulated an optimization problem for finding minimum power network subsets in these hybrid networks. Giroire et al. [10] proposed an energy-aware routing approach that takes into account the limited rule space of TCAM in SDN devices. An ILP model is presented as well as an efficient heuristic. The authors of [11] provided two greedy algorithms for minimizing the power of integrated chassis and line-cards used. To achieve this they considered an expanded network topology according to the connections between forwarding devices. Nevertheless, in all these works, dedicated links between the controller and SDN nodes were considered.

In [12] the authors proposed a model for controller-switch associations that aims to maximize the energy efficiency of the network. Although the routing of control traffic is considered in this work, they assume that controllers act as well as forwarding devices, i.e. data plane communications are routed through network controllers. Therefore, only links that belong to control paths are activated and data traffic demands are routed using these links until a MLU bound is reached. We argue that data plane traffic should not pass through network controllers, since this will represent an additional load in these devices.

The work in [5] addressed the problem of minimizing the number of required links in large-scale SDN with in-band control traffic. To accomplish this, an ILP model and a heuristic algorithm are presented, integrating the routing requirements for data and control traffic. This model also determines an optimal distribution of switches between controllers in terms of energy efficiency and load balancing. In [6] a distributed routing algorithm that optimizes the power consumption in large-scale SDN with multiple domains is proposed. The solution, called DEAR (Distributed Energy-Aware Routing), tackles the problem of minimizing the number of links that can be used to satisfy a given traffic matrix. Despite being efficient models, the complexity of considering the entire topology for the selection of the most suitable routes can be very expensive in networks with major path redundancy. By the contrary, in this work, after pruning the network topology, the number of paths and the consequent computation complexity are significantly reduced.

B. Topology-Based Solutions

The lack of awareness of traffic conditions in typical operative networks has led to several research works that, in order to reduce the number of active links, are oriented to control the network topology. Basically, these approaches modify the existing topology considering different requirements such as the resulting connectivity.

In [13] the authors present an OSPF-based routing mechanism that considers the topological information exchanged among routers. The proposed EAR algorithm is based on the definition of the "exportation" mechanism where a Shortest Path Tree (SPT) is shared between neighbour nodes. The routers with the highest node-degree, called "exporters", calculate the SPTs that are used to route the traffic and force the use of these paths to all their neighbors, so that the overall set of active links can be reduced. The exportation mechanism is enhanced in [14], where the concept of "move" was introduced turning the energy saving routing problem into a formulation of the well-known Maximum Clique Problem in an undirected weighted graph.

Authors in [15] propose a routing algorithm called Energy Saving based on Algebraic CONnectivity (ESACON), using the algebraic connectivity as a metric to control the resulting network topology. Based on this metric, ESACON is able to identify and switch off the network links that less affect the network connectivity, keeping this value over a given threshold.

Similarly, the topology-based solution reported in [16] also takes into account the algebraic connectivity as a requirement to preserve the overall network connectivity. This work also considers the edge betweenness as a metric to measure the links role in the network, placing the links least frequently used as the first candidates to be pruned. However, this approach is conceived to be implemented in a distribute way into each IP router.

The work in [17] also aims to improve the energy efficiency reducing the number of active links. For this purpose, the authors propose four different versions of the Energy Saving based on Occurrence of Links (ESOL) algorithm that show the tradeoff between complexity and efficiency in powering off a great number of links. The parameters used in this approach to select the network interfaces to be switched off are the occurrences of nodes and links in network paths.

The analysis of including QoS requirements in an energy-aware topology-based solution is discussed by the authors of [18]. Their approach, called Energy Saving IP Routing (ESIR), is also based on the concepts of SPT exportation and move but constrained to a maximum load boundary on network links.

All the previously described works mainly tend to minimize the number of active network elements in the current topology restricting the path selection to meet some specific metric bound or connectivity rate and fail to extensively examine the impact of energy-aware routing on SDN performance. Moreover, their lack of awareness about the requirements of incoming connection requests can lead to performance degradations, which is highly undesired.

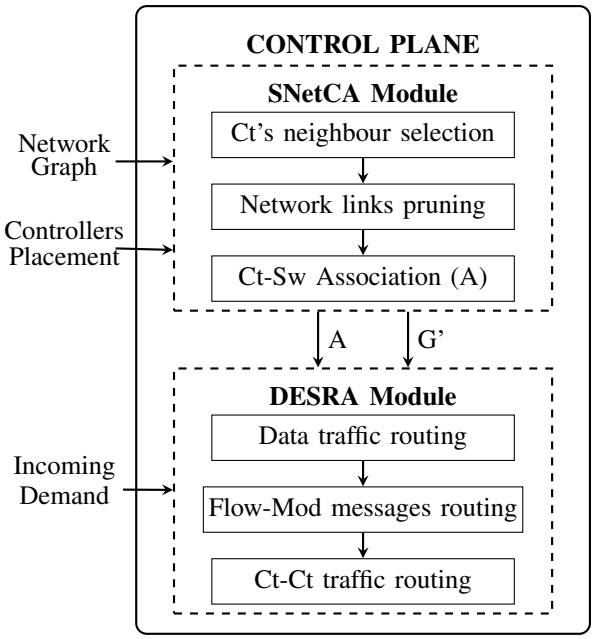


Fig. 1. Illustrative diagram of the proposed approach.

Different from the aforementioned works, the aim of this paper is to provide a low-complexity energy-aware strategy which will be used to evaluate its impact on crucial performance metrics, considering a SDN architecture with multiple controllers and in-band control traffic.

III. ENERGY-AWARE APPROACH

In this work, we present a hybrid solution for the energy efficiency problem in SDN comprising the main advantages of the two aforementioned approaches. More precisely, we exploit specific network topological properties combined with the use of traffic engineering to reduce the overall power consumption. An illustrative diagram of this strategy is shown in Fig. 1. The first component, denoted as Static Network Configuration Algorithm (SNetCA), is a topology-based solution intended to be statically activated as a planned operation. On the other hand, the traffic-based module, denoted as Dynamic Energy Saving Routing Algorithm (DESRA), is activated by the arrival of each incoming traffic demand. Therefore, an accurate prediction of incoming traffic is not needed.

In essence, this approach finds the routes between network elements that minimize the number of active links used, considering that links are shared between data and control plane traffic (i.e. in-band mode). Therefore, control paths between controllers and switches (in both senses) and between controllers are also established.

Additionally, given the controllers placement in the network topology, our model determines the ideal distribution of switches between controllers in terms of energy efficiency, considering as well a balanced load between controllers. In our energy-aware approach, the routing of additional traffic load through the controllers is avoided,

i.e., admissible control paths do not pass through any other controller that is not the source or target of the traffic and data plane communications cannot be routed through these devices.

The two main parts enclosed within the proposed energy-aware approach are described in more details in the following subsections.

A. Static Network Configuration Algorithm

In the proposed scheme the network topology can be modeled as a directed graph $G = (V, E, C)$, where V, E and C denote the set of nodes, links and controllers respectively, being $C \subset V$. We define the set of interconnection devices as $S = \{n \mid n \in V \wedge n \notin C\}$. We use X to denote the set of active links $X \subseteq E$ and U to store the utilization of network links.

By considering the typical link redundancy of backbone networks, we design a Static Network Configuration Algorithm, denoted as SNetCA, which aims to prune as many links as possible in order to stress the importance of energy saving. Additionally, the most favorable switch-controller associations in terms of energy efficiency and load balance, are determined in this stage.

The algorithm, described in the Algorithm 1 pseudocode, is composed of three steps:

1) selecting one of the controller's neighbours, as the node that will remain connected to it in the outcome topology;
2) identifying the links that do not disconnect the graph to be put into sleep mode;
3) associating each node with one controller and computing the control path between them.

The input of the algorithm is the network topology with controllers placement and its outputs are a pruned network with a reduced number of links, denoted as G', an array keeping the controller-switch associations, denoted as A and the control paths from each node to its controller, denoted as P_{sc}.

In the first step, the algorithm iterates over the set of network controllers in order to evaluate each one of its neighbours. The selection of one neighbour node for each controller is based on the betweenness centrality (B_n), which measures the intermediary role of a node in the network. In the proposed approach, we use a simplified version of this metric considering only the shortest paths from a controller to every switch.

In particular, after computing the shortest paths from one controller as single source, the algorithm determines whether a neighbour node belongs to each path and increases the B_n associated with that node (lines 6-16). For each controller a list of neighbour devices, sorted in decreasing order of B_n, is stored in L. This list is used to identify the available neighbour with the highest betweenness centrality. This node is associated with the considered controller in the current iteration and stored in A, as long as it has not been already attached to another controller.

For the remaining nodes in L, the links between them and the controller are removed from the resulting network graph. This means that they are put into sleep mode in the original graph. Notice that when a controller's neighbour is another controller, the link between them is not considered as a candidate to be pruned (lines 7-9).

In the next step, the algorithm iterates over the set of directional links in the pruned network that do not have any controller as its extreme nodes. At each iteration the

Algorithm 1 SNetCA

Require: $G = (V, E, C)$ network graph with controller placements
Ensure: $G' = (V, E', C)$ network graph with reduced number of links, A controller-switch associations, P_{sc} switch-controller control paths

1: $N_c \leftarrow$ Set of neighbours of controller $c \in C$
2: $G' \leftarrow G$
3: **for** $c \in C$ **do**
4: $B \leftarrow$ NULL ▷ Array of betweenness values
5: $SP_c \leftarrow$ Set of shortest paths from controller $c \in C$
6: **for** $n \in N_c$ **do**
7: **if** $n \in C$ **then**
8: continue
9: **end if**
10: $B_n = 0$
11: **for** $p \in SP_c$ **do**
12: **if** $n \in p$ **then**
13: $B_n = B_n + 1$
14: **end if**
15: **end for**
16: **end for**
17: $L \leftarrow B\_Sorted$
18: **for** $s \in L$ **do**
19: **if** s and c not already in A **then**
20: $A = A \cup (s, c)$
21: **end if**
22: Remove links (s, c) and (c, s) from G'
23: **end for**
24: **end for**
25: **for** $i, j \in E'$ **do**
26: **if** $i \in C$ or $j \in C$ **then**
27: continue
28: **end if**
29: $G'' \leftarrow G'$
30: Remove controllers $c \in C$ from G''
31: Remove link i, j from G''
32: **if** G'' remains strongly connected **then**
33: Remove link i, j from G'
34: **end if**
35: **end for**
36: **for** $s, c \in A$ **do**
37: PATHSELECTOR(s, c)
38: Update P_{sc}, X, U
39: **end for**
40: **for** the rest of $s \in S$ **do**
41: PATHSELECTOR(s, C)
42: Update P_{sc}, A, X, U
43: **end for**

Algorithm 2 PATHSELECTOR(a, b)

```
1: L ← ∞
2: SeP ← None
3: for p ∈ Get_All_Adimissible_Paths(G', a, b) do
4:     if b = C then
5:         if p is to an already loaded controller then
6:             continue
7:         end if
8:     end if
9:     off ← number of links in p that are not in X
10:    if off < L and p has sufficient bandwidth then
11:        L ← off
12:        SeP ← p
13:    end if
14: end for
```

Algorithm 3 DESRA

Require: G', A, d incoming traffic request
Ensure: P_{ss}, P_{cs}, P_{cc} data and control paths, X active links, U links utilization

```
1: Ct1 ← A[s_d]
2: p = PATHSELECTOR(s_d, t_d)
3: Update P_ss, X, U
4: for n ∈ p do
5:     Ct ← A[n]
6:     PATHSELECTOR(Ct, n)
7:     Update P_cs, X, U
8:     if Ct1 ≠ Ct then
9:         PATHSELECTOR(Ct1, Ct)
10:        Update P_cc, X, U
11:    end if
12: end for
```

algorithm attempts to increase the number of switched-off edges.

A new link is removed only when the resulting graph remains being strongly connected, i.e. at least one path exists between every pair of nodes in the network. To accomplish this, a temporal graph without any controller, denoted as G'', is created. This graph is used to check the required connectivity between all the forwarding devices. After validating that the possibility to reach any node in the network is not affected, the considered link is removed from the resulting graph.

The last step of the algorithm is intended to determine a control path from each forwarding device to one controller. To achieve this goal, the algorithm starts evaluating the pairs of controller-switch associations already stored in A (line 36). For each pair, an admissible control path minimizing the number of active links is selected using the method PATHSELECTOR described in Algorithm 2, which will be further explained below. As stated previously, admissible control paths do not pass through any other controller that is not the source or target of the traffic. The remaining forwarding devices are then considered. Notice that in this case the algorithm takes into account the control paths to all controllers in the network. Precisely, the path computed by the PATHSELECTOR in this step defines the controller for the rest of forwarding devices.

Using this initial control plane configuration, switches send to the controller packet_in requests when a new traffic flow arrives, as well as statistics and failure notifications. Consequently, there is an initial set of active links in the network before the ingress of traffic flows as well as some link utilization.

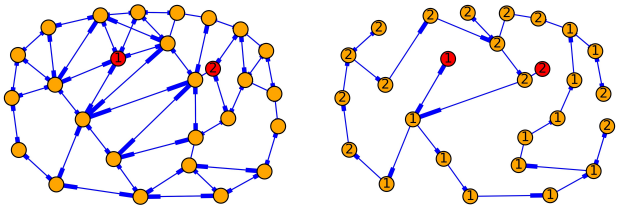
The PATHSELECTOR method, described in Algorithm 2, performs the energy-aware path selection. In essence, this function is used to select the best admissible route between a pair of nodes, in terms of minimizing the number of active links in the network. The key idea of this function is to perform a low-complexity greedy evaluation between all the admissible paths to select the most suitable route in terms of energy-efficiency, while guaranteeing a balanced load of switches between controllers and capacity constraint of links. Since this method works over the pruned network with a reduced number of links, (i.e. G'), the set of admissible paths considered is significantly smaller than in the original topology and the solution can be found with fewer iterations. When this function is called for determining the path between each forwarding device and one controller (i.e. using the set of controllers as the traffic destination) the controller load is considered (line 4 to 8). In addition, the path only can be selected if it has sufficient link capacity to route the required traffic volume.

B. Dynamic Energy Saving Routing Algorithm

When a new traffic demand arrives, a routing request is sent from the incoming node to its associated controller using the path between both devices previously computed during the static network configuration phase. Based on its global knowledge of the network topology, this controller calculates the required data path minimizing the number of links that need to be activated for this connection request and creates the flow forwarding rules. Given the multidomain scenario considered, the nodes traversed by the data traffic may be associated with different controllers.

The proposed dynamic energy-aware routing is shown in Algorithm 3. For an incoming demand d from source s_d to destination t_d, the algorithm starts storing in Ct_1 the controller associated with the source node. This controller is the main responsible of managing this traffic request. Using the PATHSELECTOR method, the most favorable admissible data path in terms of energy consumption is computed. This is done considering that admissible paths do not pass through any controller in the network. Then, a loop is used to consider the required control plane communications for each node along this path.

After determining the controller associated with each node in the data path, a control path is computed between them. These paths are used to set the flow forwarding rules in each switch using the flow_mod messages. When a node is not associated with Ct_1 an additional control message is sent from this controller to the other, in order to inform the second controller of the flow forwarding rule that need to be installed in one of its managed nodes.



(a) Original network graph. (b) Pruned network graph.

Fig. 2. SNetCA performance on the Norway topology.

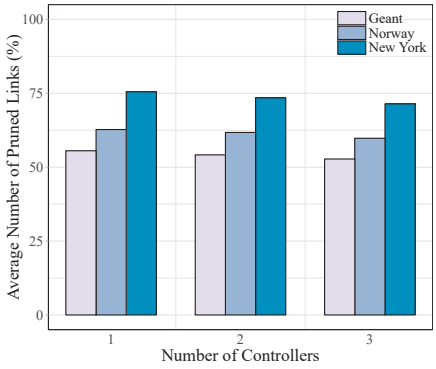


Fig. 3. Average number of pruned links in the three topologies varying the number of controllers.

IV. SIMULATIONS AND RESULTS

In this section we describe the evaluation of our energy-aware approach and analyze the results obtained. The proposed control framework described in Section III was implemented using the programming language Python to develop the heuristic algorithms. All computations were carried out on a computer equipped with 3.30 GHz Intel Core i7 and 16 GB RAM. We conducted our simulations using real network topologies and traffic demands collected from SNDlib [19], considering each router in the network as a SDN node or as a possible controller placement.

Specifically, we use three of the most link-redundant network topologies in SNDlib in order to asses the effectiveness of the proposed scheme. The mentioned topologies are: Geant ($|V| = 22$; $|E| = 72$), New York ($|V| = 16$; $|E| = 98$) and Norway ($|V| = 27$; $|E| = 102$). For the control traffic we assume an average rate of 1.7 Mbps [20]. To analyse the performance of our energy-aware approach we present the following evaluations for different amount of controllers in the network.

A. SNetCA performance

In order to evaluate the effectiveness of the proposed topology-based module, Fig. 2 shows an example of the performance of SNetCA on the Norway topology, considering two network controllers placed at nodes denoted as 1 and 2 and emphasized with a different color in the figure. The distribution of switches between controllers is depicted through the use of labels in each node, indicating the controller number to which the node is associated.

A comparison between the original network and the resulting graph illustrated in Fig. 2(a) and Fig. 2(b) respectively, shows a difference of 67 edges, which represents more than 65% of total network links. These links are pruned by our algorithm guaranteeing that the resulting graph remains being strongly connected and avoiding additional traffic load through network controllers.

Additionally, as a result of applying SNetCA on the Norway topology, switches are distributed between controllers minimizing the number of required active links and ensuring a balanced controllers load. For instance, 12 switches are associated with controller 1 while the remaining 13 are managed by controller 2.

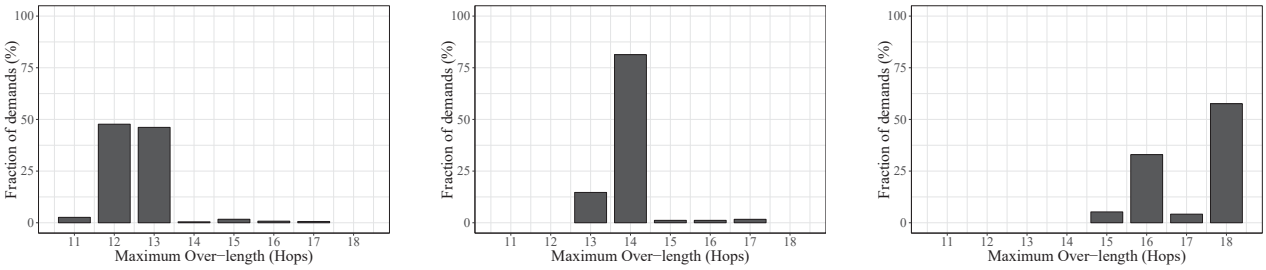
To provide a more general perspective, Fig. 3 shows, for the three considered topologies, the average number of links pruned by SNetCA. In this analysis we consider all the admissible placements of 1 to 3 controllers. Notice that a controller placement is admissible when the assumptions established in this proposal to avoid the routing of additional traffic load through network controllers can be kept (i.e., the network graph without any controller is strongly connected). As it is shown, a high number of links is pruned in all the topologies considered, which contributes directly with the energy efficiency achieved by this proposal. In general, the more redundant the network, the higher number of links can be put to sleep mode applying this strategy.

B. Impact of DESRA on Network Performance

It is to be emphasized that in our energy-aware approach quality of service (QoS) constraints and performance metric boundaries are not taken into account. This is not a limitation but a choice; since we intend to measure the impact of our proposal on the network performance metrics as a trade-off with the energy saving improvements. In fact, we are presenting an effective and easy to implement green routing mechanism that emphasizes the importance of energy efficiency in the operation of current data networks.

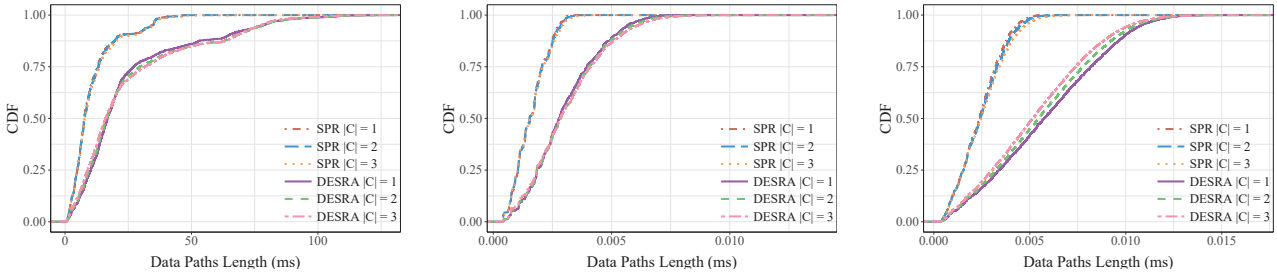
In order to assess the impact of our energy-aware approach on the network performance, we adapt two well-known state-of-the-art routing algorithms: Shortest Path Routing (SPR) and Load Balancing (LB) for their use in the considered in-band SDN environment. Being the rule space a significant issue of concern in SDN, an algorithm balancing the number of rules installed in each forwarding device, denoted as TCAM Occupation Balancing (TOB), is also included in this analysis. In essence, these algorithms are greedy heuristics that prioritize some performance metric such as: traffic latency, link utilization or TCAM occupation used in our evaluation as baselines for comparison purposes. All of them follow the assumptions established in this proposal to avoid the routing of additional traffic load through network controllers. SNetCA is still used to determine the distribution of switches between controllers.

Due to space limitation, for the different performance metrics, we may focus our attention on some specific network, but similar results have been obtained for all the considered topologies.



(a) $|C| = 1$. (b) $|C| = 2$. (c) $|C| = 3$.

Fig. 4. Distribution of maximum control traffic over-length in the Norway topology for different amount of controllers.



(a) Geant Topology. (b) New York Topology. (c) Norway Topology.

Fig. 5. CDF of data paths latency for three different topologies varying the amount of controllers.

1) Traffic Latency: In a first set of simulations, we analyze how the data and control paths latency is affected by the routing decisions made.

To evaluate the impact of our algorithm on control path delay, we collect, for each traffic demand, the length of its associated control paths and the corresponding shortest path. The notation Maximum Over-length is used to denote the maximum difference (in number of hops) between the length of the routing solution and the shortest path. Fig. 4 shows this behaviour for the Norway topology considering all possible placements for different amount of controllers. As it is shown, when the number of controllers grows, the control traffic is routed using a larger number of hops for a higher fraction of demands.

To take a closer look at the data plane, we draw in Fig. 5 the cumulative distribution function (CDF) of data paths latency for three different topologies considering all possible locations of one to three controllers. As shown in Fig. 5(a), Fig. 5(b) and Fig. 5(c), the CDFs of data paths latency for different amount of controllers are quite similar. However, we can see that under the energy-aware routing, the control path delay is affected in order to minimize the number of active links. For instance, in Fig. 5(a), only 87% of data paths exhibit delays lower than 50 ms, meanwhile all control paths in the SPR case are under this value. In general, the larger the network (in terms of geographic length), the more increase in latency is incurred.

2) Links Utilization: The selection of routing paths minimizing the energy consumption has a direct influence in the traffic load of all the network links. To better showcase this situation, we use the Geant topology and the LB algorithm. Fig. 6 provides the CDF of link utilization under both algorithms considering all possible locations of one to three controllers in this topology. As expected, the fairness of traffic distribution is altered by the energy-aware routing, since there is a subset of active links that is more overloaded than the others. Nevertheless, even in the more loaded cases the link utilization is under 60% in this topology.

3) TCAM Occupation: Intuitively, an energy-aware routing would affect the allocation of flow rules since traffic flows are redirected to minimize the number of active links. In Fig. 7, we evaluate the impact of our approach on TCAM occupation with respect to the TOB algorithm using the Geant topology and all possible locations of three network controllers. As expected, the number of installed rules is raised by the energy-aware routing in almost all the network devices (18 out of 22 nodes), being in some cases more than twice the value obtained by the TOB algorithm. However, the DESRA performance in this topology is still physically acceptable considering that a routing table can support from 750 to few thousands of rules [10].

4) Energy savings: To get a sense of the energy saving values achieved by our approach, Fig. 8 shows the average energy performance of all the considered routing models in the New York topology for the case of one centralized controller in the network. The energy savings were computed as the number of links in sleep mode over the total amount of network links. As expected, in all cases energy saving decreases while the number of demands grows, since new paths need to be established. Furthermore, the proposed routing algorithm outperforms SPR, LB and TOB in terms of energy saving. In general, DESRA achieves significant energy savings but bigger improvements with respect to the other approaches are reached when the traffic grows.

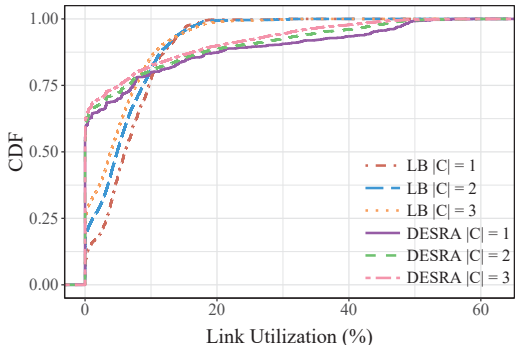


Fig. 6. CDF of link utilization in the Geant topology.

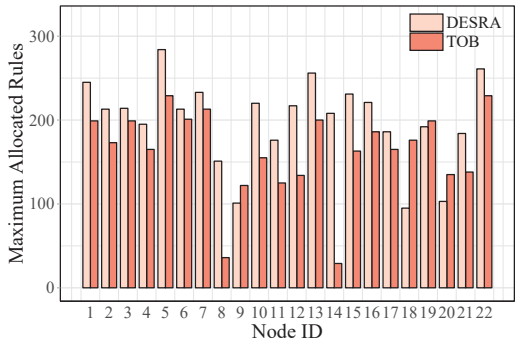


Fig. 7. Average TCAM occupation in the Geant topology with $|C| = 3$.

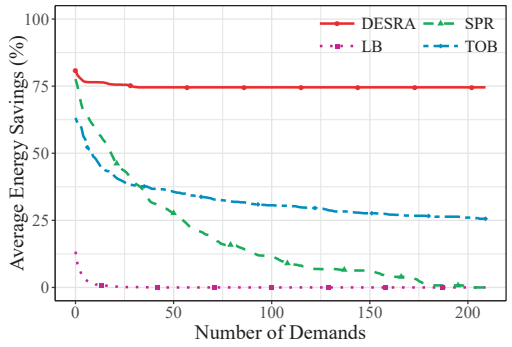


Fig. 8. Average energy savings in the New York topology with $|C| = 1$.

V. CONCLUSION

In this paper, we evaluate the performance impact of applying an energy-aware routing on SDN with in-band control traffic. To achieve such goal, we have proposed a hybrid approach comprising two heuristic algorithms: a static network configuration and a dynamic energy-aware routing. Apart of providing an effective power-aware scheme able to achieve notable improvements in terms of energy saving, the most significant added value of the proposal is the quantitative analysis presented of such significant networking concern. Extensive simulations using real topologies and traffic matrices validate that crucial network parameters such as control traffic delay, data path latency, link utilization and TCAM occupation are affected by the performance-agnostic energy-aware model. These findings confirm that energy-aware routing schemes should be designed considering specific traffic requirements and performance metric bounds. As future work, we want to provide an analysis on the impact of reducing the number of active network elements on SDN reliability, considering in-band control traffic.

ACKNOWLEDGMENT

This work has been supported by the Ministerio de Economía y Competitividad of the Spanish Government under project TEC2016-76795-C6-1-R and AEI/FEDER, UE and through a predoctoral FPI scholarship.

REFERENCES

[1] E. Gelenbe and Y. Caseau, "The Impact of Information Technology on Energy Consumption and Carbon Emissions," *ACM Ubiquity*, vol. 2015, no. June, pp. 1–15, 2015.

[2] R. S. Tucker, "Energy consumption in telecommunications," in *Proc. Optical Interconnects Conference*, May 2012, pp. 1–2.

[3] M. Gupta and S. Singh, "Greening of the Internet," in *Proc. ACM SIGCOMM'03*, 2003, pp. 19–26.

[4] D. Kreutz, F. M. Ramos, P. Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, pp. 14–76, 2015.

[5] A. Fernandez-Fernandez, C. Cervello-Pastor, and L. Ochoa-Aday, "Achieving Energy Efficiency: An Energy-Aware Approach in SDN," in *IEEE GLOBECOM'16*, Dec. 2016, pp. 1–7.

[6] ——, "Energy-Aware Routing in Multiple Domains Software-Defined Networks," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 5, no. 3, 2016.

[7] M. Zhang, C. Yi, B. Liu, and B. Zhang, "GreenTE: Power-Aware Traffic Engineering," in *Proc. IEEE ICNP'10*, 2010, pp. 21–30.

[8] B. G. Assefa and O. Ozkasap, "State-of-the-art Energy Efficiency Approaches in Software Defined Networking," in *Proc. SoftNetworking'15*, Apr. 2015.

[9] H. Wang, Y. Li, D. Jin, P. Hui, and J. Wu, "Saving Energy in Partially Deployed Software Defined Networks," *IEEE Trans. Comput.*, vol. 65, no. 5, pp. 1578–1592, 2016.

[10] F. Giroire, J. Moulierac, and T. K. Phan, "Optimizing Rule Placement in Software-Defined Networks for Energy-Aware Routing," in *Proc. IEEE GLOBECOM'14*, 2014, pp. 2523–2529.

[11] R. Wang, Z. Jiang, S. Gao, W. Yang, Y. Xia, and M. Zhu, "Energy-Aware Routing Algorithms in Software-Defined Networks," in *Proc. IEEE WoWMoM'14*, June 2014, pp. 1–6.

[12] A. Ruiz-Rivera, K. W. Chin, and S. Soh, "GreCo: An Energy Aware Controller Association Algorithm for Software Defined Networks," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 541–544, 2015.

[13] A. Cianfrani, V. Eramo, M. Listanti, M. Marazza, and E. Vittorini, "An Energy Saving Routing Algorithm for a Green OSPF Protocol," in *Proc. IEEE INFOCOM'10*, Mar. 2010, pp. 1–5.

[14] A. Cianfrani, V. Eramo, M. Listanti, and M. Polverini, "An OSPF enhancement for energy saving in IP networks," in *Proc. IEEE INFOCOM WORKSHOPS'11*, Apr. 2011, pp. 325–330.

[15] F. Cuomo, A. Abbagnale, A. Cianfrani, and M. Polverini, "Keeping the connectivity and saving the energy in the internet," in *Proc. IEEE INFOCOM WORKSHOPS'11*, Apr. 2011, pp. 319–324.

[16] F. Cuomo, A. Cianfrani, M. Polverini, and D. Mangione, "Network Pruning for Energy Saving in the Internet," *Comput. Netw.*, vol. 56, no. 10, pp. 2355–2367, July 2012.

[17] F. Cuomo, A. Abbagnale, and S. Papagna, "ESOL: Energy saving in the Internet based on Occurrence of Links in routing paths," in *Proc. IEEE WoWMoM'11*, June 2011, pp. 1–6.

[18] A. Cianfrani, V. Eramo, M. Listanti, M. Polverini, and A. V. Vasilakos, "An OSPF-Integrated Routing Strategy for QoS-Aware Energy Saving in IP Backbone Networks," *IEEE Trans. Netw. Service Manag.*, vol. 9, no. 3, pp. 254–267, Sept. 2012.

[19] S. Orlowski, M. Pióro, A. Tomaszewski, and R. Wessäly, "SNDlib 1.0-Survivable Network Design Library," *Networks*, vol. 55, no. 3, pp. 276–286, 2010.

[20] J. Li, J.-H. Yoo, and J. W.-K. Hong, "Dynamic control plane management for software-defined networks," *International Journal of Network Management*, vol. 26, no. 2, pp. 111–130, 2016.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6498

Aprovechando el Poder de las Feromonas para Mejorar la Eficiencia Energética en Redes Definidas por Software

Raúl Sánchez Romero, Jaime Galán-Jiménez
Departamento de Ingeniería de Sistemas Informáticos y Telemáticos
Universidad de Extremadura
Cáceres, España
Email: rsanchezzq@alumnos.unex.es, jaime@unex.es

***Resumen*—En los últimos años, la conciencia por la eficiencia energética se ha instalado de lleno en el seno de la sociedad. Aunque la comunidad investigadora ha realizado grandes esfuerzos en proponer soluciones para reducir el consumo de energía de las redes de comunicaciones, la mayoría de estas propuestas están pensadas para redes IP, para las cuales es necesaria la coordinación entre los distintos elementos que las componen. La aparición de las redes SDN (*Software-Defined Networking*) y el desacople del plano de datos del plano de control abre nuevos caminos para proponer algoritmos energéticamente eficientes a ejecutar por parte de un elemento lógicamente centralizado, el controlador SDN. En este artículo se propone un algoritmo basado en el comportamiento de las colonias de hormigas que permite la elección óptima, en términos de eficiencia energética, de los modos de operación de los enlaces de la red. Los resultados obtenidos tras su ejecución sobre topologías de red reales indican que es posible conseguir un ahorro de energía significativo, en torno al 30% para el caso en que la función de energía de los enlaces de la red sea de tipo lineal.**

***Palabras Clave*—Eficiencia energética, SDN, ACO, nivel de energía, función de energía.**

I. INTRODUCCIÓN

La conciencia por la mejora de la eficiencia energética ha supuesto un cambio genérico en la mentalidad de la sociedad a lo largo de los últimos años. Varios estudios indican que el consumo de energía de las TIC (Tecnologías de la Información y las Comunicaciones) supone de entre el 2% al 10% del consumo de energía a nivel mundial [1], [2]. Este consumo energético está directamente relacionado con la generación en torno al 2% del total de emisiones de CO_2 a la atmósfera [1]. Además, se prevé que la demanda de energía del sector TIC crezca de forma más rápida que la demanda de energía global [3].

Por su parte, las redes de comunicaciones suponen un 37% del total de la energía requerida por la industria TIC a nivel mundial [4], [5], siendo el equipamiento de red el principal responsable del consumo de Internet [6]. Para solventar el problema del consumo de energía en las redes de comunicaciones, la comunidad investigadora ha realizado grandes esfuerzos en proponer soluciones que permitan conseguir un ahorro de energía significativo, especialmente en periodos de tiempo en los que la actividad disminuye [7]. Partiendo de la idea básica de utilizar el mínimo conjunto de elementos de red necesarios para satisfacer la demanda de tráfico dada, se han propuesto esquemas que utilizan las técnicas *Sleeping* y *Rate Adaptation* para este propósito [6]. El tráfico de la red se redigirá para cada par origen-destino por un determinado conjunto de caminos determinados de tal forma que el consumo de la red sea mínimo y no influya negativamente sobre el rendimiento de la misma.

Sin embargo, la aplicación de estas soluciones energéticamente eficientes no es sencilla a priori debido a la naturaleza distribuida de las redes IP. La filosofía centralizada del nuevo paradigma de interconexión de redes denominado redes definidas por software (SDN, *Software-Defined Networking*) abre nuevas oportunidades para proponer algoritmos que permitan minimizar el consumo energético global de la red. La separación del plano de datos del plano de control y la visión global del estado de la red por parte del controlador, facilita la interacción entre éste y los nodos de la red (switches SDN) ya sea para realizar consultas específicas (ej. carga de los enlaces) o actuar sobre los elementos de la red (poner a dormir un enlace o activarlo).

En este trabajo se propone un algoritmo que aprovecha la naturaleza del comportamiento animal para seleccionar la configuración óptima de una red SDN en términos de eficiencia energética. En lugar de modificar dinámicamente los caminos por los que debe viajar el tráfico desde un origen hacia un destino determinados para conseguir una reducción en el consumo de energía de la red, la solución propuesta aprovecha los fundamentos

del algoritmo de optimización por colonia de hormigas (ACO, *Ant Colony Optimization*), donde lo que se persigue en cada momento es seleccionar el nivel de energía más apropiado para cada uno de los enlaces, de modo que se satisfaga la matriz de tráfico dada.

Se consideran diferentes modelos de consumo energético para los enlaces de la red (lineal, logarítmica) con el objetivo de extraer bajo qué condiciones es más apropiado utilizar uno u otro tipo. Se han realizado un conjunto de pruebas sobre topologías de red reales de distinto tamaño, cuyos resultados indican que es posible conseguir un significativo ahorro de energía en un tiempo razonable. De este modo, se demuestra que es factible instalar el algoritmo propuesto en un controlador SDN, ya que el tiempo requerido para realizar los cálculos e instalar las reglas correspondientes al encaminamiento resultante de la optimización no es considerablemente elevado.

El resto del artículo se describe como sigue. La Sección II describe los trabajos relacionados con el presentado en este artículo. El problema que se pretende resolver se describe formalmente en la Sección III, mientras que algoritmo basado en ACO para la minimización del consumo de energía en redes SDN se explica en la Sección IV. La Sección V analiza los resultados obtenidos tras las pruebas realizadas para, finalmente, indicar un conjunto de conclusiones en la Sección VI.

II. TRABAJOS RELACIONADOS

Durante la última década, tanto la industria como la comunidad académica han realizado grandes esfuerzos para enfrentarse al problema del consumo de energía en redes de comunicaciones. Comenzando con la idea básica de minimizar el número de recursos de red activos, se puede conseguir una reducción significativa del consumo energético en periodos de baja actividad o al redireccionar el tráfico hacia caminos específicos de modo que el resto de nodos y enlaces de la red puedan pasar a un estado de baja energía (dormido) [4], [8]. Cuando haya variaciones significativas, se deberá tomar la decisión de activar parte de estos elementos con el objetivo de satisfacer la demanda de tráfico y no afectar el rendimiento de la red.

En un paso más, las nuevas características de eficiencia energética proporcionadas por los dispositivos de red actuales abren la oportunidad de adaptar su consumo de energía a la carga de tráfico. La intención detrás de la propuesta del estándar EEE (*Energy-Efficient Ethernet*) fue la de reducir el consumo de energía, como mínimo, al 50% y mantener la compatibilidad requerida con el equipamiento de red existente [9]. EEE se basa en la técnica *Rate Adaptation* [6], que considera distintos perfiles de consumo para el equipamiento de red con extensiones de eficiencia energética. Con esta aproximación, es posible regular el tráfico en función del perfil energético específico para cada dispositivo de la red con el objetivo final de minimizar el consumo de energía global de la red. En [10], los autores de este artículo realizan una comparación entre los dos esquemas mencionados anteriormente (*Sleeping* y *Rate Adaptation*) para proporcionar un valor concreto al número de niveles de energía que es necesario implementar en las tarjetas de línea para conseguir un ahorro de energía significativo en la red. En concreto, se establece un valor de 4 niveles de energía como máximo para conseguir un ahorro significativo ya que, un valor mayor en el número de niveles de energía apenas se traduce en incrementos significativos.

Las redes SDN definen un nuevo paradigma donde se cambia por completo la filosofía tradicionalmente implementada por las redes IP [11]. Existen dos planos claramente diferenciados: el plano de datos y el plano de control. El primero está constituido por los switches SDN, que se limitan a redirigir los paquetes entrantes en base a la información contenida en sus tablas de flujos. Estas tablas de flujos están constituidas por un conjunto de entradas del tipo $\{match, action\}$ que indican la acción a realizar definida en $action$ (ej. reenviar el paquete por un determinado puerto de salida) si se cumple el predicado definido en $match$ (ej. paquete que ha sido enviado por un determinado nodo origen en dirección hacia otro nodo destino). Sin embargo, la lógica de control está implementada en el controlador, elemento clave del plano de control. Se trata de un dispositivo centralizado que dispone en todo momento del conocimiento global de la red y puede actuar según una lógica implementada en un algoritmo instalado en su interior. Por tanto, la aparición de las redes SDN abre nuevos caminos para proponer soluciones centralizadas energéticamente eficientes que se basen en el conocimiento adquirido por el controlador [12], [13].

Este trabajo propone una solución energéticamente eficiente para redes SDN que se basa en el comportamiento colaborativo de las hormigas (ACO). Mediante la definición de conceptos como nivel de energía y función de energía, las hormigas de la colonia representarán configuraciones de red que vendrán identificadas por el conjunto de modos de operación de los enlaces de la red. Estas soluciones potenciales deberán ser evaluadas para comprobar si finalmente son válidas en términos de un correcto encaminamiento al satisfacer la demanda de tráfico de la red. En la siguiente sección se define formalmente el problema de optimización al que se da solución en este trabajo.

III. DEFINICIÓN DEL PROBLEMA

Esta sección define y formaliza el problema de optimización que se pretende resolver con este trabajo. En concreto, el objetivo principal es minimizar el consumo de energía de una infraestructura de red SDN. Para ello, se deben definir previamente un conjunto de conceptos que serán utilizados cuando se realice el proceso de optimización.

A. Formulación del Problema

Consideremos una infraestructura de red SDN $G = (V, E)$, compuesta por un conjunto de nodos $v \in V$ conectados entre sí por un conjunto de enlaces unidireccionales $e_{(i,j)} \in E$. Cada enlace unidireccional $e_{(i,j)} \in E$ que conecta el nodo $i \in V$ con el nodo $j \in V$ tiene un

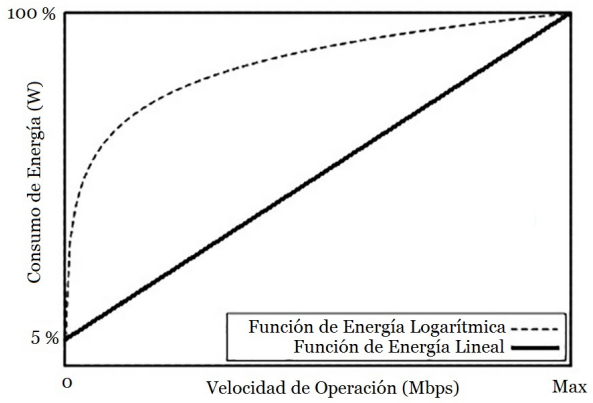


Fig. 1: Caracterización de las funciones de energía.

perfil de consumo de energía $p_{e_{(i,j)}} \in P$ y es capaz de operar utilizando $ne_{(i,j)}$ niveles de energía distintos. Se considera, además, una matriz de demanda de tráfico $M_{(i,j)}, \forall i,j \in V$ con una demanda de tráfico de $d_{(s,d)} \in M_{(i,j)}$ unidades desde el nodo origen $s \in V$ hasta el nodo destino $d \in V$.

Un enlace unidireccional se denota como $e_{(i,j)} \in E$ va dirigido desde el nodo $i \in V$ hacia el nodo $j \in V$ y se encuentra operando según un nivel de energía $ne_{e_{(i,j)}} \in N$. Por tanto, el objetivo es reducir el consumo de energía de la red, Eq. (1), en base a la minimización del consumo de energía de cada uno de los enlaces que la componen:

$$minIPC(C) = \sum_{e_{(i,j)} \in E} f_{e_{(i,j)}}(e_{(i,j)}, ne_{e(i,j)}), \quad \forall i,j \in V, \forall e_{(i,j)} \in E, \forall ne_{e_{(i,j)}} \in N \tag{1}$$

donde $f_{e_{(i,j)}}(e_{(i,j)}, ne_{e_{(i,j)}})$ calcula el consumo de energía en Vatios del enlace unidireccional $e_{(i,j)}$, que se encuentra configurado con un nivel de energía $ne_{e_{(i,j)}}$. El consumo de energía instantáneo IPC de una configuración de red $C \in G$ viene determinado por el sumatorio del consumo de energía de cada enlace, el cual depende del nivel de energía en el que se encuentre configurado.

B. Definición de Nivel de Energía

Dada una función de energía $f_{e_{(i,j)}}$, se considera nivel de energía al par (x, y), donde x corresponde a la velocidad de operación del enlace e y es el resultado de aplicar la función de energía $f_{e_{(i,j)}}$, a x, con $f_{e_{(i,j)}}(x) = y$. De este modo, un nivel de energía relaciona la velocidad de operación del enlace con su consumo energético asociado. El número mínimo de niveles de energía que puede soportar un enlace $e_{(i,j)}$ es 1 o siempre activo ($ne_{e_{(i,j)}} = 1$), y hace referencia a la capacidad nominal del enlace. Se trata por tanto de la situación en la que no existen características de eficiencia energética. El método basado en hardware *Sleeping* añade otro posible nivel de energía al que poder configurar un enlace ($ne_{e_{(i,j)}} = 2$). En este caso, los enlaces pueden encontrarse o bien dormidos o bien operando a su máxima velocidad. Los enlaces que se encuentran dormidos contribuyen al ahorro de energía al no transportar paquetes de datos, aunque permanecen operativos para ser despertados si es necesario. En este nivel de baja energía, los enlaces consumen un pequeño porcentaje, que suele ser entre el 5-10 % del consumo asociado a su nivel de energía máximo (capacidad nominal del enlace). Si por el contrario los enlaces pudiesen adaptar su velocidad de operación a la variación de la carga de tráfico a lo largo del tiempo (estándar EEE), estaríamos ante el caso de la aplicación de la técnica *Rate Adaptation* ($ne_{e_{(i,j)}} > 2$). En concreto, en este trabajo suponemos que los enlaces de la red pueden soportar un número determinado de niveles de energía diferentes.

C. Caracterización de las Funciones de Energía

Si denotamos $C_{e_{(i,j)}}$ a la capacidad nominal del enlace $e_{(i,j)} \in E$, la función de energía $f_{e_{(i,j)}}$ se define Eq. (2)

$$f_{e_{(i,j)}} : C_{e_{(i,j)}} \rightarrow w_{e_{(i,j)}} \quad \forall x \in N \exists y \in w_{e_{(i,j)}} \mid (x,y) \in N f_{e_{(i,j)}} \tag{2}$$

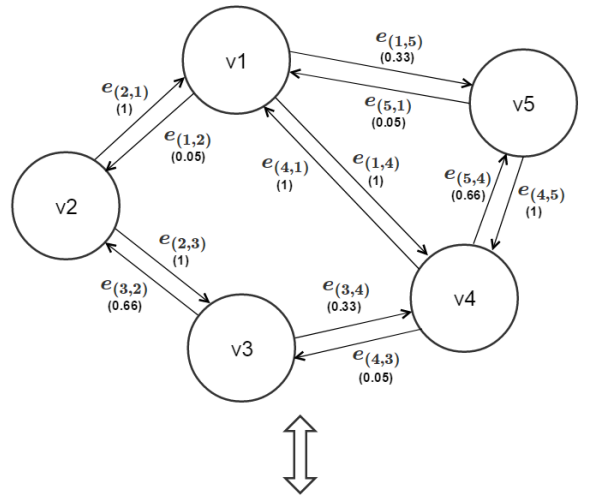
donde $w_{e_{(i,j)}}$ es la imagen de $f_{e_{(i,j)}}$, es decir, el valor del consumo instantáneo en función de la capacidad en la que opera el enlace $e_{(i,j)}$. En la Fig. 1 se muestran las 2 funciones de energía consideradas en este trabajo: i) Lineal o proporcional, $f_{e_{(i,j)}(x)} = x$, representa una función de energía tomada como función base en la que a mayor velocidad de operación, mayor consumo de energía experimentado y ii) Logarítmica, $f_{e_{(i,j)}(x)} = \log(x)$, similar a la implementada por el estándar EEE.

IV. Optimización por Colonia de Hormigas para el Ahorro de Energía

La metaheurística basada en Colonias de Hormigas ACO (*Ant Colony Optimization*) abarca un conjunto de técnicas de optimización inspiradas en el comportamiento colectivo, donde el comportamiento de las hormigas está dirigido hacia la supervivencia de la colonia como un todo que al de un simple componente individual de la colonia [14]. El algoritmo ACO es una técnica probabilística utilizada comúnmente para resolver problemas computacionales que pueden ser reducidos a la búsqueda de caminos óptimos en grafos. Es en esencia un algoritmo constructivo donde en cada iteración, cada hormiga construye una solución al problema recorriendo el propio grafo. En concreto, nuestra propuesta se basa en las variantes ACS (*Ant Colony System*) o Sistema de Hormigas [15], [16] y en la variante Sistemas de Hormigas Elitistas, EAS (*Elitist Ant System*) [17].

A. Codificación de la Hormiga

La hormiga artificial (individuo de la población) es un componente computacional simple que intenta construir soluciones válidas (una posible configuración de red) al problema de coste mínimo explorando los rastros de feromona disponibles (indica la afluencia que han tenido los movimientos por anteriores hormigas) y la información heurística (como el atractivo o la mejor conveniencia de los movimientos). Una determinada solución podría no ser válida, con lo que sufrirían cierta penalización. Una solución, s Eq. (3), es representada por una sucesión de niveles de energía, $ne_{e_{(i,j)}}$, donde cada uno de ellos



$e_{(1,2)}$	$e_{(1,4)}$	$e_{(1,5)}$	$e_{(2,1)}$	$e_{(2,3)}$	$e_{(3,2)}$	$e_{(3,4)}$	$e_{(4,1)}$	$e_{(4,3)}$	$e_{(4,5)}$	$e_{(5,1)}$	$e_{(5,4)}$
0.05	1	0.33	1	1	0.66	0.33	1	0.05	0.66	0.05	1

Fig. 2: Ejemplo solución de una hormiga con función de energía lineal y cuatro niveles de energía.

representa el nivel de energía en el que operará cada enlace de la red, $e_{(i,j)}$, identificado en s por i.

$$s = \{ne_0, ne_1, ne_2, ..., ne_{|E|}\}, \forall ne_i \in N, \forall i \in E \quad (3)$$

Por tanto, cada hormiga dentro del algoritmo tendrá por objetivo obtener una configuración de red ponderando para cada enlace de la misma el nivel de energía con el que va a operar. Para ello, la hormiga va construyendo la solución iterativa e incrementalmente recorriendo todos los enlaces de la red (de menor a mayor identificador de nodo origen).

En la Fig. 2 se muestra el ejemplo de una configuración de red como solución de una hormiga considerando una topología de red compuesta por 5 nodos y 12 enlaces unidireccionales, función de energía lineal, un conjunto de 4 niveles de energía y visibilidad proporcional. Cada enlace opera según un nivel de energía distinto dentro del rango $[0.05, 1]$ donde con un valor de 0.05 el enlace está dormido y por tanto con un 5% de consumo de energía asociado, y con un valor de 1 el enlace estaría activo (operando a su capacidad nominal) con un 100% de consumo de energía.

B. Regla de Transición

Es el modelo estocástico que emplean las hormigas para tomar la decisión de elegir un nivel de energía u otro en el enlace que están explorando, permitiendo balancear entre la exploración de nuevos caminos y la explotación de los conocimientos acumulados hasta el momento. Para este proceso se usan dos tipos de información.

- **Información memorística**, que corresponde con el aporte o nivel de feromonas, y que mide la deseabilidad aprendida del movimiento. Dado un enlace $e \in E$, cada nivel de energía posible bajo el que pueda operar $ne_e \in N$ tendrá un aporte de feromonas $\tau_{(e,ne_e)} \in F$ donde F es la matriz de feromonas. Este rastro de feromonas será modificado a lo largo de la ejecución del algoritmo aplicando la política de actualización y evaporación de las feromonas.
- **Información heurística** o alta visibilidad, que mide la preferencia heurística del movimiento, donde cada nivel de energía posible bajo el que pueda operar cada enlace tendrá una visibilidad $\eta_{(e,ne_e)} \in H$ donde H es la matriz de visibilidad. H tendrá unos valores iniciales y no modificables en la ejecución del algoritmo.

$$P^k_{(e,ne_e)} = \begin{cases} \frac{[\tau_{(e,ne_e)}]^\alpha * [\eta_{(e,ne_e)}]^\beta}{\sum_{ne_e \in N_e} [\tau_{(e,ne_e)}]^\alpha * [\eta_{(e,ne_e)}]^\beta} & \text{si } ne_e \in N_e \\ 0 & e.c.o. \end{cases} \quad (4)$$

La regla de transición se basa en la Eq. (4). Para una hormiga k que esté ubicada en el enlace $e \in E$, el nivel de energía $ne_e \in N_e$ tendrá una probabilidad de ser seleccionado $P^k_{(e,ne_e)}$. Además se introduce el parámetro α que estima la importancia relativa de la información memorística o del camino recorrido en función del rastro de feromonas depositado y β como la importancia de cercanía en términos de visibilidad o preferencia heurística.

C. Actualización y Evaporación de las Feromonas

La actualización de las feromonas se realiza mediante el Sistema de Hormigas-ciclo de ACS [15], [16], donde la deposición de las feromonas se lleva a cabo una vez que todas las hormigas de cada generación han obtenido una solución. Primero se produce un decaimiento o evaporación al valor de todos los rastros de feromonas, Eq. (5), con un factor constante γ para evitar un incremento ilimitado de rastro de feromonas y para permitir olvidar las malas decisiones tomadas.

$$\tau_{(e,ne_e)} = (1 - \gamma) * \tau_{(e,ne_e)} \quad (5)$$

Así, para el nivel de energía ne del enlace e, su valor asociado de rastro de feromonas será reducido una cantidad $\gamma \in (0, 1]$. Posteriormente se realiza una retroalimentación positiva a los rastros de feromonas Eq. (6). Para ello se recorren todas las soluciones de las hormigas que han sido válidas SV y se depositan feromonas a cada nivel de energía.

$$\begin{aligned} &\tau_{(e,ne_e)} = \tau_{(e,ne_e)} + \Delta\tau^k_{(e,ne_e)}, \forall ne_e \in s_k, \forall s_k \in SV \\ &\Delta\tau^k_{(e,ne_e)} = f(C(s_k)) = \tfrac{1}{C(s_k)} \end{aligned} \quad (6)$$

donde $\Delta\tau^k_{(e,ne_e)}$ es la cantidad de feromonas a retroalimentar, que depende de la calidad de la solución $C(s_k)$ de la hormiga k, al nivel de energía ne_e, que pertenece a la solución s_k.

$$\tau_{(e,ne_e)} = \tau_{(e,ne_e)} + \Delta\tau^S_{(e,ne_e)}, \forall ne_e \in S \quad (7)$$

Además, al aplicar EAS, se proporciona un peso adicional que refuercen los niveles de energía de aquellos enlaces que pertenecen a la mejor solución S encontrada hasta el momento Eq. (7).

Algoritmo 1 Pseudo-codigo de ACO genérico.

Require: Grafo dirigido: $G = (V, E, M)$
1: $inicializarParametros(G)$
2: $inicializarRastrosFeromonas()$
3: **for** $g = 1 : Generaciones$ **do**
4: $SV = construirSolucionesPorHormigas(G, g)$
5: $actualizarFeromonas(SV)$
6: **end for**
7: **return** S

Algoritmo 2 Construir soluciones por hormigas.

Require: Grafo dirigido, generación: $G = (V, E, M), g$
1: Soluciones válidas: $SV = []$
2: **for** $k = 1 : Poblacion$ **do**
3: $inicializarHormiga(k, g)$
4: **for** $e = 1 : E$ **do**
5: $P^k = calcularProbabilidadesTranscicion(e, N)$
6: $ne_e = aplicarPoliticaDecision(P^k)$
7: $s_k(e) = ne_e$
8: **end for**
9: **if** $comprobarValidez(s_k)$ **then**
10: $incluirSolucionValida(s_k, SV)$
11: $S = actualizarSolucionGlobalSiMejor(s_k)$
12: **end if**
13: **end for**
14: **return** SV

D. Comprobar la Validez de una Solución

Terminado el ciclo de construcción de la solución de una hormiga, es necesario comprobar su validez. Este proceso se realiza en dos pasos. Primero se comprueba que $\forall d_{(s,d)} \in M_{(i,j)}$, es posible encaminar todas las demandas de la matriz de tráfico dada aplicando Dijkstra [18] de modo que ninguno de los enlaces de la red se encuentra sobrecargado. Si la solución de la hormiga cumple dichos criterios de validación, se comparará con la mejor solución S encontrada hasta el momento. En el caso de que la solución de la hormiga sea energéticamente más eficiente que S, la sustituirá como nueva mejor solución encontrada.

E. Pseudocodigo ACO

El pseudo-codigo del algoritmo ACO genérico puede observarse en el Algoritmo 1, mientras que en la Fig. 3 se presenta de forma gráfica como sería la ejecución de la primera generación del algoritmo ACO. Dada una topología de red y sus demandas de tráfico, se inicializan los parámetros del algoritmo ACO y las feromonas (Estado 1 de la Fig. 3). Además, se construyen tantas configuraciones de red (soluciones de hormigas) como tamaño tenga la población por cada una de las generaciones y se almacenan aquellas configuraciones válidas, SV. Posteriormente, se actualizan los niveles de feromonas mediante las soluciones válidas, SV, encontradas en la generación actual. Al final de la ejecución del algoritmo se obtiene la mejor configuración de red encontrada en términos de eficiencia energética.

El Algoritmo 2, representado por el Estado 2 de la Fig. 3, realiza el proceso de construcción de soluciones de las hormigas donde por cada enlace, aplica la regla de transición a los niveles de energía mediante los valores

Algoritmo 3 Actualizar feromonas.

Require: Soluciones válidas: SV
1: ▷ $evaporarFeromonas(\gamma)$
2: **for** $feromona = f : F$ **do**
3: $f = f - \gamma$
4: **end for**
5: ▷ $retroalimentarFeromonas(SV)$
6: **for** $solucion = s : SV$ **do**
7: **for** $nivelEnergia = ne : s$ **do**
8: $F[ne] = F[ne] + f(C(s))$
9: **end for**
10: **end for**
11: ▷ $reforzarMejorSolucion(S)$
12: **for** $nivelEnergia = ne : S$ **do**
13: $F[ne] = F[ne] + f(C(S))$
14: **end for**

heurísticos y los rastros de feromonas, para obtener las probabilidades de elección de cada nivel de energía para el enlace. Posteriormente, con las probabilidades anteriores, aplica la política de decisión para seleccionar el definitivo nivel de energía para el enlace en cuestión (Estado 2.1 de la Fig. 3). Ponderados todos los enlaces con niveles de energía en la red, la hormiga devuelve s como solución potencial. Por último, si esta configuración de red s es válida, se compara con la mejor solución encontrada S hasta el momento y, en el caso de que presente un menor consumo de energía, pasaría a convertirse en la mejor solución encontrada, S.

Por último, en el Algoritmo 3, Estado 3 de la Fig. 3, se desglosa el proceso de actualización de la matriz de feromonas. Primero se realiza el decaimiento de todos los rastros de feromonas mediante la constante γ. En el siguiente paso se realiza una retroalimentación positiva a aquellos rastros de feromonas que pertenezcan a las soluciones válidas para, finalmente, reforzar aquellos rastros de feromonas que pertenecen a la mejor solución encontrada.

V. Resultados Experimentales

En esta sección se analizan los resultados obtenidos tras realizar un conjunto de pruebas en las que se aplica el algoritmo ACO propuesto sobre topologías de red de distinto tamaño.

A. Topologías y tráfico

Con el fin de evaluar nuestra propuesta, se ha optado por tres escenarios de red reales: NSFNet (14 nodos y 42 elaces), Géant (23 nodos y 74 enlaces) y Germany50 (50 nodos y 176 enlaces), con una carga de tráfico entorno al 50%. Tanto la información correspondiente a las topologías como las matrices de tráfico utilizadas se han extraído de [19].

B. Parámetros y valores de ACO

En la Tablas I se desglosan los parámetros utilizados por el algoritmo ACO para dar solución al problema de eficiencia energética en redes SDN. Se puede observar que se consideran dos tipos de matrices de visibilidad: Natural y Proporcional. A continuación se explican cada una de ellas.

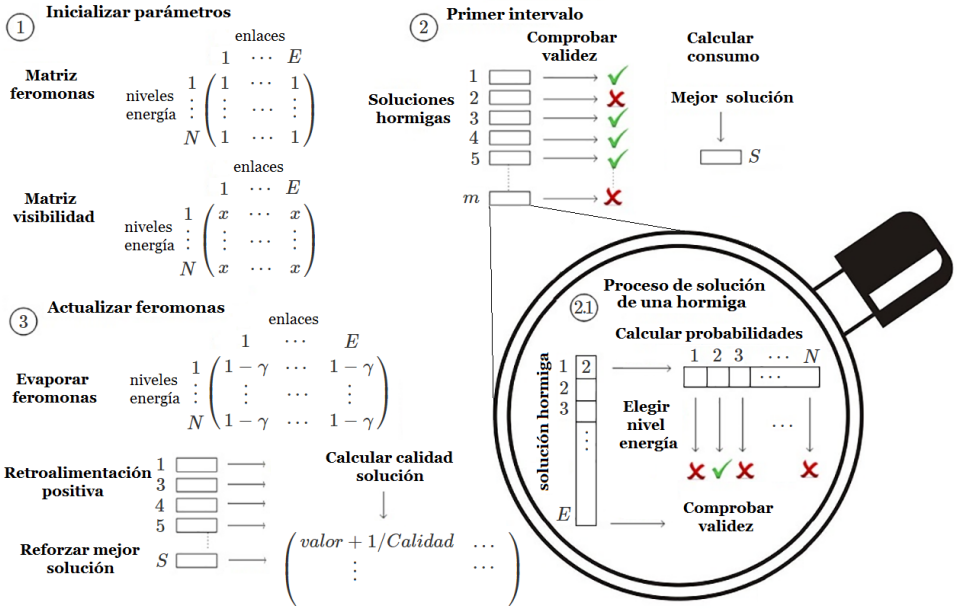


Fig. 3: Esquema del comportamiento del algoritmo ACO para el problema.

Tabla I: Parámetros de EEACO y valores considerados.

Parámetro	Descripción	Valores
m	Tamaño de población	20, 50, 80
g	Número de generaciones	5, 10, 20
α	Importancia relativa de las soluciones	1
β	Importancia de cercanía en términos de visibilidad	[2,5]
γ	Decaimiento de feromonas	(0,1]
N	Niveles de energía	2, 3, 4
H	Matriz de visibilidad	Natural, Proporcional
F_0	Matriz de feromonas inicial	Todos a 1
f	Función de energía	Lineal, Logarítmica
z	Repeticiones por prueba	5
T	Topologías consideradas	NSFNet, Géant, Germany50

Para componer la matriz de visibilidad Natural se aplica Eq. (8), donde los valores heurísticos tendrán el mismo valor que el nivel de energía considerado.

$$H_{(e,ne_e)} = ne_e \tag{8}$$

En cambio, para componer la matriz de visibilidad Proporcional se aplica Eq. (9). En este caso, los valores heurísticos están comprendidos entre $[0.05, 1]$. En la Fig. 4 se muestra un ejemplo de matrices de visibilidad obtenidas al aplicar las ecuaciones anteriores para una red compuesta por 5 enlaces que pueden operar según un conjunto de 4 niveles de energía.

$$H_{(e,ne_e)} = \begin{cases} 0.05 & \text{si } ne_e = 1 \\ \left(\frac{1}{N-1}\right) * (ne_e - 1) & \text{si } ne_e > 1 \end{cases} \tag{9}$$

C. Análisis de Resultados

En esta sección se analizan los resultados obtenidos tras ejecutar el algoritmo ACO sobre los tres escenarios de red considerados. La metodología lleva a cabo para la obtención de resultados consiste en generar una muestra de 5 repeticiones de ejecución del algoritmo ACO por cada combinación de parámetros considerados en la Tabla I con un nivel de confianza del 95%.

En las Figs. 5 y 6 se pueden observar los resultados obtenidos. En concreto, cada gráfica muestra los porcentajes de ahorro de energía obtenidos por el algoritmo ACO por cada función de energía y topología consideradas.

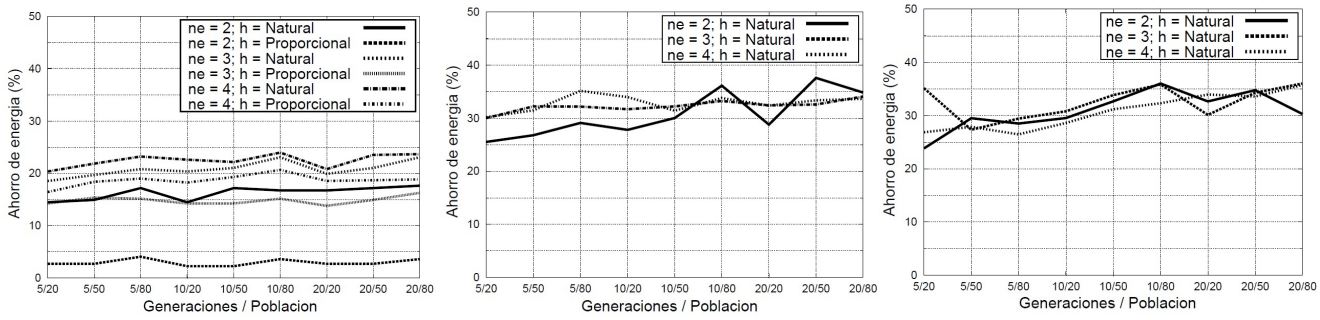
En la Fig. 5a se presentan los porcentajes de ahorro para las pruebas realizadas sobre la topología NSFNet con función de energía lineal. En ella, se puede observar que la visibilidad Natural se comporta mejor que la visibilidad Proporcional, llegando a unos valores máximos de ahorro energético cercanos al 25% para la visibilidad Natural, en comparación con el 20% de ahorro energético en el caso de la visibilidad Proporcional. Además, podemos observar que la mejor configuración en términos de eficiencia energética para la topología NSFNet con función de energía lineal, es considerar un conjunto de 4 niveles de energía y visibilidad Natural.

Para el caso de la función de energía logarítmica y la topología NSFNet, Fig. 6a, se pueden observar resultados similares a los obtenidos por las pruebas realizadas con función de energía lineal, es decir, la visibilidad Natural se comporta mejor que la visibilidad Proporcional. En cambio, si consideramos la función logarítmica, podemos observar que a mayor conjunto de niveles de energía, menor es el ahorro energético que se obtiene. Además, este comportamiento se repite para los dos tipos de visibilidad considerados. También cabe destacar que independientemente de la función de energía considerada, a

$$\text{niveles energía} \begin{pmatrix} & \text{enlaces} & & & & \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 & 4 \end{pmatrix} \qquad \text{niveles energía} \begin{pmatrix} & \text{enlaces} & & & & \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.33 & 0.33 & 0.33 & 0.33 & 0.33 \\ 0.66 & 0.66 & 0.66 & 0.66 & 0.66 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

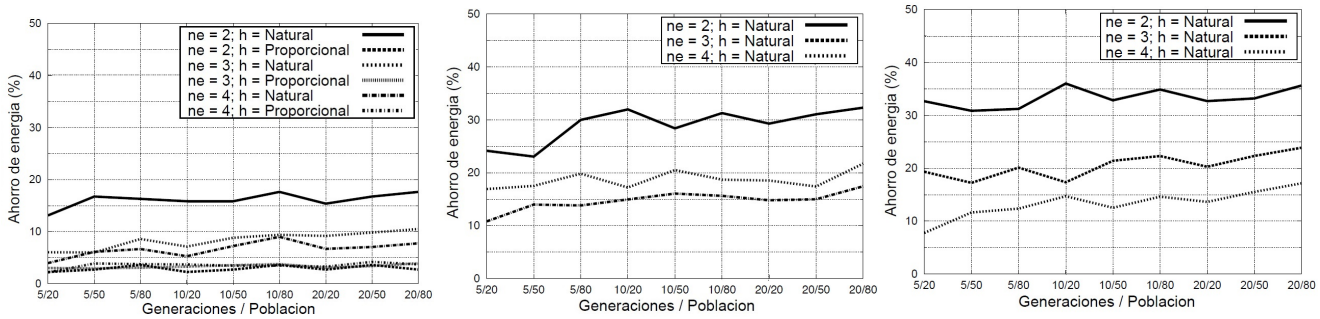
(a) Natural (b) Proporcional

Fig. 4: Ejemplo matrices de visibilidad.



(a) NSFNet (b) Geant (c) Germany50

Fig. 5: Porcentaje de ahorro de energía en función del número de niveles de energía por enlace con función de energía lineal.



(a) NSFNet (b) Geant (c) Germany50

Fig. 6: Porcentaje de ahorro de energía en función del número de niveles de energía por enlace con función de energía logarítmica.

mayor número de generaciones y número de hormigas o población, mayor es el ahorro energético resultado del algoritmo ACO. Los resultados aplicando las topologías Géant y Germany50 tienen las mismas consideraciones que los anteriormente analizados para NSFNet, con la diferencia de que en estas dos topologías solamente se tiene en cuenta la visibilidad Natural, que es la visibilidad que mejor se ha comportado para la topología NSFNet. Por regla general, al considerar la función de energía lineal, a mayor conjunto de niveles de energía, mayor es el ahorro energético obtenido por el algoritmo ACO.

Por último, el tiempo de cómputo y procesamiento del algoritmo ACO es proporcional al número de generaciones y población tomadas como parámetros de entrada. Es decir, a mayor número de generaciones (iteraciones del algoritmo) y a mayor tamaño de la población, mayor será el tiempo de cómputo requerido. Otro factor que afecta directamente al tiempo de cómputo es el escenario de red considerado, ya que una topología de mayor tamaño tiene una repercusión operacional y temporal mayor. La relación de tiempo de cómputo de Germany50 es aproximadamente 12 veces superior al tiempo requerido para dar solución en la topología Géant y ésta, a su vez, es 7 veces superior a NSFNet. En la Tabla II se muestra un resumen de los resultados de las pruebas donde por cada topología se muestran los porcentajes de ahorro máximo y ahorro medio de consumo según la función de energía considerada para los enlaces.

Considerando los resultados obtenidos con la función de energía logarítmica podemos observar que a mayor

Tabla II: Porcentajes de ahorro energético por topología.

Topología	% Max F.E.Lineal	% Medio F.E.Lineal	% Max F.E.Logarítmica	% Medio F.E.Logarítmica
NSFNet	23,98	19,87	17,64	10,39
Géant	37,58	31,95	31,96	20,83
Germany50	36,01	31,37	36,00	22,37

conjunto de niveles de energía menor es el ahorro energético obtenido. Este comportamiento radica en la propia definición de la función de energía logarítmica en comparación con la función de energía lineal, ya que habría que realizar una comparativa más justa de ambas funciones (Fig. 1) para que sus áreas delimitadoras con el eje de abscisas fueran iguales.

VI. Conclusiones y Trabajos Futuros

En este trabajo se propone una solución al problema del consumo de energía en redes SDN. El algoritmo propuesto se basa en el comportamiento de las colonias de hormigas (ACO) para encontrar la configuración de red óptima en términos de eficiencia energética. Mediante la utilización del rastro de feromonas, las hormigas (o soluciones al problema) van creando soluciones potenciales en función del modo de operación activo de los enlaces (definidos en el artículo como niveles de energía). Además, se consideran dos funciones de energía diferentes: i) lineal o proporcional, tomada como base, donde a mayor velocidad de operación del enlace mayor consumo de energía; y ii) logarítmica, que simula el comportamiento del estándar EEE.

Tras realizar un conjunto de pruebas sobre topologías de red reales (NSFNet, Géant y Germany50), para las cuales varían los parámetros del algoritmo ACO, se obtienen un conjunto de resultados donde se puede observar que es posible conseguir un ahorro de energía significativo en torno al 30% dependiendo de la función de energía utilizada. La inclusión y aplicación del algoritmo propuesto no modificaría la arquitectura SDN tradicional, ya que únicamente bastaría con instalarlo en el controlador SDN de modo que se ejecute cada vez que se active un disparador (o trigger) en función de la superación de un umbral (como por ejemplo un incremento o decremento de tráfico significativo).

Como trabajo futuro, se plantea la posibilidad de evaluar el delay que puede producirse al cambiar de configuración de red en función del tráfico. Para ello, se podrían utilizar herramientas de emulación de redes SDN como Mininet y de generación de tráfico como iPerf.

Agradecimientos

Este trabajo ha sido financiado, en parte, por el Ministerio de Economía, Industria y Competitividad (TIN2014-53986-REDT, TIN2015-67083-R y TIN2015-69957-R (MINECO/FEDER)), por el proyecto 4IE (0045-4IE-4-P) financiado por el Programa Interreg V-A España-Portugal (POCTEP) 2014-2020, por la Junta de Extremadura, Consejería de Economía e Infraestructuras (GR15098), y por el Fondo Europeo de Desarrollo Regional (FEDER).

Referencias

[1] E. Gelenbe and Y. Caseau, “The Impact of Information Technology on Energy Consumption and Carbon Emissions,” Ubiquity, vol. 2015, no. June, pp. 1–15, 2015.

[2] L. Neves et al., “Smart 2020 report—Enabling the low carbon economy in the information age,” The Climate Group, 2008.

[3] B. Aebischer and L. M. Hilty, ”The Energy Demand of ICT: A Historical Perspective and Current Methodological Challenges. Springer” International Publishing, 2015.

[4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, “Power awareness in network design and routing,” in Proc. INFOCOM, Phoenix, AZ, USA, Apr. 2008, pp. 1130–1138.

[5] M. Gupta and S. Singh, “Greening of the Internet,” in Proc. SIGCOMM, Karlsruhe, Germany, Aug. 2003, pp. 19–26.

[6] S. Nedevschi, L. Popa, G. Iannaccone, D. Wetherall, S. Ratnasamy, ”Reducing network energy consumption via sleeping and rate-adaptation,”Proc. 5th USENIX Symp. on Networked Systems Design and Implementation (NSDI ’08), San Francisco, CA, 2008, pp. 323-336.

[7] F. Idzikowski, L. Chiaraviglio, A. Cianfrani, J. L. Vizcaíno, M. Polverini, and Y. Ye, “A survey on energy-aware design and operation of core networks,”IEEE Commun. Surveys Tut., vol. 18, no. 2, pp. 1453–1499, 2nd Quart., 2015.

[8] L. Chiaraviglio, M. Mellia, F. Neri, ”Reducing Power Consumption in Backbone Networks”, Proc. 2009 IEEE Internat. Conf. on Communications (ICC 2009), Dresden, Germany, June 2009.

[9] S. M. Kerner, ”Energy Efficient Ethernet hits standards milestone — InternetNews:The Blog — Sean Michael Kerner”. Internetnews blog. Último acceso 25/7/2017.

[10] J. Galán-Jiménez, A. Gazo-Cervero, ”Using bio-inspired algorithms for energy levels assessment in energy efficient wired communication networks, Journal of Network and Computer Applications, Vol 37, (2014) pp 171–185.

[11] D. Kreutz et al., “Software-defined networking: A comprehensive sur- vey,” Proc. IEEE, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[12] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, “Elastictree: Saving energy in data cen- ter networks.” in NSDI, vol. 10, 2010, pp. 249–264.

[13] J. Wang, X. Chen, C. Phillips, and Y. Yan, “Energy efficiency with qos control in dynamic optical networks with sdn enabled integrated control plane,” Computer Networks, vol. 78, pp. 57 – 67, 2015, special Issue: Green Communications.

[14] M. Dorigo, 1992. Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italy.

[15] M. Dorigo, V. Maniezzo & A. Colorni, 1996. ”Ant System: Optimization by a Colony of Cooperating Agents”, IEEE Transactions on Systems, Man, and Cybernetics–Part B, 26 (1): 29–41.

[16] M. Dorigo, & T. Stützle. (2003). The ant colony optimization metaheuristic: Algorithms, applications and advances. In Glover, F. & Kochenberger, G. (Eds.), Handbook of Metaheuristics, 251-285. Kluwer Academic Publishers.

[17] E. Bonabeau, M. Dorigo, G. Theraulaz (1999). Swarm Intelligence: From Natural to Artificial Systems. New York: Oxford University Press.

[18] E. W. Dijkstra (1959). A note on two problems in connexion with graphs. Numerische Mathematik 1: 269-271. doi:10.1007/BF01386390.

[19] SndLib. Data Source working Optimizaction of Telecommunications Networks. URL: http://sndlib.zib.de/home.action

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6613

Red SDN para el Control Energéticamente Eficiente de un Aula Remota para la Elaboración de Prácticas Reales a Distancia

Marina Terrón-Camero[1], Sandra Sendra[1,2], Jorge Navarro-Ortiz[1], Jaime Lloret[2]

[1] Departamento de teoria de la señal, telemática y comunicaciones, ETS Ingenierías Informática y de Telecomunicación. Universidad de Granada

C/ Periodista Daniel Saucedo Aranda, s/n. E-18071 Granada

[2] Instituto de Investigación para la Gestion Integrada de zonas Costeras. Universitat Politècnica de València, Valencia

C/Paraninf, 1. 46730. Grao de Gandia (Valencia)

mtcmarina@correo.ugr.es, ssendra@ugr.es, jorgenavarro@ugr.es, jlloret@dcom.upv.es

***Resumen*- La demanda de cursos de redes a distancia y semipresenciales está aumentando día a día en el sector de las tecnologías de la información y comunicación (TIC). Para realizar este tipo de cursos se precisa tener laboratorios específicos que permitan la conexión remota, por lo que requieren que estén continuamente encendidos para su acceso. Este hecho implica un consumo de energía bastante elevado, aun cuando el laboratorio no está siendo utilizado. Además, el sector de las TIC es uno de los que presenta un consumo mayor a nivel mundial. Al ver la necesidad de controlar el consumo de estas instalaciones, se propone la implementación de un laboratorio real energéticamente eficiente, controlado mediante una red definida por software (SDN) de gestión. Se basa en un controlador y switches Openflow que en función del tipo de peticiones que reciba por parte de los usuarios remotos, distribuirá, de manera eficiente, el acceso a los diferentes laboratorios. Para ello, se implementa un algoritmo de control que ejecuta el controlador, y éste, actuará sobre diversos switches OpenFlow. Estos switches son los encargados de encender y apagar las diversas secciones de los laboratorios. Finalmente se simula el funcionamiento de este laboratorio para los 2 tipos de consumo de energía, es decir, usando la red de gestión SDN y cuando los laboratorios permanecen encendidos a la espera de las peticiones de conexión por parte de los usuarios. Los resultados demuestran que para el funcionamiento del laboratorio durante una semana, podríamos obtener un ahorro de energía cercano al 47%.**

***Palabras Clave*- OpenFlow, prácticas de redes, eficiencia energética, redes definidas por software (SDN), Green Networking.**

I. INTRODUCCIÓN

Con la evolución de la Internet de las cosas (IoT) y el aumento del número de dispositivos conectados a Internet, la demanda de la gente para tener conectividad a Internet en todas partes está aumentando. Para poder satisfacer esta demanda, necesitamos desplegar más redes e interconectarlas. Esto implica un importante aumento en el consumo de energía. El desarrollo de nuevas tecnologías y redes de comunicación para dar soporte a todas estas nuevas aplicaciones ha incrementado el consumo de energía eléctrica y la huella de carbono. Hoy en día, el consumo de energía en el área de telecomunicaciones se calcula que es el 4% del consumo mundial de electricidad [1]. El número de usuarios de Internet y el tráfico enviado a través de la red ha aumentado drásticamente y esto está directamente relacionado con el consumo de energía en las telecomunicaciones. De hecho, el consumo mundial de energía casi se ha duplicado desde 1990 [2].

Las redes verdes (o *Green Networking*, en inglés), se han convertido en uno de los principales temas de interés a nivel de investigación. El término *Green Networking* es un amplio término referido a los

procesos utilizados para optimizar la red o hacerla más eficiente, a la vez que conservan el ancho de banda. Como consecuencia reducen de forma indirecta el coste económico. Además, el término *'green'* puede llevar a confusión, debido a que también se suele usar para referirse a la energía sostenible, ahorro de agua, gas y electricidad y todo aquello que se considera *'smart'*, es decir, el uso de redes y tecnología inteligentes para la reducción del impacto medioambiental respecto a previas tecnologías.

Relacionado con este tema podemos encontrar diversas iniciativas, que van desde el desarrollo de protocolos especialmente diseñados para reducir el consumo de la red [3] [4] hasta la organización de las redes en determinadas topologías en función de la carga de red en los nodos, para mejorar la estabilidad de la red [5].

Sin embargo a nivel de red, hablar de *Green networking* implica hablar del uso de una tecnología que tiene beneficios medioambientales y que puede ser usada para soportar prácticas medioambientales más sostenibles. Esto se consigue buscando modos de funcionamiento menos costosos energéticamente, eligiendo cableado y materiales optimizados para este fin y en la elección de hardware con características energéticas reducidas [6].

Finalmente, debemos tener en cuenta que el rápido desarrollo que el sector de las tecnologías de la información y comunicación (TIC), hace que se requieran mayor cantidad de personal cualificado y con amplios conocimientos en redes. Esto ha implicado que los cursos formativos han ido poco a poco adaptándose a las necesidades de los estudiantes. Esta adaptación, se ha realizado desde diferentes vertientes, pero una de ellas y tal vez la más importante, sea la implementación de cursos a distancia [7] que posibilitan la realización de prácticas remotas [8]

Ante la necesidad de implementar este tipo de aulas y teniendo en cuenta las diferentes investigaciones relacionadas con técnicas de ahorro de energía en redes, en este artículo presentamos la implementación de un laboratorio real que permitirá la realización de prácticas a distancia. El conjunto de laboratorios es controlada por una red de gestión basada en switches *OpenFlow* y un controlador que será el encargado de registrar las peticiones de conexión y establecer las reglas necesarias para redirigir a los usuarios a un laboratorio u otro y de conectar y desconectar los laboratorios según sean necesarios. El diseño parte de la idea de funcionamiento de los *elastic tree* diseñadas para la reducción del consumo de energía en los data center [9]. Los resultados nos muestran valores de ahorro de energía superiores al 40% en algunos casos.

El resto del artículo se estructura como sigue. La sección II recoge algunas de las principales iniciativas relacionadas con el ahorro de energía en las redes. En la sección III, se explica la arquitectura propuesta y el algoritmo implementado para la implementación de una red de gestión definida por software (SDN) para el control energéticamente eficiente del laboratorio docente remoto. Los resultados obtenidos tras la simulación del escenario se muestran en la sección IV. Finalmente, las conclusiones y trabajos futuros se muestran en la sección V.

II. Trabajos Previos

Actualmente la sociedad científica está muy concienciada en el desarrollo y mejora de técnicas para la reducción del consumo de energía en las redes. Además, cada vez más se está intentando agregar estas iniciativas en los planes de estudios de grados relacionados con las TIC [10].

Tras analizar el tipo de arquitecturas desarrolladas para la realización de prácticas remotas, como por ejemplo, las presentadas por I. Santana et al. [11] y J. Sáenz [12], observamos que existen diversas propuestas para la realización de prácticas remotas. Sin embargo, en ninguna de ellas hemos encontrado la aplicación de una red de gestión para el control de acceso, y que además sea capaz de reducir el consumo de energía. Por ello, en esta sección nos centraremos en algunas de las propuestas más interesantes sobre reducción del consumo de energía en redes.

Las mejoras en términos de ahorro de energía pueden ser implementadas a distintos niveles, es decir, desde la modificación e inhabilitación del hardware inutilizado hasta la creación de nuevos protocolos de enrutamiento y MAC [13]. También es común encontrarnos con nuevos desarrollos de redes capaces de auto organizarse en función del tipo de variables sensadas, como es el caso de las redes basadas en grupo propuestas por J. Lloret et al. [14]. Sin embargo estas técnicas están siendo aplicadas sobre topologías tradicionales y en ningún caso se están aplicando sobre laboratorios de docencia.

Hoy en día, la tendencia de mejoras en las redes es bien distinta. Actualmente, se está tendiendo al uso de las SDN y la virtualización las redes [15]. A modo de resumen, podemos decir que las *SDNs* tienen como objetivo el facilitar la implementación y desarrollo de servicios de red de una manera determinista, dinámica y escalable, buscando reducir el papel del administrador de la red en las tareas de gestión de dichos servicios a bajo nivel.

J. M. Jimenez et al. [16] muestra una comparación de rendimiento entre Mininet y una red real cuando se entregan flujos multimedia. Los autores realizan la comparación en términos de ancho de banda consumido, delay y jitter. El estudio muestra que existen algunas diferencias importantes cuando se comparan estos parámetros, por lo que los autores presentan este trabajo como base para mostrar la diferencia con implementaciones reales cuando se usa Mininet.

D. Sarabia-Jácome et al [17] presentan una evaluación práctica donde se mide el consumo de energía en una red virtualizada. Para ello, hemos desarrollado una topología compuesta por varios enrutadores virtuales que emulan el núcleo de una red a través de la cual hemos enviado varios flujos de datos.

Las mediciones se han realizado utilizando *HTTP* y tráfico de flujo de vídeo mientras se ejecuta *Open Shortest Path First (OSPF)* y *Routing Information Protocol versión 2 (RIPv2)* como protocolos de enrutamiento. El resultado muestra que ambos protocolos presentan un comportamiento similar cuando se envía tráfico de flujo mientras *RIPv2* exige más uso de la CPU cuando se envía tráfico http. Por último, el consumo de energía en el tráfico http es siempre mayor.

La reducción del consumo de energía en las *SDN*, es hoy en día un tópico candente del que encontramos muy pocas referencias. Por ejemplo, B.B. Rodrigues et al. [18] exponen las dificultades que podemos encontrarnos al implementar un entorno *SDN* donde emular protocolos de ahorro de energía a diferentes capas de la red. Los autores proponen una solución llamada *GreenSDN*, basadas en el entorno de emulación *Mininet* y el controlador *POX OpenFlow* usando tres protocolos de red diseñados para el ahorro de energía. Los autores comparan los ratios de ahorros de energía obtenidos cuando estos protocolos son empleados. Los resultados muestras que la peor de las opciones, ofrecen ahorros de energía superiores al 15% y que este ahorro es más notable cuando la velocidad de transferencia de datos es mayor.

Finalmente, cabe destacar el concepto de *elastic tree* que está siendo aplicado para la reducción del consumo de energía global de un data center. Los *Elastic tree* [9] se aplican principalmente en el control de energía en redes de data center. Creemos que este mismo concepto puede ser aplicado a cualquier tipo de red donde exista una gran densidad de dispositivos de red. Nosotros, además, añadimos una red de gestión *SDN* basada en un controlador y *switches OpenFlow*, que serán los encargados de controlar en encendido y apagado de los laboratorios en función de las peticiones recibidas por parte de los usuarios. Todo esto nos permitirá reducir considerablemente el consumo de energía en la red entera.

III. Arquitectura propuesta y Algoritmo de funcionamiento

Cuando implementamos prácticas remotas para los cursos de redes debemos asegurarnos que todos los usuarios registrados en los cursos tienen acceso a todos los equipos. Para ello, debemos tener una arquitectura física con suficientes equipos disponibles y preparados para ser usados. Sin embargo esto acarrea un consumo de energía considerable, debido al propio funcionamiento de *switches*, *routers* y sistemas de refrigeración. Por tanto, necesitamos un modo de controlar este consumo energético.

En esta sección se muestra una arquitectura mixta formada por una red de gestión *SDN* y los laboratorios desplegados para la realización de las prácticas remotas. La sección muestra la arquitectura desplegada y el algoritmo de funcionamiento que regulará el encendido y apagado de los laboratorios.

A. Conceptos teóricos básicos: Elastic tree.

Los *Elastic Tree* se definen como un controlador de la potencia a lo largo de la red, el cual ajusta dinámicamente el conjunto de todos los elementos activos de la red, refiriéndonos tanto a enlaces como switches, para satisfacer los cambios en la carga de tráfico de la red.

Los objetivos que persiguen los *Elastic Tree* es optimizar la energía a lo largo de la red que monitoriza las condiciones de tráfico del centro de datos de forma continua. En los *Elastic tree* se elige un conjunto de elementos de la red que deben estar activos para el correcto funcionamiento del sistema y la tolerancia a los fallos, por lo que apaga o 'duerme' los enlaces y switches que no son necesarios ante la carga de tráfico dada en ese instante. Los *elastic tree* se dividen básicamente en 3 módulos lógicos que podemos fácilmente identificar con nuestra arquitectura (Ver Figura 1).

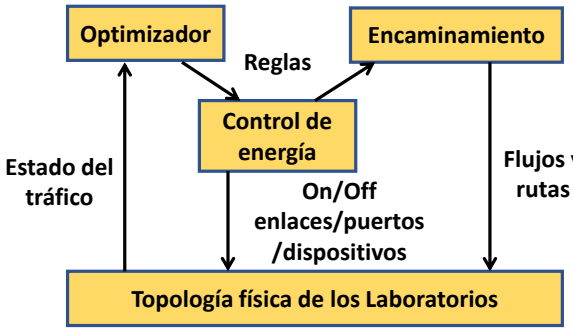


Fig. 1. Arquitectura Elastic tree aplicado a nuestra topología.

Aunque este tipo de sistemas suele aplicarse a escenarios típicos de data centers, nosotros deseamos aplicarlo a nuestra arquitectura y comprobar las mejoras en consumo de energía que su aplicación puede aportarnos. Los resultados demuestran que para las cargas producidas en centros de datos, *Elastic Tree* puede ahorrar hasta el 50% de la energía de la red, mientras que mantiene la capacidad de manejar las oleadas de tráfico.

B. Arquitectura y red SDN de gestión para el control de los laboratorios docentes.

El escenario que se propone para este trabajo es un laboratorio de redes en la modalidad de docencia virtual. Este laboratorio será accedido por parte de los alumnos, desde e.g. sus casas, para poder desarrollar diferentes prácticas de redes.

Tal como se muestra en la Figura 2, la red a la que los alumnos accederán a través de Internet está dividida en dos partes funcionales. Por un lado, una red *SDN* que será la encargada de gestionar la distribución de alumnos para el uso del equipamiento de las prácticas. Por otro lado, las redes de los laboratorios que contendrán los equipos que usarán los alumnos.

La red SDN de gestión está formada por un *frontend*, que actuará como elemento de entrada, y una red de distribución compuesta por una serie de *switches*

que encaminarán al alumno hacia los equipos que debe utilizar para su práctica.

La red para prácticas estará formada por bloques idénticos (en el ejemplo, compuestos de 6 *routers*, 6 PCs y los *switches* de interconexión necesarios). Todos tendrán una configuración inicial idéntica (mismas conexiones, direcciones IP, etcétera). Cada alumno conectado al sistema recibirá un bloque libre que utilizará en exclusiva para la realización de su práctica. En el ejemplo de la Figura 2, se ha supuesto que se dispone de 6 laboratorios con equipamiento para prácticas, estando cada uno formado por 6 armarios que contienen 1 bloque cada uno. Desde un punto de vista energético, cada laboratorio dispone de un sistema de climatización que se encenderá cuando alguno de sus armarios esté en uso. Para la asignación de estos bloques, será la red *SDN* de gestión la que se encargue de asignar bloques a los alumnos de manera totalmente transparente. Todos los alumnos se conectarán de igual forma y verán la misma configuración de partida, independientemente del bloque asignado.

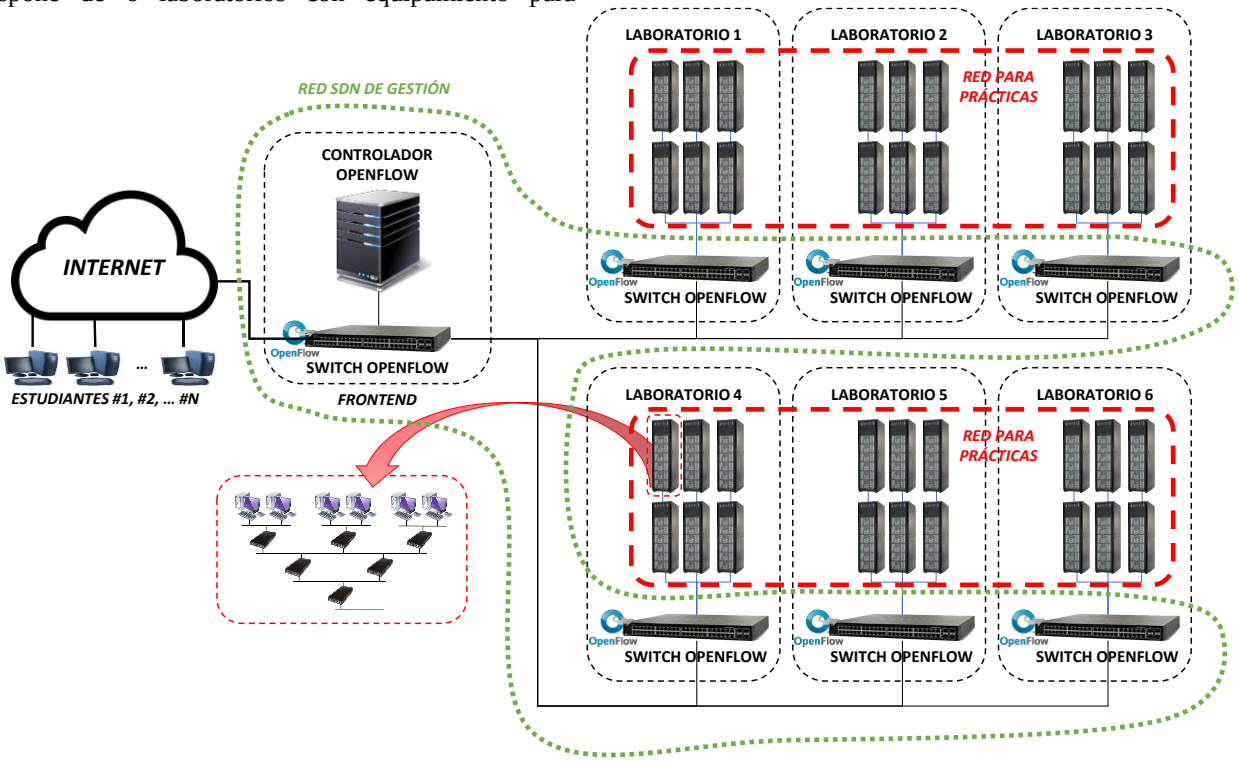


Fig. 2. Laboratorio de redes para docencia remota.

El primer elemento de la red de gestión será el *frontend*. Éste está formado por un *switch* con soporte de *OpenFlow* que está conectado a un controlador. Este controlador será el que implemente la inteligencia del sistema, modificando las reglas de flujos en los switches de la red de distribución. Para ello, cuando un nuevo usuario se conecta al sistema, el *switch OpenFlow* de entrada reenviará la petición al controlador, al tratarse de un flujo aún no incluido en su tabla de flujos. El controlador utilizará una tabla donde guardará qué equipos de los laboratorios están siendo utilizados, y le asignará uno de los bloques que aún no han sido utilizados. Para ello, el controlador modificará las tablas de flujos de los *switches* de la red de distribución, de manera que el nuevo flujo (identificado e.g. por la dirección IP del alumno) será encaminado de forma transparente a su bloque correspondiente. Una vez finalizada la práctica, el bloque volverá al pool de bloques que pueden utilizar alumnos nuevos.

De esta manera, el alumno sólo necesita conocer la dirección IP (o el nombre de dominio) del *frontend.*

C. Algoritmo de control.

Para realizar el control de toda la arquitectura de manera correcta debemos tener en cuenta varios aspectos. El primero de ellos es que deseamos dar servicio a todos nuestros alumnos; el segundo objetivo es que deseamos reducir el consumo energético global, por último y como aplicación futura, podríamos hacer una subdivisión de armarios, por tipo de práctica.

La Figura 3 muestra el diagrama de funcionamiento de la red de gestión. Como podemos ver, el controlador está a la espera de peticiones de conexión. Cuando se realiza la petición por parte del alumno mediante el protocolo seguro *Secure Shell (SSH),* el sistema analizará si este usuario está autorizado en el sistema o no. Tras esta comprobación, se le dará acceso un armario del laboratorio. Para ello, la red de gestión SDN comprueba si existe disponibilidad en el laboratorio actual y si no lo hay, el sistema deberá poner en marcha el siguiente laboratorio. El controlador periódicamente chequea si el usuario sigue activo o no, para liberar dicho armario. Finalmente la red de gestión

comprobará si siguen usuarios conectados a dicho laboratorio y si no es así, el laboratorio pasará a un estado *sleep*.

Cuando un laboratorio es puesto en marcha, también debe encenderse la refrigeración para mantener las condiciones ambientales idóneas para el correcto funcionamiento.

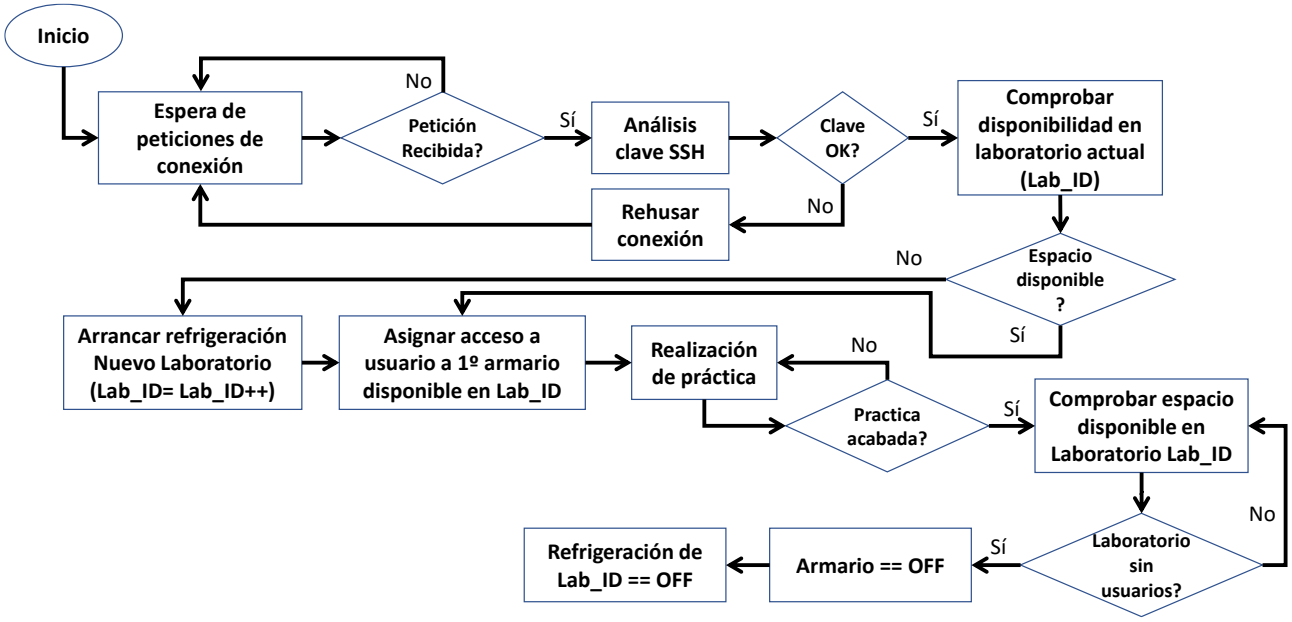


Fig. 3. Algoritmo de funcionamiento de la red de gestión SDN.

IV. Resultados

Para comprobar la viabilidad de la propuesta, hemos implementado y programado el algoritmo explicado en la sección anterior, así como la implementación de los algoritmos de decisión de los *elastic tree*. Como herramienta de simulación hemos empleado Matlab. Esta sección muestra los resultados de simulación, cuando tenemos un laboratorio sin aplicar el balanceo de energía que los *elastic tree* y cuando es aplicado.

El escenario desarrollado se basa en el comportamiento del laboratorio durante una semana. Se considera que a este laboratorio pueden conectarse estudiantes en cualquier franja horaria, al azar y de manera aleatoria. Hemos tomado como referencia de consumos de energía de los dispositivos, los especificados por B. Heller et al. en [9]. En la simulación se muestran las mejoras introducidas gracias al uso de *Green Networking* en el escenario presentado (6 laboratorios, 6 armarios por laboratorio, un *switch* más por laboratorio, un refrigerador por laboratorio, y el *switch* del pasillo conectado al controlador).

Para su puesta en marcha se tienen en cuenta el número de módulos activos, entendiendo como módulo cada uno de los armarios del laboratorio. Dependiendo del número de módulos requeridos, se encienden más o menos *switches*, y cuando todos los módulos de un laboratorio están en funcionamiento, en caso de necesitar más módulos se pasa al siguiente laboratorio y así sucesivamente. Si el laboratorio no está en uso no se enciende el módulo de refrigeración, y los *switches* se encuentran en *sleep mode*. El *switch* del pasillo siempre está en uso.

En total, consideramos 36 módulos. Consideraremos que se registra una carga baja cuando se está trabajando con 1 laboratorio, carga media cuando se trabaja con hasta 3 laboratorios y, por último, la carga muy alta considera el uso de todos los laboratorios.

La Figura 4 muestra el consumo energía de los laboratorios sin considerar el balanceo de energía. Podemos observar que el consumo para los 3 momentos del día se mantiene en todos los casos, entre los 40 y 50 kW, ya que en ningún caso, se contempla la posibilidad de que los dispositivos no utilizados sean desconectados.

La Figura 5 muestra el consumo energía de los laboratorios cuando se considera el balanceo de energía y se aplica nuestra propuesta. Lo primero que observamos es que los consumos de energía varían entre días y momentos del día. Esto se debe precisamente a que en este tipo de sistemas puede resultar complicado saber cuántas peticiones de conexión vas a tener por día. En cualquier caso, podemos ver el martes, presenta un consumo similar al registrado en la Figura 4, sin embargo, días como el domingo, el consumo de energía no supera en ningún caso los 30 kW.

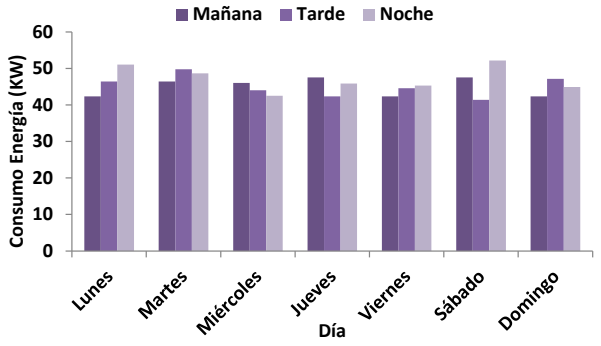


Fig. 4. Consumo energía de los laboratorios sin considerar el balanceo de energía.

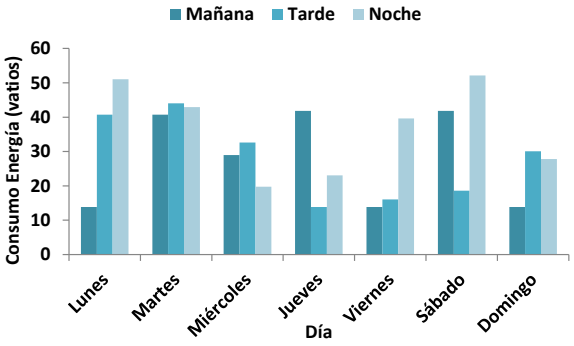


Fig. 5. Consumo energía de los laboratorios usando el balanceo de energía.

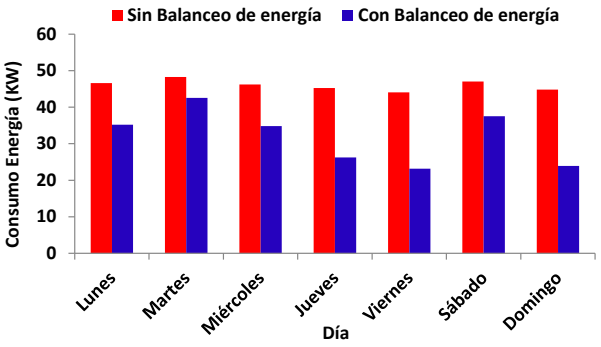


Fig. 6. Valor medio consumo energía por día para ambos casos.

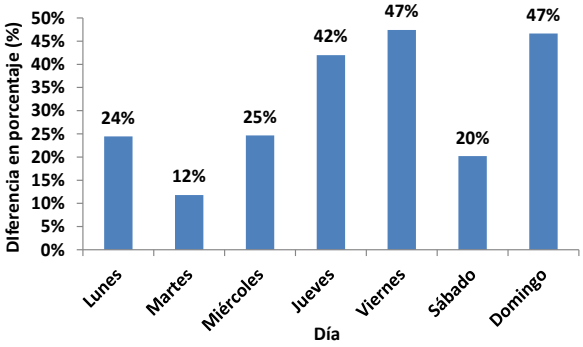


Fig. 7. Porcentaje de ahorro de energía medio por día.

Para comparar la energía consumida diariamente, calculamos una estimación del consumo medio por día. La Figura 6 muestra los resultados obtenidos. En este caso, se ve claramente que mientras que cuando no se aplica nuestra propuesta, el consumo medio diario se sitúa en torno a los 46 kW, cuando aplicamos el balanceo de energía mediante nuestra propuesta, obtenemos valores medios comprendidos entre los 23 kW y los 42 kW.

Por último, en la Figura 7 calculamos el ahorro de energía que implicaría tener implementada nuestra propuesta. Vemos que podemos conseguir ahorros de un 47% en el consumo de energía. En una semana se observa un 31% de ahorro de energía. Este hecho implica un considerable ahorro económico y una reducción importante en la huella de carbono que emitimos en la producción de dicha energía eléctrica.

V. Conclusiones

Con el paso de los años y el aumento del número de dispositivos de red, cada vez más se está demandando personal altamente cualificado en temas de gestión de redes. Por ello, con mayor frecuencia, estamos recibiendo peticiones de cursos a distancia y que ofrezcan la posibilidad de practicar con equipos reales sin restricciones horarias. Sin embargo, tener una serie de laboratorios disponibles 24h implica un elevado consumo de energía y por ello creemos necesaria la mejora de este hecho.

Por ello, en este artículo hemos propuesto la implementación de un laboratorio docente que permita el acceso desde internet para la realización de prácticas reales a distancia. Además este laboratorio está gestionado por una red *SDN* basada en *OpenFlow* que gestiona las peticiones de conexión por parte de los alumnos y los redirige a los laboratorios disponibles, de una manera eficiente. Esta eficiencia viene dada por aprovechar los armarios y equipos disponibles hasta completar un laboratorio completo y mantener apagados aquellos que no sea necesario utilizar.

Para comprobar la viabilidad de nuestra propuesta hemos simulado el funcionamiento de una arquitectura formada por 6 laboratorios con 6 armarios, que reciben peticiones por parte de usuarios de manera aleatoria. La idea básica del algoritmo propuesto se basa en los *elastic tree*, una técnica muy empleada en la gestión energéticamente eficiente de los data center, activando y desactivando enlaces.

Como resultado hemos podido observar mejoras en el consumo energético de hasta el 47%. Dicho valor depende en gran medida del número de usuarios conectados, que requieran su utilización.

Como líneas futuras, deseamos implementar laboratorios específicos clasificados según el tipo de prácticas, ya que, en cursos como los impartidos por *Cisco Networking Academy*, diferencian entre prácticas básicas, prácticas de encaminamiento, prácticas de conmutación y prácticas de seguridad. Además, nos gustaría medir el consumo real de energía de los dispositivos empleados para virtualizar los diferentes equipos [17] y ejecutar *OpenFlow*.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado por el "Ministerio de Ciencia y Tecnología", a través de la a través de la convocatoria 2014. Proyectos I+D - Programa Estatal de Investigación Científica y Técnica de Excelencia, a través del "Proyecto Subprograma Estatal de Generación de Conocimiento" (TIN2014-57991-C3-1-P) y por el "Ministerio de Economía y Competitividad" a través de la convocatoria 2016. Proyectos I+D+I - Programa Estatal de Investigación, Desarrollo e innovación orientada a los retos de la sociedad, (TEC2016-76795-C6-4-R).

REFERENCIAS

[1] G. Koutitas, and P. Demestichas, "A review of energy efficiency in telecommunication networks", Telfor journal, 2010. Vol. 2, No.1, pp. 2-7.

[2] Infoplease, "World Energy Consumption and Carbon Dioxide Emissions, 1990–2025". Available at: https://www.infoplease.com/science-health/energy/world-energy-consumption-and-carbon-dioxide-emissions-1990-2025. [Last access: May 12, 2017]

[3] R. Azizi. "Consumption of Energy and Routing Protocols in Wireless Sensor Network". Network Protocols Algorithms, 2016, Vol 8, No 2, pp76-87.

[4] M. Su-Qin, G. Yu-Cui, L. Min, Y. Yu, C. Ming-Zhi, "A Cluster Head Selection Framework in Wireless Sensor Networks Considering Trust and Residual Energy", Ad Hoc and Sensor Wireless Networks, 2015, Vol. 25, Num. 1-2, pp. 147-164.

[5] M. Fereydooni, M. Sabaei and G. Babazadeh, "Energy Efficient Topology Control in Wireless Sensor Networks with Considering Interference and Traffic Load", Ad Hoc and Sensor Wireless Networks, 2015, Vol. 25, Num. 3-4, pp. 289-308.

[6] S. Andrade-Morelli, E. Ruiz-Sánchez, S. Sendra, J. Lloret, "Router Power Consumption Analysis: Towards Green Communications", Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2012,Vol. 113, No.-, pp 28-37

[7] S. Sendra, J. M. Jiménez, L. Parra, J. Lloret, "Blended Learning in a Postgraduate ICT course", en las actas del 1st International Conference on Higher Education Advances, HEAd'15. 24-26 de Junio de 2015, València (España)

[8] J. Lloret, J.M. Jimenez, J. R. Diaz, G. Lloret. "A Remote Network Laboratory to Improve University Classes". The 5th WSEAS/IASME International Conference on engineering education (EE'08), Heraklion, Creta. (Grecia), 22 - 24 July, 2008.

[9] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, N. McKeown. "ElasticTree: Saving Energy in Data Center Networks". En las actas del 7th USENIX conference on Networked systems design and implementation (NSDI 2010), 28 – 30 de abril de 2010, San Jose, (California – EEUU).

[10] F. Sánchez Carracedo, D. López Álvarez, J. García Almiñana. "El desarrollo de la competencia Sostenibilidad y Compromiso Social en la Facultat d'Informàtica de Barcelona". En las Jornadas de Enseñanza Universitaria de la Informática, Universidade de Santiago de Compostela. Escola Técnica Superior d'Enxeñaría. 2010. 7 de julio de 2010, Santiago de Compostela (España). Pp. 249-256.

[11] I. Santana, M. Ferre, E. Izaguirre, R. Aracil, L. Hernandez, "Remote Laboratories for Education and Research Purposes in Automatic Control Systems," en IEEE Transactions on Industrial Informatics, 2013, vol. 9, no. 1, pp. 547-556.

[12] J. Sáenz, J. Chacón, L. De La Torre, A. Visioli, S. Dormido, "Open and Low-Cost Virtual and Remote Labs on Control Engineering", IEEE Access, 2015, Vol.3, No.-, Pp. 805-814

[13] S. Sendra Compte, J. Lloret, M García Pineda, J. F. Toledo Alarcón, Power saving and energy optimization techniques for Wireless Sensor Networks", Journal of Communications. 2010, Vol. 6, No. 6, pp. 439-459.

[14] M. Garcia, S. Sendra, J. Lloret, R. Lacuesta. "Saving energy with cooperative group-based wireless sensor networks". En Cooperative Design, Visualization, and Engineering. CDVE 2010. Lecture Notes in Computer Science, vol 6240, pp. 73-76. Springer, Berlin, Heidelberg.

[15] H. Wang, Y. Li, D. Jin, P. Hui, and J. Wu, "Saving energy in partially deployed software defined networks". En IEEE Transactions on Computers, 2016, vol. 65, No. 5, pp.1578-1592.

[16] J. M. Jimenez, O. Romero, A. Rego, A. Dilendra, J. Lloret, "Study of multimedia delivery over software defined networks". Network Protocols and Algorithms, 2016, 7(4), pp. 37-62.

[17] D. Sarabia-Jácome, A. Rego, S, Sendra, J. Lloret, "Energy Consumption in Software Defined Networks to Provide Service for Mobile Users", en las actas del 13th International Wireless Communications and Mobile Computing Conference, 26-30 de Junio de 2017, Valencia (España).

[18] B. B. Rodrigues, A. C. Riekstin, G. C. Januário, V. T. Nascimento, T. C. M. B. Carvalho, C. Meirosu, "GreenSDN: Bringing energy efficiency to an SDN emulation environment,", En las actas de 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM2015), 11-15 de Mayo de 2015, Ottawa (Ontario), 2015, (pp. 948-953).

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6490

PaCoVNE: Power Consumption Aware Coordinated VNE with Delay Constraints

Khaled Hejja, Xavier Hesselbach
Dept. Ingeniería Telemática
Universitat Politècnica de Catalunya
C/ Jordi Girona, 1-3 - Edif.C3 - Campus Nord - 08034 Barcelona - Spain
{khaled.hejja, xavier.hesselbach}@upc.edu

***Abstract*—This paper introduces a more efficient embedding approach, called power consumption aware and coordinated VNE heuristic, denoted as (PaCoVNE). It embeds both virtual nodes and edges, simultaneously, and within one stage, while satisfying CPU and BW constraints, minimizes power consumption of the whole substrate network, and considers end-to-end delay as a major constraint. Performance of the new heuristic was compared to the energy aware algorithm OCA/EA-RH, for off-line scenario using homogeneous configurations, with and without end-to-end delay. The paper also presents simulation results once without end-to-end delay, and also when it was included.**

Keywords: Virtual Network Embedding, Power Consumption, Coordinated, Delay

I. INTRODUCTION

Networks virtualization has become an integral component of future Internet, offering network operators a way to overcome ossification of the Internet, by consolidating many of their equipments onto standardized high volume components located at centralized data centers [2]-[4]. More specifically, the key advantageous of network virtualization are basically related to efficiently utilizing physical network resources through sharing them among several virtual networks (VN), as well as providing more flexibility to manage, expand, or shrink the physical network according to VNs' characteristics.

However, allocating enough resources to satisfy all requirements of a virtual network request (VNR), on top of a substrate network (SN) that has limited residual capacities, is a very challenging task in network virtualization [5]. To realize that, VNE process is usually divided into two sub-problems, the first one is allocating virtual nodes onto physical nodes, which is known as virtual node mapping (VNM) stage. The other one, is virtual edge mapping (VEM), which embeds virtual edges onto physical paths connecting corresponding nodes in the physical network. Along such process, VNE usually trades off between minimizing embedding costs through utilizing less SN resources, and maximizing revenues through accepting as much as possible VNRs, while maintaining acceptable quality of services (QoS).

Generally, VNM and VEM stages can be carried out in two strategies, uncoordinated or coordinated [5]. Regarding the uncoordinated case, VNM and VEM used to be solved independently without any coordination between the two stages, raising the possibilities of higher VNRs rejections. This is because VEMs could be mapped on longer physical paths, therefore, utilizing additional resources, consuming more power, and adding more delay due to passing through hidden hopes [6]. The other strategy, is performing both VNM and VEM in two separate, but coordinated stages, where VNM is performed according to predefined VEM constraints to guide allocating the virtual nodes [7]. However, even through there is a sort of coordinating VNM with VEM, still, virtual nodes could be embedded at physical nodes that could be farther away from each other, enforcing edges to be mapped at longer physical paths, resulting on similar disadvantageous as in the uncoordinated scenario. Furthermore, regardless of the used strategy, VNE used to be constrained by CPU and BW resources, but occasionally considering power consumption, and almost very seldom adding delay as an additional constraint. Thus, it could be possible that, the lack of considering more constraints throughout the VNE process, would result on a degraded QoS for the whole embedding process, including raising operational costs, consuming more power, as well as generating less revenues.

In view of that, this paper introduces the PaCoVNE approach, as a fully coordinated VNE algorithm. It performs virtual nodes and edges embeddings simultaneously and in one stage, according to the following constraints combined: CPU, throughout, power consumption and end-to-end delay. The core of PaCoVNE approach is based on formulating VNR's demands and SN paths' resources into two separate sets, called (Segments), one for VNR and another one for a precisely selected SN path. The ***VNR segment*** (Seg^r) is defined as a set of parameters, grouped as one entity, representing demands of virtual nodes and edges. While ***SN path's segment*** (Seg^S) is defined as a set of parameters; also grouped as one entity, representing resources of the physical nodes and edges belonging to a specific selected SN path. Both, VNR and SN segments must be identical in terms of number of nodes and edges in order to compare them element by element. Subsequently, PaCoVNE starts VNE process to minimize total power consumption in the whole SN, by comparing each element in the VNR segment to its corresponding element in the SN path segment, then deciding if SN path has enough CPU and throughput resources to accommodate the VNR, while considering end-to-end delay.

Main contributions:

1) PaCoVNE heuristic is introduced as a one stage coordinated VNE approach constrained by CPU, BW, and end-to-end delay to minimize total power consumption of the whole SN.
2) Analysis of PaCoVNE was performed for off-line scenario, using homogeneous and heterogeneous SN settings.
3) Comparison was conducted against one of the most refer-

enced energy efficient embedding algorithms, the energy aware relocation heuristic (OCA/EA-RH) given by [9].

Rest of the paper is organized as follows: Section II provides related work. System model is introduced in section III, followed by ILP problem formulation in section IV. Design of the proposed PaCoVNE heuristic is shown in section V, and performance evaluation is presented in VI. Then results and discussion are included in section VII, while section VIII concludes the paper and highlights some future work.

II. RELATED WORK

One of the main benefits of network virtualization is its ability to consolidate network resources by hosting them on the same substrate resource, which allows for reducing energy consumption and cost [9],[13]. In most cases, saving energy in networks has been devoted to the reduction of energy consumption in a single networking device or parts of a device, and not power saving in the whole network, where unused resources could be but into sleeping mode or turned off completely. Other approaches performed VNE on small parts of the SN, then widen the area if no sufficient power resources were found on SN. Moreover, virtual resources can be migrated to balance the overall load in an energy efficient way, thus reducing the total power consumption of the network without compromising QoS or VNRs' acceptance ratio. More details about most related and recent literature about energy aware VNE approaches are summarized in the following paragraph:

A modified VNE algorithm was presented by [8], which prefers SN nodes consuming less power and selects edges in an energy efficient path, then in [9], they developed a scalable energy-aware reconfiguration heuristic approach, including embedding cost and load balancing. The heuristic considers a set of embedded VNRs as input to perform an energy efficient relocation of resources, without impacting the acceptance ratio. [10] proposed to maximize the accepted VNRs while minimizing the energy cost of the whole system. They followed two observations, first embed VN nodes on SN nodes that has lowest electricity price, second embed VN nodes on an already active SN as much as possible, then put other nodes that has no load into sleeping mode.

Moreover, [11] developed an embedding algorithm that embeds a subset of VNRs into a subset of cleanest SN resources in terms of CO_2 emissions resulting from the energy usage, while satisfying the VNR constraints. They constrained the VNE process by introducing link delay, packet loss, used energy source, VNR priority and location. The authors showed that the embedding guarantees reduced number of substrate resources and cost, faster embedding time, and reduction of carbon footprint of the VNE operation. While in [12], the authors designed an MILP and a real time heuristic algorithm that considers granular power consumption of all devices in an IP over WDM network. They tried to consolidate the nodes embeddings by filling the ones with the least residual capacity before switching on others, as well as consolidating more than virtual node at the same data center to minimize additional hop counts. And in [13] The authors considered an energy efficient VNE in the IP network over the WDM optical network, by adapting a feedback control approach performing the embedding on a smaller set of SN resources. A limited mappable area consisting of a selection of candidate nodes is located first, then they check if VN embedding was successful, if not, then a feedback control approach is triggered to search for a wider mappable area, and the whole process repeats again. In this way, they managed to increase number of hibernated links and nodes, resulting on reducing energy consumption by the SN.

III. SYSTEM MODEL

The aim of this paper is to perform coordinated VNE that minimizes total power consumption in the whole SN. Consequently, following paragraphs explain the overall design model for VNE system, starting by defining SN model and introducing its notations. Then VN's model definition and notation will be explained, as well as defining the used power consumption model by PaCoVNE.

A. Substrate Network Model:

The physical network $G^S = (N^S, E^S)$ is modeled as a weighted directed graph, where: i and $j \in N^S$ are SN nodes, and $(i,j) \in E^S$ is an edge connecting nodes i and j. Each node $i \in N^S$ is associated with pw_i^{idle} representing average power value when the sever is idle, PW_i^{Busy} average power value when the server is fully utilized, PC_i total power consumption of i, as well as cpu_i^a representing current available CPU capacity, cpu_i consumed CPU capacity, and CPU_i as the maximum CPU capacity at node i. μ_i is a fractional value (consumed to maximum CPU capacity, which could reach a value of 1 maximum) representing the processing utilization of node i defined in the range (0-1), zero if node i is not loaded, up to 1 if its 100% loaded. Each substrate edge (i,j) is associated with bw_{ij}^a, representing current available bandwidth capacity, bw_{ij} as consumed bandwidth capacity, BW_{ij} for maximum bandwidth capacity, d_{ij}^a as current end-to-end delay in SN edge (i,j), while $f_{i,j}^a$ is current traffic flow defined as the total throughput from SN node i to j. $P^S = \{(i,j)\}$ represents a set of all directed paths connecting all pairs of SN nodes i and j with a set of edges $\{(i,j)\}$. And substrate path $P_{sd} = \{(s,n),...,(k,l),...,(m,d)\} \in P^S$ is an end-to-end path constructed of more than one physical edge, where (s,n) is the first physical edge connecting the source node s to its adjacent node n, (k,l) is an intermediate physical edges, and (m,d) is the last physical edge connecting destination node d to its previous node m. Finally, total end-to-end delay in P_{sd} is the sum of delays of each edge (i,j) between the source node s and the destination d, and is given by $d_{sd}^a = \sum_{\forall\ (i,j) \in E^S} d_{ij}^a$.

B. Virtual Network Model:

Similar to the substrate network, the virtual network is modeled as a weighted directed graph $G^V = (N^V, E^V)$, where u and $v \in N^V$ are virtual nodes, and $(u,v) \in E^V$ is a virtual edge. VNR^r is a virtual network request number r out of R total VNRs. Each virtual node $u \in N^V$ is associated with cpu_u^r, representing the demanded CPU capacity, and each virtual edge (u,v) connecting a pair of virtual nodes u and v is also associated with bw_{uv}^r as the demanded bandwidth capacity. Lastly, d_{uv}^r represents the maximum allowed end-to-end delay demanded by virtual edge (u,v).

C. Power Consumption Model:

A comprehensive survey for state of the art power consumption models were presented in [1]. Accordingly, this paper identified the linear power model introduced by [14], which defined a formula to estimate the power consumption of network's servers PC including its idle power. The model approximated the aggregate behavior of a server system while being active, by measuring the total power consumption of the server PC_i against it's CPU utilization. In addition to idle power, the model includes total power consumed by the server when loaded as shown in Fig.(1). The formula is given as follows:

$\forall\ i \in N^S$

$$PC_i = pw_i^{idle} + [PW_i^{Busy} - pw_i^{idle}] \times \mu_i \tag{1}$$

$$\mu_i = \left(\frac{cpu_i}{CPU_i} \times 100 \right) \tag{2}$$

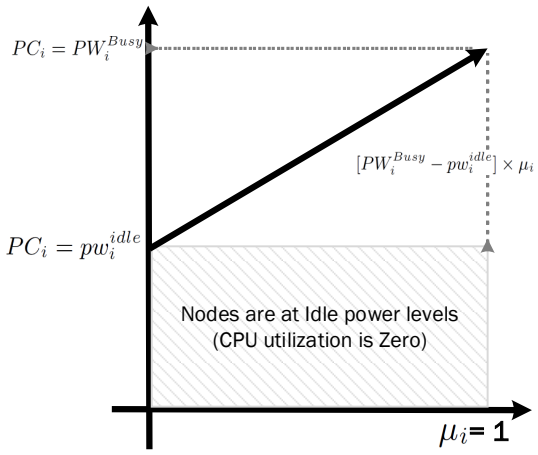


Fig. 1. Power Consumption Model

IV. PROBLEM FORMULATION

VNE problems are traditionally modeled as an optimization problem of objective function with positive integer and linear variables, usually referred to as integer linear programming (ILP) problem. However, optimal solution for the VNE as an ILP problem, implies introducing binary constraints to connect one edge only for each node, then, mapping all virtual nodes and edges on their physical counterparts having enough resources to accommodate their demands. Accordingly, virtual edges associated with bandwidth constraints is usually treated as a commodity between pair of nodes, and therefore, embedding a virtual edge optimally is similar to finding an optimal flow for the commodity in any network model [6],[15],[17].

To formally introduce the VNE problem as an ILP, following paragraphs define and formulate VNR and SN path's segments, in addition to the objective function and its constraints, as follows:

A. Segments formulation:

To solve VNE problem in one stage by fully coordinating nodes and edges embedding together at the same time, segment based formulations for the VNR and SN path are defined and formulated as follows:

Definition: Segment is defined as a set of parameters, grouped as one entity, for VNRs, Seg^r representing demands of virtual nodes and edges in VNR number r, and for a specific SN path, Seg^S represents resources of the physical nodes and edges in the selected SN path.

1) VNR segment formulation (Seg^r): each VNR is reformulated into a segment listing its CPU, BW, and delay demands together as a set. Eq(3) shows general design of Seg^r. Starting by the processing power capacity of its virtual nodes denoted by cpu_u^r for the source node, cpu_w^r for all intermediate nodes, and cpu_v^r for the destination node. Next, the segment lists all virtual edges' resources, including bandwidth capacity per each edge denoted by bw_{uo}^r for the edge connecting source virtual node u to next virtual node o, then it lists all bandwidth capacities for all intermediate virtual edges including virtual edge bw_{wx}^r, connecting intermediate virtual node w to next virtual node x, in addition to virtual path bw_{pv}^r connecting destination virtual node v to its previous virtual node p, and finally Seg^r lists the demanded end-to-end delay d_{uv}^r between the virtual source node u and its virtual destination node v.

$$Seg^r = \{cpu_u^r, .., cpu_w^r, .., cpu_v^r, bw_{uo}^r, .., bw_{wx}^r, .., bw_{pv}^r, d_{uv}^r\} \quad (3)$$

o, p, w, x are virtual nodes $\in$ VNR^r

2) SN path segment formulation (Seg^S): Similarly, P_{sd} segment is shown in eq(4). The segment lists all P_{sd} resources, namely: current available processing power capacities for all nodes in the path, given as cpu_s^a, cpu_k^a, and cpu_d^a for source, all intermediate, and destination physical nodes respectively. In addition, Seg^S lists current available bandwidth capacities for all of its edges starting by bw_{sn}^a, connecting source node s to next physical node n, all edges connecting intermediate nodes including bw_{kl}^a, and bw_{md}^a connecting destination node d to its previous physical node m. Lastly, Seg^S segment lists its end-to-end current delay d_{sd}^a between P_{sd} source and destination nodes.

$$Seg^S = \{cpu_s^a, .., cpu_k^a, .., cpu_d^a, bw_{sn}^a, .., bw_{kl}^a, .., bw_{md}^a, d_{sd}^a\} \quad (4)$$

k, l, n, m are physical nodes $\in P_{sd}^a$

B. Objective function definition and formulation:

Following the same analogy of estimating the power consumption of SN nodes, formula shown in eq.(1) will be applied to formulate the objective function as an ILP optimization problem. The main target is to minimize overall power consumption in the whole substrate network, by putting into sleeping mode all non utilized SN resources that are at idle power consumption, while accommodating VNR's demands. The rational behind that, is that for all SN nodes that are at idle mode, still, they are consuming considerable amount of power, even if their consumed CPU were zero. This is because when nodes are at idle mode, the power consumed by chassis (backplane) and cooling systems could be at least 40% or higher of the total power [19]. Accordingly, setting them into sleeping mode will result on minimizing the total substrate network's power consumption.

1) Objective Function: To make sure that a specific SN node is active and hosting at least one virtual node, variable x_i^{ur} is used in the ILP objective function formulation, which takes a binary value of (1) if substrate node i is active and assigned to host the virtual node u, and (0) otherwise. The objective function is shown in eq.(5) as follows:

$$\forall u \in N^V and \; \forall r \in R$$

$$\min PC_i = \sum_{\forall i \in N^S} (pw_i^{idle} + [PW_i^{Busy} - pw_i^{idle}] \times \mu_i) \times x_i^{ur} \quad (5)$$

C. Constraints definition and formulation:

Objective function solution will be constrained by capacity, flow and domain constraints as shown bellow. However, power consumption constraint was intentionally omitted, since it relies on CPU utilization of each node, nevertheless, it will be satisfied if constraints (6) and (7) were satisfied.

1) Capacity constraints: To ensure current available CPU processing power capacity in substrate node i is greater than or equal to demanded capacity by virtual network node u, constraint (6) is defined as follows:

$$\forall \; i \in N^S \quad cpu_i^a \geq cpu_u^r \quad (6)$$

To ensure total consumed CPU processing power capacity at substrate network node i, is less than or equal to maximum CPU capacity at that SN node, constraint (7) is defined as follows:

$$\forall \; u \leftarrow i \quad \sum_{r \in R} cpu_u^r \leq CPU_i \quad (7)$$

Note: $u \leftarrow i$ means that virtual network node u is hosted at substrate network node i.

To ensure that current available bandwidth capacity on substrate network edge (i, j) is greater than or equal to demanded

bandwidth capacity by virtual network edge (u, v), constraint (8) is defined as follows:

$$\forall\ (i,j) \in P_{sd} \quad bw_{ij}^{a} \geq bw_{uv}^{r} \tag{8}$$

To ensure that total consumed bandwidth capacity in substrate network edge (i, j), is less than or equal to maximum bandwidth capacity at that edge, constraint (9) is defined as follows:

$$\forall\ (u,v) \leftarrow (i,j) \quad \sum_{r \in R} bw_{uv}^{r} \leq BW_{ij} \tag{9}$$

Note: $(u, v) \leftarrow (i, j)$ means that virtual network edge (u, v) is embedded on the substrate network edge (i, j).

To ensure that current end-to-end delay in substrate network path P_{sd} is less than or equal to maximum allowed delay d_{uv}^{r} by VNR^{r}, constraint (10) is defined as follows:

$$d_{sd}^{a} \leq d_{uv}^{r} \tag{10}$$

2) Flow constraints: To ensure that a flow getting in a substrate node must go out, the following constraints has to be satisfied:

$$\sum_{\forall n \in N^{S}} f_{sn}^{a} - \sum_{\forall n \in N^{S}} f_{ns}^{a} = bw_{uo}^{r} \tag{11}$$

$$\sum_{\forall m \in N^{S}} f_{dm}^{a} - \sum_{\forall m \in N^{S}} f_{md}^{a} = -bw_{pv}^{r} \tag{12}$$

$$\sum_{\forall k,l \in N^{S}} f_{kl}^{a} = \sum_{\forall k,l \in N^{S}} f_{lk}^{a} \tag{13}$$

Constraint (11) ensures that the total flow getting out of source node s is the demanded flow bw_{uo}^{r}, while constraint (12) ensures that total flow getting into destination node d is the forwarded flow bw_{pv}^{r}, and constraint (13) ensures that all demanded flow is transfered from source to destination node, and nothing remains at any intermediate node within SN path P_{sd}.

3) Domain constraints: To solve the problem as ILP, constraint (14) is defined as follows:

$$\forall\ i \in N^{S}\ x_{i}^{ur} \in \{0, 1\} \tag{14}$$

To ensure each virtual node is mapped only to one substrate node, constraint (15) is defined as follows:

$$\forall\ u \in N^{V} \sum_{\forall i \in N^{S}} x_{i}^{ur} = 1, \tag{15}$$

To ensure virtual nodes from the same VNR are mapped to different substrate nodes, constraint (16) is defined as follows::

$$\forall\ i \in N^{S} \sum_{\forall u \in N^{V}} x_{i}^{ur} \leq 1, \tag{16}$$

V. HEURISTIC DESIGN

Optimal solution for VNE is known to be NP-Hard and computationally intractable, since it can be reduced to multi-way separator problem, which is NP-Hard by itself [7]. As a summary, [18] listed some of the main reasons highlighting why solving VNEs is challenging, such as: randomness of the arrival of VNRs depending on users' demands, topology and resources constraints by each VNR, and limited SN resources. However, the virtual edges embedding problem is what makes the VNE problem exceptionally an NP-hard, because it could be mapped to one or more physical edges that are not necessarily physically connected. Even for offline VNE case, given that all nodes were embedded, still virtual edge embedding stage can be reduced to the unsplittable flow problem, which is NP-hard [15],[16]. Consequently, solving VNE problem in polynomial time is not possible.

Therefore, majority of VNE approaches followed heuristic or meta-heuristic algorithms to solve VNE optimization problems in a reasonable polynomial time [5]. For example, one of the most referenced VNE heuristic approaches is the algorithm presented by [7]. It coordinates node and edge embedding, through mapping virtual nodes onto substrate nodes in a way that facilitates mapping of virtual edges. Nevertheless, the authors performed VNM and VEM in two interrelated stages. First they designed a node embedding algorithm to embed the virtual nodes on a suitable physical nodes, which could be separated a part from each other. Second, once node mapping was successful, they triggered another algorithm to embed the associated virtual edges on substrate paths, noting that it mostly would include hidden nodes to be used as hops. However, other ideas could be explored to better coordinate embedding VNM and VEM stages, and at the same time avoid including non necessary hidden hops and edges beyond VNRs needs.

Therefore, this paper proposed the PaCoVNE algorithm as a new heuristic methodology to solve VNE optimization problem more efficiently. Its main strength, is that it coordinates node and edge embedding in one step, based on matching each element in VNR^{r} segment, Seg^{r}, against their counterparts in the SN path's segment, Seg^{S}, considering the following four constrains, namely: CPU and BW capacity constraints, in addition to power consumption and end-to-end delay constraints.

A. Heuristic code explained:

Pseudo-code for PaCoVNE heuristic is shown in Algorithm 1 bellow, and is explained by the following main four steps:

1) ***Initialization****:* it starts by generating SN topology, lists all its possible paths, and categorizes them into types according to number of nodes and edges per each SN path. Notice that, number of lists and paths per list varies depending on the size and topology of SN. Since SN topology is physically fixed in real life, the main elements formulating any SN path (number and connectivity of SN nodes and edges) are also fixed and does not change, but only their capacities varies due to consumption. Therefore, to avoid searching for SN paths while VNE algorithm is running, and in contrary to most available heuristics in literature, this paper performs the initialization step in advance ahead of VNRs' arrival. This is one advantage behind PaCoVNE's speed of performing VNE in real-time, given it mainly focuses on the actual mapping process itself. To facilitate recalling a specific list of SN paths by PaCoVNE algorithm whenever it receives a new VNR, these lists will be saved and categorized per path type in a data base repository, including number of nodes, edges, and connectivities for each path.

2) ***Segmentation and ranking****:* this is the differentiating aspect of PaCoVNE heuristic compared to others, mainly because it facilitates accommodating VNRs one by one and embed their nodes and edges in one step and in full coordination between VNM and VEM. First the heuristic formulates VNR^{r} segment Seg^{r}. Then, to formulate the candidate SN path segment Seg^{S}, it recalls the appropriate list of SN paths that has similar number of nodes and edges as that of VNR^{r}. Next, it ranks them according to their CPU utilization, and ends by formulating SN segment for the top ranked path.

3) ***Embedding decision****:* compares each element in the SN segment Seg^{S} to its counterpart in the Seg^{r}, one-by-one. Accordingly, if SN segment has enough resources to accommodate all demands of VNR^{r}, PaCoVNE selects the path of SN segment Seg^{S} to host VNR^{r}. Decision matrix for the embedding process is shown in eq.(17) bellow:

$$if\ cpu_{i}^{a} - cpu_{u}^{r} \geq 0\ and$$

$$if\ bw_{ij}^{a} - bw_{wx}^{r} \geq 0\ and$$

$$if\ d_{sd}^{a} \leq d_{uv}^{r} \tag{17}$$

*4) **Updating**:* once a successful embedding occurs, the heuristic updates all changed SN resources and moves to next VNR. However, in case that SN segment does not have enough resources to accommodate VNR demands, the heuristic jumps to the next ranked path, and follow on from step 5. This process keeps on going until no more VNRs to be handled.

Algorithm-1, PaCoVNE Pseudo-Code

1) Input: G^V.
2) **for** each VNR$^r \in R$ **do**
   -Formulate VNR^r parameters into segment Seg^r according to eq.(3).
3) For the set of all saved SN paths P^S:
   List all SN paths matching VNR^r size.
   Rank them in descending order based on μ_i according to eq.(2)
4) For top ranked SN path P_{sd}, formulate its segment Seg^s according to eq.(4).
5) **Compare** Seg^r against Seg^S
   Check for CPU, BW and Delay constraints according to eq.(17).
6) **If** satisfied,
   embed VNR^r on P_{sd}.
   else go to next ranked SN path, step-4.
7) **for** all SN nodes and edges **do**
   Update CPU and BW resources.
   Remove the embedded VNR^r from VNRs list.
8) **for** idle SN nodes **do**
   Turn-off to save power.
9) Evaluate Metrics.
10) **If** VNRs list not empty, **go** to next VNR step-2.

B. PaCoVNE Computational Time Complexity:

In this paper, regardless the number of VNRs and based on the adjacency matrix of SN, searching and listing all types of paths will consume $O(|N^S| + |E^S|)$ processing time, depending on total number of nodes N and edges E formulating the SN [17]. This step is performed and saved only once before the arrival of any VNR. Therefore, it will not have any impact on the real computational time complexity of the VNE process.

However, the actual VNE process starts when the first VNR arrives at the SN. Therefore, in order to evaluate computational time complexity of PaCoVNE at worst case, the focal computational component of the heuristic is determined based on the time consumed while sorting all listed SN paths that has the same number of nodes and edges as the VNR^r. The larger the number of listed paths, the more computational time is consumed by the working machine.

Accordingly, for each VNR^r, the PaCoVNE adopted (Bubble Sort) algorithm to rank all SN paths in descending order [17]. Thus, at the worst case, the PaCoVNE algorithm will have a quadratic computational time complexity in the order of $O(n^2)$, where n is number of paths.

C. Illustrative Example:

A detailed example to explain the proposed heuristic is shown in fig.(2). It applies PaCoVNE on a SN of four nodes as shown in stage A, then it evaluates how to accommodate VNR^1, by sorting all listed SN paths based on the total sum of CPU utilizations 'μ' for each path. As shown in stage B, the PaCoVNE concludes by embedding VNR^1 on nodes 2 and 3, along path P_{23}, which had enough resources to accommodate its demands. In this case, the heuristic managed to save 21% of the total consumed power in the whole SN, by turning-off nodes 0 and 1, since they were idle. Stage C introduced VNR^2, the PaCoVNE decides that even though P_{23} is still the top ranked path, based on its CPU utilization, but since it does not have enough BW resources to accommodate the demanded BW by VNR^2, it jumps to next ranked SN path, P_{02}, which satisfies all demands of VNR^2.

TABLE I
SIMULATION SETTINGS FOR OFF-LINE HOMOGENEOUS

Parameter	SN	VNR
Nodes	50	15
$CPU\ max$	100	2.1
$BW\ max$	100	2.3
$Delay\ max$	250	$100 - 250$
PW^{Busy}	524	
pw^{idle}	$PW^{Busy} * 0.4$	
Loads	$0.2 - 0.9$	
Runs/load	50	
α	0.6	
β	0.23	
p_{wax}	0.2	

Therefore, PaCoVNE assigns P_{02} to accommodate VNR^2, then it keeps node 1 turned-off, since its the only idle node, resulting on saving 12% of the total consumed power by the whole SN.

VI. PERFORMANCE EVALUATION

In this paper, off-line version of PaCoVNE heuristic was tested using homogeneous and heterogeneous settings, once with end-to-end delay, and another time without it. The homogeneous version was compared to one of the most referenced heuristics, the energy aware relocating algorithm 'OCA/EA-RH' developed by [9]. Then, for the heterogeneous scenario, PaCoVNE was compared to its homogeneous version.

A. Simulation Settings:

For the off-line homogeneous scenario without delay, PaCoVNE was compared to OCA/EA-RH heuristic, which only used VNRs of 15 nodes; denoted as ($VNRs_{15}$), and therefore, the same simulation settings will be applied for PaCoVNE as well [9]. Specifically, the SN will handle a set of 80 VNRs, for different average loads, denoted by $\rho \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The value of each load ρ, reflects a ratio between VNRs demands to SN capacities, and therefore, loading the SN with X%, means this is the average of loading each node by X% load as well. The values of SN's nodes cpu_i^a and bw_{ij}^a resources were set to uniformly distributed values equal to 100. For each VNR, cpu_w^r and bw_{wx}^r values were estimated from the average embedding cost figure of [9], and are given as follows: $cpu_w^r = 2.1$ and $bw_{wx}^r = 2.3$. Finally, maximum power consumption by each SN node PW_i^{Busy} was set to 524 watts, while its idle power pw_i^{idle} was set as (PW_i^{Busy} * 0.4) [19]. For SN edges, end-to-end delay d_{ij}^a, was set equal to $250ms$ as a limit [20],[21]. While virtual network delays d_{wx}^r, was selected pseudorandomly between $100 - 250ms$. Table (1) summarizes all simulation settings to compare PaCoVNE against OCA/EA-RH for the off-line and Homogeneous scenario.

SN topologies were generated as directed graphs, through Waxman algorithm according to the following parameters: $\alpha = 0.6$, $\beta = 0.23$, and mean probability of creating an edge between any two SN nodes, denoted as p_{wax} was set equal to 0.2. Important to notice that, these parameters differ from what [9] used, since the aforementioned parameters will provide average edges at each SN node of 6, instead of 12 as used by [9]. This caused PaCoVNE heuristic to rank much less number of paths, yet, it produced better results compared to OCA/RA-EH. To overcome the probabilistic nature of Waxman topology generation, the set of 80 VNRs were run for 50 times per each ρ load value.

B. Heuristic work-flow:

Initialization: based on the SN adjacency matrix, the heuristic lists all SN paths of 15 nodes, denoted as $P_{15} \in P^S$, this is only performed once and saved at the beginning. These paths can then be used for any number of $VNRs_{15}$. This is important, since

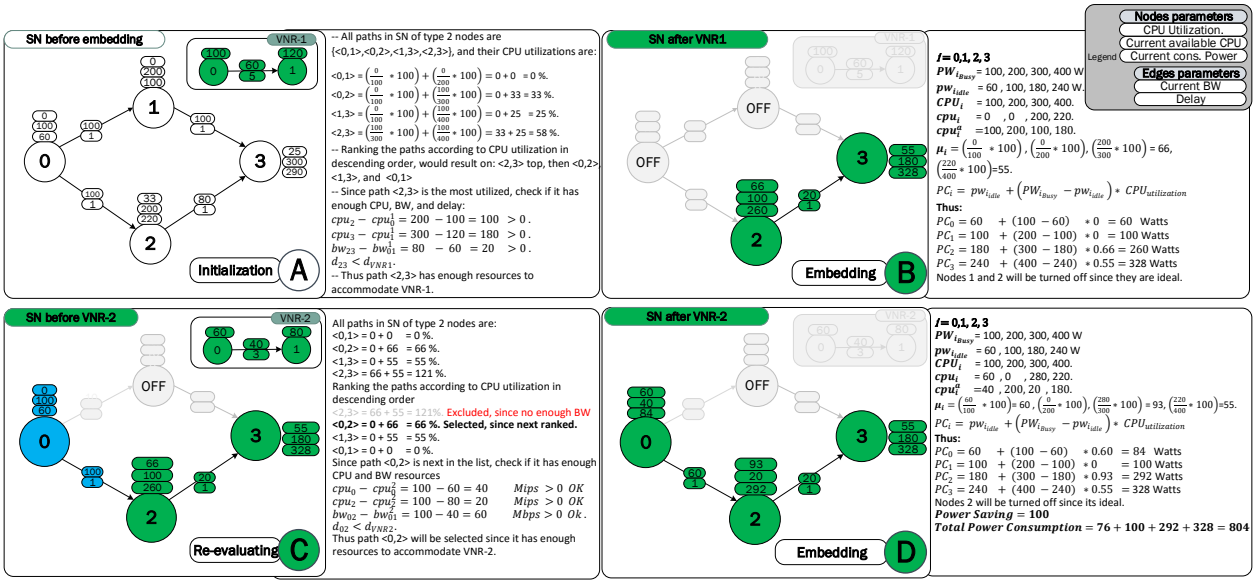


Fig. 2. Numerical example showing basics of PaCoVNE

the heuristic will focus on the actual embedding process itself, and not on searching for the best path at the arrival of each new VNR, which saved PaCoVNE's run time considerably. Moreover, its important to mention that PaCoVNE heuristic can generate all types of SN paths at the beginning, and accordingly, it can handle any type of VNRs regardless of how many nodes they may contain.

VNR segment formulation: Since OCA/EA-RH heuristic used VNRs$_{15}$, then PaCoVNE formulates Seg^{r15} as defined in eq.(3).

SN path segment: the heuristic selects the path of highest μ, denoted as $P^a \in P_{15}$, and formulates its segment Seg^{a15} as defined in eq.(4).

Ranking: For each path $P^a \in P_{15}$, PaCoVNE calculates and sums its μs, then it ranks the paths based on the value of its μ from highest to lowest.

Embedding: The algorithm compares both segments according to eq.(17), if the conditions are satisfied, then it embeds VNR^{r15} on SN path P^a.

Turning off idle SN nodes: Once VNR^{r15} is embedded successfully, PaCoVNE identifies all idle SN nodes and turns them off to save power consumption. Next it updates all SN elements based on that.

C. Evaluation Metrics:

The PaCoVNE heuristic will be evaluated according to following metrics:

- *Average Total power consumption, PW:* defined as the total power consumed by all SN nodes after each VNR embedding, and averaged over the total number of VNRs R [9],

$\forall \rho \in$ *Loads*,

$$PW = \frac{1}{R}\Big(\sum_{\forall r \in R}\sum_{\forall i \in N^S} PC_i\Big) \tag{18}$$

- *Average Saved Power, PS:* the amount of saved power after embedding each VNR, using the proposed power reduction strategy. Calculated by subtracting total power consumed by all SN nodes without power reduction strategy PW^-, from total power consumed by all active SN nodes after applying power reduction strategy PW^+. The results will be averaged over the total number of VNRs R.

$\forall \rho \in$ *Loads*,

$$PS = \frac{1}{R}\sum_{\forall r \in R}\Big(\sum_{\forall i \in N^S} PW^- - \sum_{\forall i \in N^S} PW^+\Big) \tag{19}$$

- *Average Acceptance Ratio, AR:* is a ratio to represent how PaCoVNE algorithm is performing, calculated for each load value ρ, by dividing number of successfully embedded VNRs by total number of VNRs R [7],[9].

$\forall \rho \in$ *Loads*,

$$AR = \frac{1}{R}\ Total\ Number\ of\ Embdded\ VNRs * 100 \tag{20}$$

- *Average Cost of embedding VNRs, EC:* is the sum of total consumed SN resources CPU and BW while embedding each VNR. Tuning parameters to represent relative costs per each SN resource, denoted as α for SN nodes' cost, and β for SN edges, were both set equal to one [7],[9].

$\forall \rho \in$ *Loads*,

$$EC = \frac{1}{R}\sum_{\forall r \in R}\Big(\sum_{\forall (i,j) \in E^S}(\beta * bw_{ij}) + \sum_{\forall i \in N^S}(\alpha * cpu_i)\Big) \tag{21}$$

- *Average CPU utilization, CPU_{util}:* it represents SN nodes' utilization trend after all simulation iterations. Its defined as ratio between consumed CPU cpu_i, and maximum CPU resources, averaged overall VNRs for each load ρ [7].

$\forall \rho \in$ *Loads*,

$$CPU_{util} = \frac{1}{R}\sum_{\forall r \in R}\Big(\sum_{\forall i \in N^S}\frac{(CPU_i - cpu_i^a)}{CPU_i} * 100\Big) \tag{22}$$

- *Average BW utilization, BW_{util}:* it represents utilization of SN edges after all simulation iterations. And is defined as ratio between consumed bw_{ij}, and the maximum BW, averaged overall VNRs for each load ρ [7].

$\forall \rho \in$ *Loads*,

$$BW_{util} = \frac{1}{R}\sum_{\forall r \in R}\Big(\sum_{\forall (i,j) \in E^S}\frac{(BW_{ij} - bw_{ij}^a)}{BW_{ij}} * 100\Big) \tag{23}$$

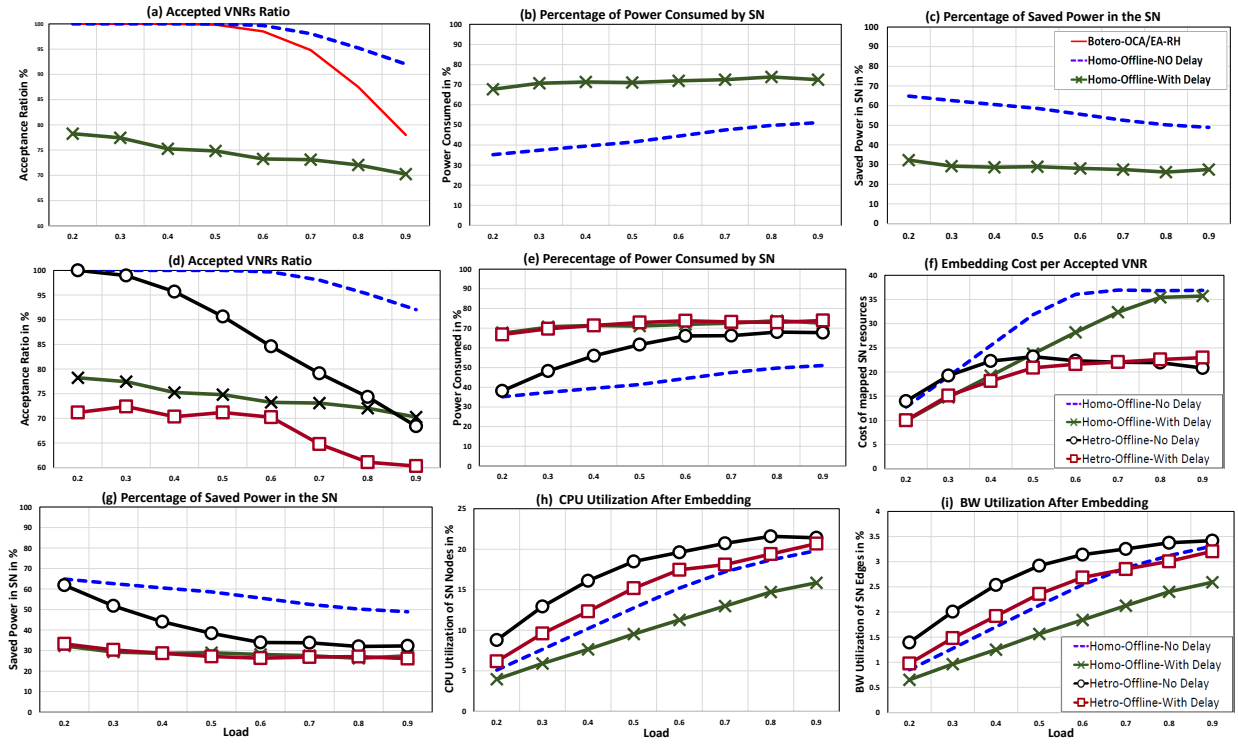


Fig. 3. Comparison Results of PaCoVNE Homogeneous against OCA/EA-RH and PaCoVNE Heterogeneous

VII. RESULTS AND DISCUSSION

A. *Off-line Homogeneous scenario:*

Simulation results in fig.(3a) shows that PaCoVNE performed similar or better than EA-RH in terms of acceptance ratio for lower loads, and is much better for higher loads, thanks to the one stage, full coordinated embedding and segment formulation of PaCoVNE, which makes sure to allocate virtual nodes and their associated edges together, and at the same time, thus, increasing the acceptance ratio. In comparison, the OCA/EA-RH relocates least stressed virtual nodes and their associated edges to other suitable active SN nodes that are more stressed, using the cost-based VNE approach of [7]. Then in a separate phase, OCA/EA-RH relocates least stressed edges to shortest energy path.

However, in terms of power consumption at SN nodes, PaCoVNE can not be compared to OCA/EA-RH, since both algorithms used different formulas to calculate the consumed power per each SN node. Nevertheless, fig.(3b) shows that SN's power consumption is still high, giving that PaCoVNE model includes idle power in addition to power consumption when SN nodes were loaded according to their CPU utilization. This entails the importance of considering idle power as a main component for increasing power consumption of SN's nodes, even if they do not process any data.

Moreover, regarding saved power results shown in fig.(3c) clarifies that, when the load was 0.2 PaCoVNE managed to save 65% of SN's total power, and when the load was much increased to 0.9 it saved 49%, implying that, in a range of loads between 0.2 to 0.9, PaCoVNE would save in average 57% of SN's total power consumption, by putting idle nodes into sleeping mode, while maintaining high VNE acceptance ratios across almost all loads. These results highlights the benefits of using PaCoVNE's new segmentation strategy to fully coordinate VNE, also pinpoints the obvious impact of idle power consumption on the overall SN's power consumption, thus, reducing it would ultimately reduce SN costs. Indeed, important to point out that in real life conditions, putting idle nodes into sleeping mode as a power reduction strategy, could affect service maintainability of SN, especially considering on-line scenarios. Therefore, other strategies could be explored as well.

TABLE II
SIMULATION SETTINGS FOR OFF-LINE HETEROGENEOUS SCENARIO

Parameter	SN	VNR
Nodes	50	15
CPU_{max}	$Random$ 40 – 100	$Random$ 1.5 - 2.1
BW_{max}	$Random$ 40 – 100	$Random$ 1.6 - 2.3
$Delay_{max}$	Random 100 – 250	Random 100 – 250
PW_{Busy}	524	
pw_{Idle}	PW_{Busy} * 0.4	

In the case of including end-to-end delay, fig.(3a, 3b, and 3c) shows the obvious impact of end-to-end delay. In comparison to PaCoVNE homogeneous without delay, acceptance ratio was degraded by 24% in average for all loads, increased power consumption by 57%, and reduced saved power by 51%. These results implies the significance of including end-to-end delay as a main constraint to embed VNRs, and how negatively it would impact the whole VNE process.

B. *Off-line Homogeneous against Heterogeneous:*

The rational behind comparing PaCoVNE using homogeneous to heterogeneous configuration is to give some insights about how PaCoVNE would behave on semi-real life conditions, where SN resources usually differ in size and capacity, in addition to including end-to-end delay. Table-2 summarizes the heterogeneous simulation settings.

Fig.(3d, 3g, 3h, and 3i) shows simulation results considering heterogeneous conditions, indicating the out-performance of homogeneous-PaCoVNE in terms of acceptance ratio, saved power, CPU and BW utilizations with and without end-to-end delay. In terms of power consumption and embedding cost as shown in fig.(3e and 3f), heterogeneous-PaCoVNE performed

much worse than homogeneous in both of them, mainly due to PaCoVNE's rapid tendency to utilize the SN resources. This is clearly translated into less accepted VNRs as loads increases. In addition to that, almost same conclusion can be deduced when end-to-end delay was applied, showing that the resultant metrics for the heterogeneous performed even much worse than the homogeneous across all metrics and loads, doubling down the significance of including end-to-end delay as a main VNE constraint.

VIII. CONCLUSIONS

This paper introduced the PaCoVNE heuristic, which performed VNE in a more efficient methodology than other algorithms in the literature. Performance of the heuristic was evaluated using homogeneous and heterogeneous configurations, once without considering end-to-end delay, as a constraint, and also when delay was included. Simulation results showed that, for the homogeneous scenario and when end-to-end delay was not included, the new heuristic managed to save considerable amount of substrate network's power consumption by 57% in average, for a range of loads between 0.2 to 0.9, through putting idle nodes into sleeping mode, while maintaining high VNE acceptance ratios, thanks to the new coordinated VNE approach. However, when end-to-end delay was factored in, PaCoVNE performance resulted on both, less saved power and acceptance ratio in comparison to homogeneous without delay. Suggesting that, introducing end-to-end delay, as in the real world and as a major constraint, had clear impact on the whole VNE process. On the other hand, when PaCoVNE in homogeneous setting was compared to heterogeneous version, the heuristic's performance degraded across all evaluation metrics, and specifically when end-to-end delay was included. Thus, doubling on the critical importance of considering delay as a major guiding principle to perform the VNE process in acceptable levels that could be applicable to real world applications.

The following points are the main outcomes of this paper:

1) PaCoVNE provided a new and better strategy to fully coordinated VNM and VEM simultaneously and in one step, thanks to the segmentation design concept.
2) The new strategy resulted on clear enhancements on VNE acceptance ratio in comparison to literature, fundamentally due to the advantageous of one stage embedding.
3) Most significant, was PaCoVNE's capabilities to save a very considerable amount of total power consumption of SN elements. Mainly due to the very precise embeddings, which allowed for efficiently distributing VNRs on the most powerful SN nodes, and consequentially, enabled a better identification methodology for the more idle SN nodes to turn them off.
4) However, when end-to-end delay was included, it significantly impacted VNE process, as reflected by lower acceptance ratios. Suggesting the importance of including end-to-end delay as a major VNE constraint.
5) Depending on the size of the VNR, the time consumed by the PaCoVNE heuristic to embed the VNR successfully varies significantly. The larger the number of nodes per a VNR, the more time it takes to embed it. This suggests that for large networks, the PaCoVNE should partition the VNRs and physical paths into smaller portions to speed up the embedding time. Thus, even thought the solution for some partitions may be sufficient, but it may not be as sufficient when aggregating the solutions for all partitions of the selected substrate network path.

As a future work, the authors are planning to extend the application of PaCoVNE to work for the online scenarios. Also, in addition to CPU utilizations to rank SN paths, other criterion can be studied, such as: edges utilizations, or their propagation delay. Moreover, other non linear parameters can be considered to evaluate the performance of the PaCoVNE, namely, what would be the impact of jitter, packet-loss, and grade of service on the VNE process given the segmentation strategy used by the PaCoVNE.

IX. ACKNOWLEDGMENT

This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under project TEC2016-76795- C6-1-R and AEI/FEDER, UE.

REFERENCES

[1] M. Dayarathna, Y. Wen and R. Fan, "Data Center Energy Consumption Modeling: A Survey," in IEEE Communications Surveys and Tutorials, vol. 18, no. 1, pp. 732-794, Firstquarter 2016.
[2] ESTI, Network Functions Virtualisation, Introductory White Paper, October,2012.
[3] 5G PPP Architecture Working Group, "View on 5G Architecture," Version 1.0, 2016.
[4] Rachid El Hattachi, and Javan Erfanian, NGMN 5G White Paper, 2015.
[5] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer and X. Hesselbach, "Virtual Network Embedding: A Survey," in IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 1888-1906, 2013.
[6] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: Substrate support for path splitting and migration," ACM SIGCOMM CCR, vol. 38, noi. 2, pp. 17-29, 2008.
[7] M. Chowdhury, M. R. Rahman and R. Boutaba, "ViNEYard: Virtual Network Embedding Algorithms With Coordinated Node and Link Mapping," in IEEE/ACM Transactions on Networking, vol. 20, no. 1, pp. 206-219, Feb. 2012.
[8] J. Botero, X. Hesselbach, M. Duelli, D. Schlosser, A. Fischer, and H. de Meer, "Energy efficient virtual network embedding," Communications Letters, IEEE, vol. 16, no. 5, pp. 756-759, 2012.
[9] J. Botero, X. Hesselbach, "Greener networking in a network virtualization environment," Computer Networks, vol 57, issue 9, pp. 20121-2039, 2013.
[10] Sen Su, Zhongbao Zhang, Alex X. Liu, Xiang Cheng, Yiwen Wang, and Xinchao Zhao, "Energy-Aware Virtual Network Embedding," IEEE/ACM Transactions on Networking, vol. 22, no. 5, pp. 1607-1620, 2014.
[11] Nizar Triki, Nadjia Kara, May El Barachi, Souad Hadjres, "A green energy-aware hybrid virtual network embedding," Computer networks, Vol. 91, pp. 712-737, 2015.
[12] Leonard Nonde, Taisir E. H. El-Gorashi, and Jaafar M. H. Elmirghani, "Energy Efficient Virtual Network Embedding for Cloud Networks," Journal of Lightwave Technology, Vol. 33, No. 9, pp. 1828-1849, 2015.
[13] Xiaohua Chen, Chunzhi Li, and Yunliang Jiang, "A feedback control approach for energy efficient virtual network embedding," Computer Communications, Vol. 80, pp. 16-32, 2016.
[14] X. Fan, W. D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in Pro. 34th Annu. ISCA, pp. 13-23, 2007.
[15] Bradley, Hax and Magnanti, "Applied Mathematical Programming," Chapters-8 and 9, Addison-Wesley, 1977.
[16] S. G. Kolliopoulos and C. Stein, "Improved approximation algorithms for unsplittable flow problems," Proceedings 38th Annual Symposium on Foundations of Computer Science, Miami Beach, FL, pp. 426-436, 1997.
[17] J. Kleinberg and E. Tardos, "Algorithms Design," Addison-Wesley, 2009.
[18] Ilhem Fajjari, "Resource Allocation Algorithms for Virtual networks within Cloud Backbone Network," PhD Thesis, Pierre et Marie Curie University, France, 2012.
[19] Telecommunications Infrastructure Standard for Data Centers. http://www.tia-942.org/.
[20] ITU, Draft new Report ITU-R M. [IMT-2020.TECH PERF REQ], Minimum requirements related to technical performance for IMT-2020 radio interface(s)", ITU, Document 5.40-E, 22 February, 2017.
[21] G. Almes, S. Kalidindi, M. Zekauskas, amd A. Morton, A One-Way Delay Metric for IP Performance Metric (IPPM), IETF, RFC-7679, 2016.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6502

Definición de Testbeds Virtualizados Utilizando Perfiles de Actividad de Red

David Muelas, Javier Ramos, Jorge E. López de Vergara
HPCN, Departamento de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Francisco Tomás y Valiente, 11, 28049 Madrid, España
Email: {dav.muelas, javier.ramos, jorge.lopez_vergara}@uam.es

***Resumen*—Un problema recurrente para los profesionales de la Ingeniería Telemática es la escasez de despliegues de tecnologías emergentes y las restricciones de acceso a redes operativas. Por ello, en este trabajo presentamos un método para la generación automática de carga siguiendo perfiles de actividad que facilita la replicación del comportamiento típico de una red. Este método se basa en un nodo de control que configura agentes de generación de tráfico, para aprovechar las capacidades de las plataformas de virtualización de red. Evaluamos esta propuesta en un caso de estudio que considera un despliegue de Voz sobre IP (*Voice over IP*, VoIP) en un servidor de propósito general, usando Mininet como entorno de virtualización ligera. Los resultados muestran que el método propuesto replica fidedignamente la dinámica de red especificada, y que los recursos físicos consumidos permiten su uso en equipamiento de coste reducido.**

***Palabras Clave*—redes virtualizadas, redes definidas por software, experimentación, testbeds, evaluación de prestaciones, VoIP, Mininet.**

I. INTRODUCCIÓN

Los avances en las tecnologías de telecomunicaciones han permitido una evolución muy rápida de los elementos y comportamientos presentes en las redes actuales [1]. Como consecuencia, las herramientas clásicas de simulación, como OMNeT++ o ns-3, no son particularmente aptas para la innovación en entornos emergentes, como los orientados a la Internet de las Cosas (*Internet of Things*, IoT) o a las Redes Definidas por Software (*Software Defined Networks*, SDN) [2]. Por ello, desde el punto de vista de la evaluación de nuevos protocolos, metodologías y elementos de red, se hace necesario desarrollar nuevos enfoques que se adapten a estos entornos reales.

Las oportunidades que ofrece la Virtualización de Funciones de Red (*Network Function Virtualization*, NFV) [3], [4] ha atraído la atención de la comunidad dedicada a la investigación en Ingeniería Telemática por la flexibilidad que ofrecen los despliegues de red virtualizados. No obstante, pese a la profusión de soluciones para encarar diversos retos de la gestión y operación de redes de comunicaciones, su aplicación a la definición de *testbeds* realistas y de bajo coste no ha sido particularmente amplia. Este tipo de plataformas resultan totalmente necesarias a la hora de garantizar tanto la repetibilidad de los experimentos como la disponibilidad de infraestructura de pruebas para la investigación de entornos emergentes, tales como las redes inalámbricas de sensores [5].

Además, el acceso a entornos de red operativos es en muchos casos muy restringido, debido a la importancia que tienen estas infraestructuras. Esto dificulta la evaluación de nuevas soluciones con comportamientos que representen la realidad, de forma que muchas veces la detección de problemas que no aparecen en evaluaciones sintéticas es inviable —por ejemplo, los errores de marcado de tiempo reportados en [6] que afectan a motores de captura de altas prestaciones aparecen cuando la tasa del tráfico disminuye.

Por todo ello, en este trabajo exploramos la definición por software de *testbeds* virtualizados con el fin de facilitar la experimentación en entornos emergentes de red. Para ello, definimos un método para automatizar la acción de agentes dentro de una red de comunicaciones, replicando una dinámica de red determinada. Además, evaluamos Mininet [7], [8] como entorno de virtualización ligera, proporcionando una caracterización de métricas de rendimiento, lo cual define los escenarios y situaciones donde esta herramienta puede ser usada. Nuestro trabajo sigue la filosofía con la que fue desarrollada Mininet, que surgió como una herramienta que facilitara la experimentación en entornos emergentes de red [9], [10], [11].

Para responder a la cuestión de si es posible emular despliegues que representen un amplio espectro de escenario reales en condiciones controladas y con un coste reducido, en este trabajo *(i)* presentamos una metodología y arquitectura para la generación automática de carga de red siguiendo perfiles de actividad, para facilitar la definición de *testbeds* virtualizados; *(ii)* describimos la

implementación de esta arquitectura en Mininet, para reducir el coste de aplicación y maximizar su versatilidad; y *(iii)* comprobamos la viabilidad de la solución tanto en términos del consumo de recursos físicos en diversas topologías con baja y alta actividad de red, como de las características de la carga generada.

El resto del artículo se estructura del siguiente modo. La Sección II recopila diversos trabajos relacionados con el nuestro, motivando las características y decisiones de diseño de nuestra propuesta. Posteriormente, en la Sección III se describe nuestro método de generación de carga, primero en su versión más general y luego en un caso particular aplicado al estudio de despliegues de Voz sobre IP (*Voice over IP*, VoIP). En la Sección IV se incluyen los resultados obtenidos en un entorno virtualizado con Mininet, mostrando la viabilidad de este enfoque para la definición de *testbeds*. Finalmente, la Sección V recopila las principales conclusiones de este trabajo, y plantea las líneas de trabajo futuro que estamos empezando a explorar.

II. TRABAJO RELACIONADO

En esta sección presentamos una selección de resultados previos que motivan nuestro trabajo. Todos ellos ilustran la necesidad de mejorar la definición de *testbeds* virtualizados, por la importancia que tienen durante la evaluación de nuevas herramientas y métodos en un amplio abanico de entornos de red emergentes.

En [12], [13], [14] se presentan metodologías para la generación de tráfico sintético según los parámetros extraídos de tráfico real. Particularmente, los autores de [12] definen un método para extraer las características del tráfico de red, y generar conexiones que presentan las mismas características estadísticas. Por su parte, en [13], se ofrece una revisión más exhaustiva de la literatura relativa a generadores de carga de red, motivando las decisiones de diseño de una arquitectura que comparte ciertos rasgos con la solución que presentamos en este trabajo, aunque se restringe a la generación de flujos de datos realistas. Más recientemente, [14] incluye un análisis más próximo al nuestro, analizando las características de flujos de datos procedentes de fuentes específicas.

La evaluación de estas soluciones permite comprobar que las características del tráfico sintético son indistinguibles estadísticamente hablando de las del tráfico original. No obstante, nuestra propuesta se centra en replicar la dinámica de la actividad de la red en términos de conexiones activas, con el fin de definir comportamientos complejos. Este aspecto facilita la evaluación de nuevos protocolos y herramientas, y además permite que el tráfico sintético pueda ser generado en base a la actividad observada de distintas aplicaciones reales. Por ello, nuestra propuesta se podría utilizar para extender esos trabajos, con el fin de enriquecer el comportamiento dinámico de la carga generada.

Por su parte, en [15], [16] se discute la adecuación de la virtualización de elementos de red para la evaluación de aplicaciones multimedia. Estos trabajos se pueden ver como un antecedente directo de nuestra solución, aunque las características tecnológicas de las soluciones planteadas en ellos están lejos de los últimos desarrollos de virtualización de redes. Estos factores constriñen, por tanto, la aplicabilidad del sistema propuesto en la actualidad, aunque es un trabajo que motiva los desarrollos en la línea de la solución que proponemos.

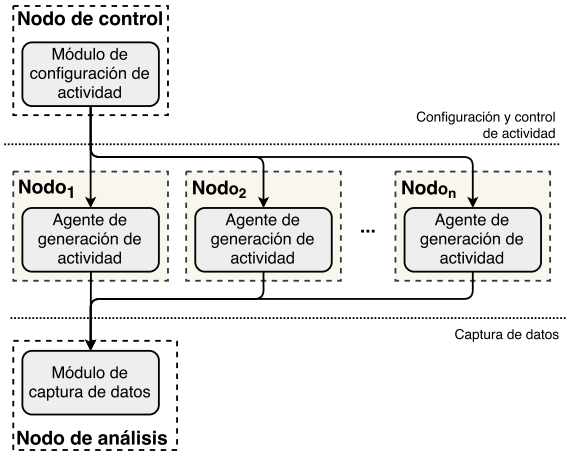


Fig. 1. Esquema de la arquitectura propuesta.

El uso de Mininet como plataforma de emulación de redes para validación y experimentación es una de las motivaciones iniciales de su desarrollo. Experiencias previas, como la recogida en [17], ya muestran su idoneidad para el estudio de redes en operación, siempre y cuando se acepte cierta pérdida de precisión en los resultados obtenidos [1], [18]. Por otro lado, trabajos como [10], [11] muestran la utilidad de esta herramienta a la hora de experimentar sobre entornos emergentes de red. Todos estos resultados previos motivan la exploración de las posibilidades que ofrecen los *testbeds* virtualizados para facilitar el acceso a entornos controlados y que representen de manera fidedigna los futuros despliegues de red.

III. GENERACIÓN DE CARGA A PARTIR LA DINÁMICA DE LA ACTIVIDAD

La metodología que proponemos para generar de forma automática carga de red realista se basa en el uso de agentes que emulen la actividad de usuarios que son controlados para replicar el comportamiento agregado de una red. La carga generada consiste en paquetes de red con el fin de que la solución sea útil en el mayor número de casos posible.

A. Definición general del método

La Fig. 1 representa la arquitectura de nuestra solución, formada por un nodo de control que configura y que activa agentes de generación de actividad (tráfico) en el resto de nodos. Finalmente, para observar el comportamiento global agregado, se incluye otro nodo que captura y analiza el tráfico generado. Este diseño permite que asumamos sin pérdida de generalidad que tanto la configuración como la generación de actividad se refieren a únicamente una aplicación. En otro caso, se puede replicar la arquitectura propuesta por cada aplicación que se quiera incluir en el *testbed* y agregar posteriormente el tráfico resultante.

Para caracterizar el comportamiento de la dinámica de red, es necesario que exista un patrón más o menos estable de la actividad de red. De hecho, cambios sostenidos en estos valores pueden ser indicativos de cambios de uso en la red [19], lo que entra en contradicción con la caracterización del comportamiento dinámico de la red. Por ello, para aplicar nuestro método requerimos que las siguientes hipótesis se cumplan:

- Existe una línea base para la evolución temporal típica del número de conexiones activas en la red [20].
- Existe un proceso característico para la aparición de nuevas conexiones por unidad de tiempo que se puede ajustar utilizando su esperanza.
- La distribución de la duración de la conexión no cambia con el tiempo.

A partir de la primera hipótesis, se sigue que es posible definir una línea base de alta dimensionalidad, basada en una función $F(t), t \in \mathbb{T}$, con $\mathbb{T} \subset \mathbb{R}$ un compacto correspondiente al dominio temporal de la dinámica. Esta línea base se puede definir a partir del análisis de la dinámica de una una red real [20] con el fin de replicar su comportamiento; o para acomodarse a una situación controlada como en el caso de estudio propuesto.

Por otro lado, como en este caso se están modelando las conexiones activas tal y como las vería un elemento de red que recibiese todo el tráfico, en este sistema no aparece ningún efecto de encolado —y por lo tanto, la duración de las conexiones y el tiempo que están activas coincide. Además, de la tercera hipótesis se sigue que la esperanza de la duración de las conexiones (W) debe ser constante.

A partir de $F(t)$ y W, se quiere definir una aproximación de la esperanza del proceso de nuevas conexiones para poder modular la actividad del *testbed* y ajustarla a la dinámica esperada. Siguiendo la demostración de la Ley de Little [21], utilizamos una descomposición del compacto $\mathbb{T}$ en una sucesión de compactos $\{\mathbb{T}_i\}$ tales que su unión es $\mathbb{T}$ y que son disjuntos dos a dos. Ahora, como W es constante es posible definir el número esperado de nuevas conexiones en base a la expresión de la Ec. 1:

$$\lambda(t) = \frac{\sum_{t \in \mathbb{T}_i} F(t)}{W}, \ \forall t \in \mathbb{T}_i \tag{1}$$

Por su parte, el proceso de nuevas conexiones $A(t), t \in \mathbb{T}$ se ajusta de modo que se cumpla la Ec. 2:

$$\mathbb{E}[A(t)] = \lambda(t), \forall t \in \mathbb{T} \tag{2}$$

Finalmente, el nodo de control configura un número aleatorio de nuevas conexiones entre los *hosts* presentes en la topología virtual, generado según el proceso ajustado en cada unidad de tiempo. Después, las nuevas conexiones son activadas, establecidas y mantenidas de forma autónoma por los nodos de generación de carga.

B. Adaptación para generación de tráfico de VoIP

Para adaptar el método general previo al caso de generación de tráfico de VoIP, vamos a considerar el modelo clásico de telefonía en el que el número de nuevas conexiones sigue una distribución de Poisson, de modo que el número de nuevas conexiones $A(t)$ en cada instante t cumple la Ec. 3:

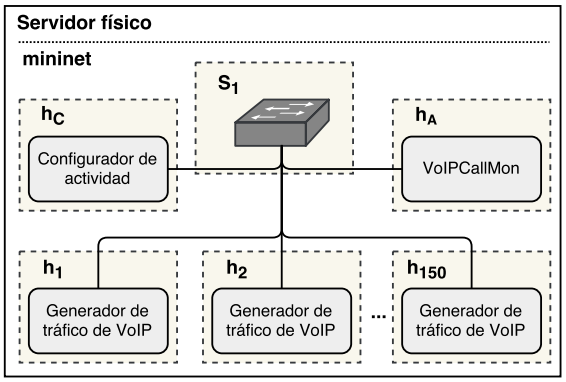


Fig. 2. Topología virtual definida para la evaluación del método de generación de carga.

$$A(t) \sim \text{Poi}(\lambda(t)), t \in \mathbb{T} \tag{3}$$

y la duración de cada conexión k sigue una distribución exponencial, de modo que se cumple la Ec. 4:

$$Dur(\mathcal{C}) \sim \text{Exp}(1/W(t)), t \in \mathbb{T} \tag{4}$$

para todas las conexiones $\mathcal{C}$ iniciadas en el instante t.

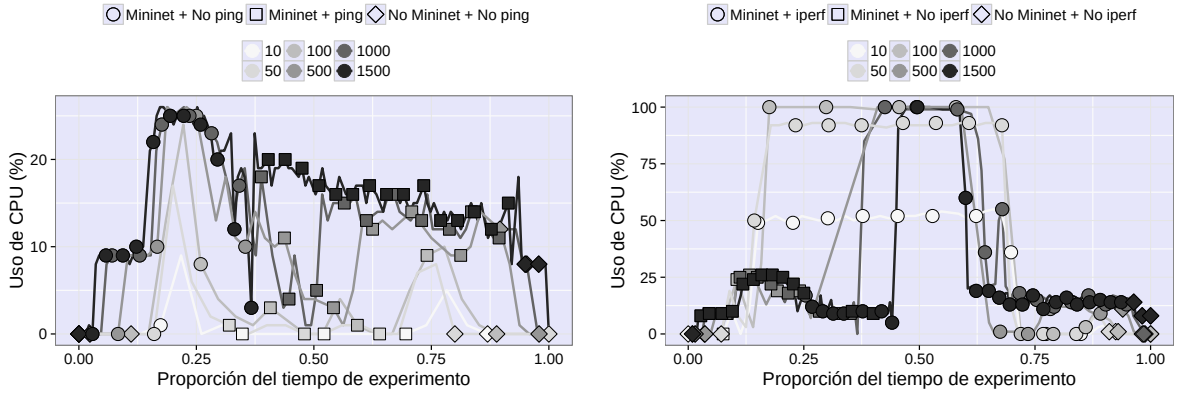
En lo referente a las conexiones, deben generarse dos flujos de datos por cada sentido de una llamada —uno correspondiente a la señalización y otro para los datos multimedia. La combinación de protocolos de señalización y transporte multimedia ha sido seleccionada para representar entornos habituales tanto en redes empresariales como domésticas. En particular, los protocolos de señalización utilizados son el Protocolo de Inicio de Sesión (*Session Initiation Protocol*, SIP) y el Protocolo de Control de Llamada Skinny (*Skinny Call Control Protocol*, SCCP). El protocolo para transporte de datos multimedia es el Protocolo de Transporte de Tiempo Real (*Real-time Transport Protocol*, RTP), utilizando la definición de carga correspondiente al códec G.711.

IV. RESULTADOS EXPERIMENTALES

A. Definición de los experimentos

A continuación se ofrece una descripción completa del *hardware* y la topología virtual considerada durante la realización de nuestros experimentos. Asimismo, el software empleado para la generación de carga se encuentra disponible bajo petición para cualquier persona interesada en su uso o modificación.

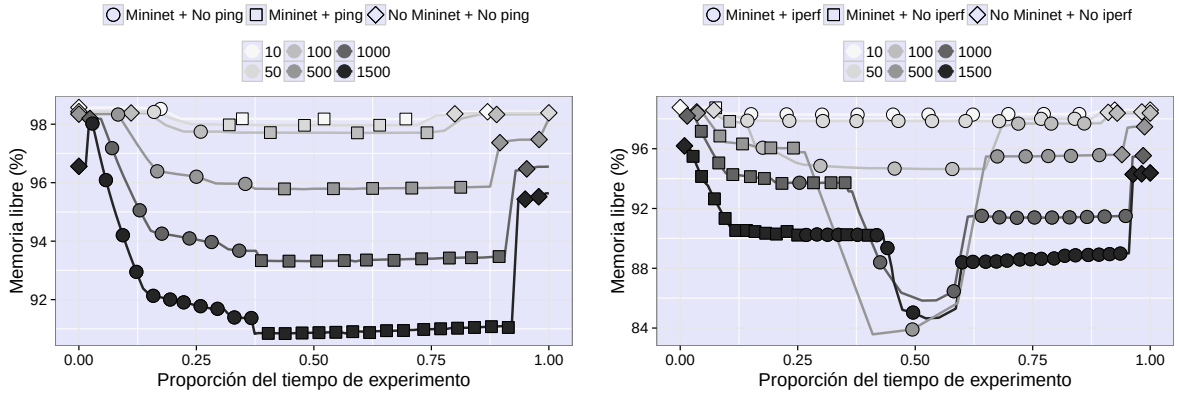
Todos los experimentos han sido ejecutados en un servidor de propósito general equipado con dos procesadores Intel Xeon E5-2620 (6 *cores* por procesador) con una frecuencia de 2.10 GHz y 32 GB de memoria RAM. Para evitar efectos no controlados producidos por la virtualización de *hardware*, las características de *Hyper-Threading* han sido desactivadas. El sistema operativo de este servidor se corresponde con una distribución Ubuntu 14.04.1 instalada con un *kernel* de Linux versión 4.4.0-45. Las pruebas se han ejecutado usando la versión 2.2.2



(a) Entorno de baja carga de red

(b) Entorno de alta carga de red

Fig. 3. Uso de CPU en base al número de nodos de generación de carga, 10 *switches*.



(a) Entorno de baja carga de red

(b) Entorno de alta carga de red

Fig. 4. Uso de memoria en base al número de nodos de generación de carga, 10 *switches*.

de Mininet descargada desde el repositorio Git de la herramienta[1] y utilizando la configuración por defecto.

Mininet utiliza *namespaces* de red para virtualizar los distintos elementos que conforman una topología —*hosts*, *switches*, etc.. En nuestro caso consideraremos Open vSwitch (OvS) como elemento de interconexión de todos los *hosts* presentes en las topologías propuestas. Por ello, como primer conjunto de resultados, analizamos la ocupación de recursos físicos del servidor a medida que se incrementa el número de *hosts* con una topología lineal de 10 *switches* interconectándolos. Con ello, caracterizamos el comportamiento del *testbed* en términos de recursos, asegurando que este escenario no presenta cuellos de botella o limitaciones que impidan la generación de carga.

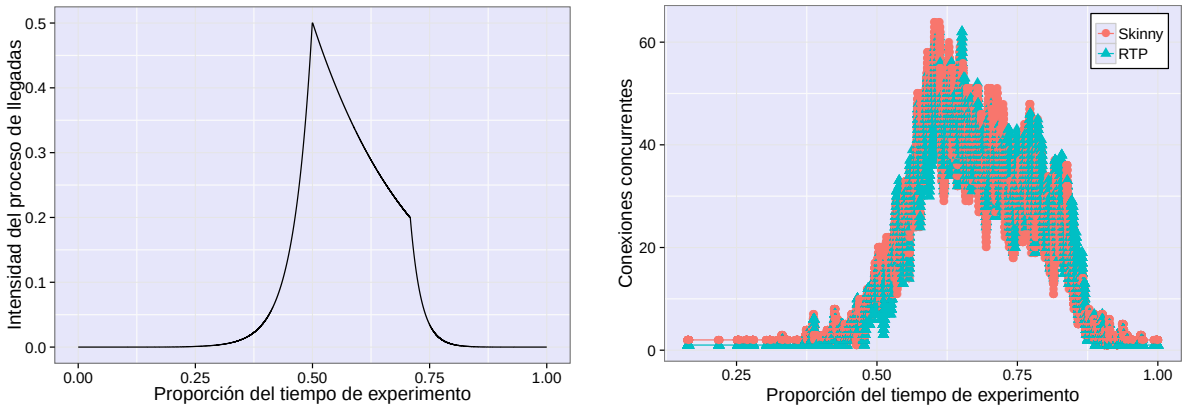
Por otro lado, la topología definida con Mininet correspondiente a las pruebas de generación de carga se muestra en la Fig. 2, indicando los enlaces (sin ningún tipo de limitación) establecidos entre los distintos elementos de red implicados. Con el fin de simular una Red de Área Local (*Local Area Network*, LAN) de terminales de telefonía, se utilizan 150 *hosts* de Mininet $\{h_i\}, i = 1 \ldots 150$, controlados por un *host* que orquesta el número de conexiones establecidas, h_C, y que se interconectan a través de un único *switch* virtual, s_1. Para el análisis de tráfico, se introduce otro *host* h_A que recibe todo el tráfico, que ejecuta una instancia de la herramienta de análisis de tráfico VoIPCallMon [22] —de hecho, el *testbed* se utilizó para evaluar el comportamiento de la herramienta antes de su despliegue en un entorno operativo.

[1] git://github.com/mininet/mininet

B. Comportamiento del entorno de virtualización

En primer lugar, caracterizamos el uso de recursos físicos que supone utilizar Mininet, para obtener evidencias de su viabilidad como plataforma de *testbeds* virtualizados. La Fig. 3 incluye los resultados de consumo de CPU, mientras que la Fig. 4 refleja el consumo de memoria. Ambos recursos son monitorizados en entornos de baja y alta carga de red (definidos usando `ping` e `iperf`, respectivamente), y variando el número de *hosts* activos en la topología emulada desde 10 hasta 1500. De esta forma, barremos todas las posibles situaciones de interés para el caso de estudio que nos proponemos.

En el caso del uso de CPU, se observa que durante la creación de la topología virtual ((Mininet + No ping) no se consume más del 25% de la CPU del equipo. La mayor parte de este consumo se debe a la creación e interconexión de los *switches*. Cuando se usa `ping`, la carga de CPU disminuye substancialmente, ya que la carga de red que genera es un muy limitada —1 paquete ICMP cada segundo. En el caso de `iperf`, la carga de CPU crece hasta el 100% en algunos casos como consecuencia de las elevadas tasas de transmisión y recepción (∼10



(a) Intensidad del modelo de nuevas conexiones.

(b) Dinámica de conexiones de señalización y tráfico multimedia.

Fig. 5. Caracterización de la carga generada por el sistema.

Gb/s). Además, el efecto de incorporar nuevos *hosts* es más notable que en el caso anterior, por el incremento asociado del número de instancias de `iperf` en ejecución.

Los resultados de consumo de memoria muestran resultados similares. Durante la creación de la topología de Mininet se observa que, a mayor número de *hosts*, mayor consumo de memoria. No obstante, el consumo de memoria no resulta excesivo, ya que para 1500 *hosts* la memoria utilizada no representa más del 6% del total disponible. Analizando las diferencias entre los entornos de baja y alta carga, se observa que utilizando `ping` apenas se incrementa el uso de memoria al lanzar los procesos una vez ha sido creada la topología. Por el contrario, cuando se crean los procesos de `iperf`, se incrementa sustancialmente el consumo de memoria, suponiendo hasta un consumo extra del 12% en el peor de los casos.

Las conclusiones que se extraen de estos experimentos son: (i) que la carga de CPU es altamente dependiente del tipo de proceso de red que ejecutemos en los *hosts* de Mininet y (ii) que el consumo de memoria depende en gran medida del número de *hosts* usado en el escenario. Además, podemos observar que, en términos de memoria, emular una red compleja con 1500 *hosts* y 10 *switches* no consume más de un 20% de la memoria de la máquina.

C. Carga generada

Tras contar con evidencia suficiente de la viabilidad de emulación del despliegue, estudiamos el comportamiento de una red de teléfonos con soporte para VoIP emulada, incluyendo 150 terminales. En nuestro caso, el objetivo es estudiar la funcionalidad y estabilidad de VoIPCallMon a la hora de monitorizar una red local de VoIP. Por ello, fijamos como requisito del *testbed* que se tenga una concurrencia de 60 llamadas en el período de máxima actividad y que muestre la dinámica de actividad típica de una red empresarial, marcada por actividad en horario laboral y hora más cargada alrededor de las 12 del mediodía.

Para ello, fundamentándonos en experiencias previas de monitorización, prefijamos un perfil de concurrencia diario basado en cuatro puntos temporales que lo parametrizan:

- *Hora de comienzo de actividad:* 6 de la mañana.
- *Hora de finalización de actividad:* 8 de la tarde.
- *Hora de máxima actividad (H1):* 12 del mediodía.
- *Transición de actividad de media tarde a cierre (H2):* 5 de la tarde.

y dos adicionales que nos dan la concurrencia en H1 y H2. Posteriormente, definimos una función a trozos que sigue ese perfil, y la transformamos para obtener el patrón de nuevas conexiones (llamadas) según muestra la Fig. 5(a).

Utilizando el método descrito en la Sección III, generamos actividad que replique este perfil para un día de actividad, y analizamos el tráfico en el nodo que recibe el agregado ejecutando VoIPCallMon. La generación de tráfico se ha acometido utilizando la librería de Python Scapy, por su versatilidad a la hora de conformar paquetes de tráfico. La utilización de esta librería se fundamenta en que SCCP es un protocolo propio de teléfonos de VoIP de Cisco. Esto ocasiona que sea la alternativa más simple para poder evaluar un sistema de monitorización con soporte para estos terminales, si no se tiene acceso a un despliegue real. Por otro lado, y tal y como se ha mencionado anteriormente, las llamadas generadas utilizan G.711 como códec para el audio transmitiendo un paquete por cada 20ms de muestras. En base a medidas empíricas, se ha seleccionado 100s como duración media, ya que este valor representa lo esperable en entornos de oficina.

Las series temporales de conexiones activas detectadas por esta herramienta se muestra en la Fig. 5(b), mostrando el buen ajuste a los requisitos para la dinámica de evaluación. Además, al simular distintas trayectorias sobre el dominio $\mathbb{T}$, se observa que la actividad en instantes equivalentes en distintas realizaciones de la dinámica presenta un comportamiento estable adaptado a los parámetros antes indicados —la Fig. 6 ilustra esta idea sobre 30s del período de más actividad de 40 trayectorias, mostrando gráficos de cajas para cada día y la función de medias. La variación de la función de medias viene dada por la varianza del proceso de generación de nuevas conexiones —igual a la media, por seguir una distribución de Poisson.

V. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo presenta evidencias de la viabilidad de la definición de *testbeds* para entornos emergentes de red usando Mininet como plataforma de virtualización ligera.

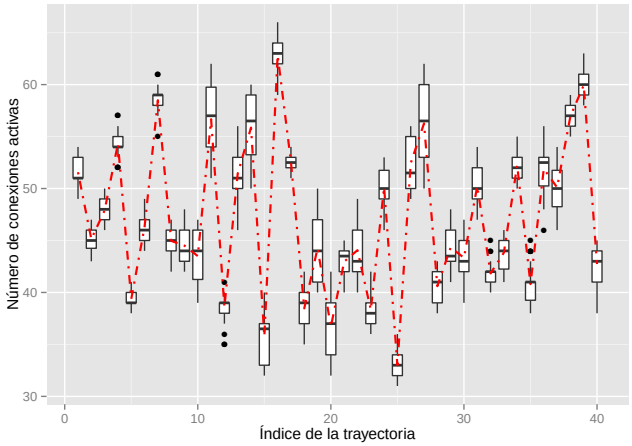


Fig. 6. Evolución de las conexiones en 30s del período de máxima actividad, 40 trayectorias. Cada diagrama de caja muestra los valores para un mismo día, y la línea roja discontinua la función de medias.

Hemos caracterizado Mininet en términos de consumo físico de recursos, analizando la carga de CPU y uso de memoria en entornos de baja y alta actividad de red. Nuestro resultados indican que es posible desplegar topologías con más de 1000 elementos de red en servidores de propósito general, siendo la carga de CPU muy dependiente de la actividad de los nodos.Con estos resultados, comprobamos la viabilidad del uso de esta plataforma para un caso de estudio que ilustra un método de generación de carga capaz de replicar perfiles de actividad no estacionarios. Este caso de estudio plantea la emulación de un despliegue empresarial de VoIP, mostrando que nuestro método genera tráfico según el perfil prefijado.

Como trabajo futuro, planteamos la extensión de la metodología propuesta a otro tipo de tráfico para facilitar, por ejemplo, la evaluación de mecanismos de transmisión de vídeo desde dispositivos empotrados, el funcionamiento de redes de sensores, o la introducción de métodos de validación de parámetros de calidad. Asimismo, estamos estudiando el comportamiento de esta plataforma introduciendo penalizaciones en los enlaces desplegados, mediante el uso del comando `tc` de Linux; y las posibilidades que ofrecen OvS y el uso de OpenFlow.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional a través de los proyectos TRÁFICA (MINECO / FEDER TEC2015-69417-C2-1-R) y RACING DRONES (MINECO / FEDER RTC-2016-4744-7).

REFERENCIAS

[1] D. Pediaditakis, C. Rotsos, and A. W. Moore, "Faithful Reproduction of Network Experiments," in *Proceedings of the Tenth ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ser. ANCS '14, 2014, pp. 41–52.

[2] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A Survey on Software-Defined Networking," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 27–51, 2015.

[3] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[4] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[5] J. Horneber and A. Hergenröder, "A Survey on Testbeds and Experimentation Environments for Wireless Sensor Networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1820–1838, 2014.

[6] V. Moreno, P. M. S. del Río, J. Ramos, J. J. Garnica, and J. L. García-Dorado, "Batch to the Future: Analyzing Timestamp Accuracy of High-Performance Packet I/O Engines," *IEEE Communications Letters*, vol. 16, no. 11, pp. 1888–1891, 2012.

[7] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: rapid prototyping for Software-Defined Networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. ACM, 2010, pp. 19:1–19:6.

[8] J. Yan and D. Jin, "VT-Mininet: Virtual-time-enabled Mininet for Scalable and Accurate Software-Define Network Emulation," in *Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research*, ser. SOSR '15, 2015, pp. 27:1–27:7.

[9] N. Handigol, B. Heller, V. Jeyakumar, B. Lantz, and N. McKeown, "Reproducible Network Experiments Using Container-based Emulation," in *Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '12, 2012, pp. 253–264.

[10] B. Lantz and B. O'Connor, "A Mininet-based Virtual Testbed for Distributed SDN Development," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 365–366, 2015.

[11] L. Baldesi and L. Maccari, "NePA TesT: network protocol and application testing toolchain for community networks," in *2016 12th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*, 2016, pp. 1–8.

[12] M. C. Weigle, P. Adurthi, F. Hernández-Campos, K. Jeffay, and F. D. Smith, "Tmix: A Tool for Generating Realistic TCP Application Workloads in Ns-2," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 3, pp. 65–76, 2006.

[13] A. Botta, A. Dainotti, and A. Pescapé, "A tool for the generation of realistic network workload for emerging networking scenarios," *Computer Networks*, vol. 56, no. 15, pp. 3531 – 3547, 2012.

[14] P. Rygielski, V. Simko, F. Sittner, D. Aschenbrenner, S. Kounev, and K. Schilling, "Automated Extraction of Network Traffic Models Suitable for Performance Simulation," in *Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering*, ser. ICPE '16, 2016, pp. 27–35.

[15] C. Bachmeir, P. Tabery, S. Uzumcu, and E. Steinbach, "A scalable virtual programmable real-time testbed for rapid multimedia service creation and evaluation," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 3, 2003, pp. III–257–60 vol.3.

[16] W. Fuertes and J. E. López de Vergara, "An emulation of vod services using virtual network environments," *Electronic Communications of the EASST*, vol. 17, 2009.

[17] M. Raza, S. Chowdhury, and W. Robertson, "SDN based emulation of an academic networking testbed," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2016, pp. 1–6.

[18] J. M. Jimenez, J. O. R. Martínez, A. Rego, A. Dilendra, and J. Lloret, "Study of multimedia delivery over software defined networks," in *Network Protocols and Algorithms*, vol. 7, no. 4. Macrothink Institute, 2015, pp. 37–62.

[19] F. Mata, J. L. García-Dorado, and J. Aracil, "Detection of traffic changes in large-scale backbone networks: The case of the Spanish academic network," *Computer Networks*, vol. 56, no. 2, pp. 686 – 702, 2012.

[20] D. Muelas, J. E. López de Vergara, J. R. Berrendero, J. Ramos, and J. Aracil, "Facing Network Management Challenges with Functional Data Analysis: Techniques & Opportunities," *Mobile Networks and Applications*, pp. 1–13, 2016.

[21] J. D. Little, "Little's Law as Viewed on Its 50th Anniversary," *Operations Research*, vol. 59, no. 3, pp. 536–549, 2011.

[22] J. L. García-Dorado, P. M. Santiago del Río, J. Ramos, D. Muelas, V. Moreno, J. E. López de Vergara, and J. Aracil, "Low-cost and high-performance: VoIP monitoring and full-data retention at multi-Gb/s rates using commodity hardware," *International Journal of Network Management*, vol. 24, no. 3, pp. 181–199, 2014.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6514

Mecanismos de nivel de transporte para la optimización de envíos en base al ancho de banda estimado sobre Long Fat Networks

Alan Briones, Guillermo Dobao, Ramon Martín de Pozuelo, Agustín Zaballos, Guiomar Corral
Grupo de Investigación en Internet Technologies y Storage (GRITS)
Universidad Ramon Llull – La Salle
08022
<abriones, gdobao, ramonmdp, zaballos, guiomar> @salleurl.edu

***Resumen*- Este artículo investiga los mecanismos de diferentes protocolos de transporte en transferencias sobre redes de alta capacidad y alto retardo, conocidas como *Long Fat Networks* (LFNs), para un envío eficiente de los datos. *Transport Control Protocol* (TCP) presenta limitaciones de rendimiento y flexibilidad. En la literatura se pueden encontrar diferentes propuestas de variantes del comportamiento de TCP, protocolos como *Stream Control Transmission Protocol* (STCP) o soluciones que proporcionan una comunicación confiable y mecanismos de control de congestión sobre *User Datagram Protocol* (UDP). En este artículo se presentan una serie de mecanismos de nivel de transporte para la optimización de transferencias de datos sobre redes LFN. Estos mecanismos ofrecen un rendimiento elevado utilizando todo el ancho de banda disponible del enlace mediante un proceso de cálculo del estado de la red y un control de congestión activo para la utilización de todo el *bandwidth*, a la vez que reactivo en caso de producirse pérdidas para evitar congestiones en la red. El objetivo es demostrar la eficiencia de dichos mecanismos, así como su adaptabilidad y *aggressive friendliness* respecto a otros protocolos de transporte mediante el despliegue de una serie de pruebas expuestas en este artículo.**

***Palabras Clave*- protocolo de transporte, long fat networks, pérdidas de paquetes, ancho de banda, retraso, control de congestión.**

I. Introducción

Internet ha ido cambiando a lo largo de los años. En un mundo cada vez más interconectado, en el que los usuarios requieren y consumen cada vez más información y la necesitan de manera inmediata, se ha producido un cambio de paradigma respecto a cómo se concibieron inicialmente las redes y su uso.

Este aumento de uso está ligado en gran parte al incremento de la oferta de servicios multimedia. El amplio abanico de contenidos, tanto por la cantidad ofrecida como por el tamaño de los mismos, ha evidenciado la necesidad de disponer de redes con mayor ancho de banda para conexiones *end-to-end*, donde los equipos finales están separados por grandes distancias.

Dentro de la clasificación de enlaces de alta capacidad, existe un tipo de red conocida como *Long Fat Network* (LFN) [1]. La principal característica que define este tipo de redes es su alto *Bandwidth* (BW) y unos valores elevados de *Round Trip Time* (RTT). Una red es considerada LFN si su *Bandwidth-Delay Product* (BDP) es mayor a 12500 bytes (10^5 bits). Por ejemplo, un enlace de 1 Gbps y 1 ms de RTT, obtiene un BDP de 10^6 bits, siendo clasificada como LFN.

Estas características de las LFNs provocan que *Transmission Control Protocol* (TCP), el protocolo de transporte más utilizado en la red, no obtenga un buen rendimiento, lo que estimuló la definición de una extensión del protocolo [2]. Las principales problemáticas surgen debido al propio diseño de TCP, como la limitación de la ventana de congestión, la cual solo permite una ventana máxima de 65 KB debido a que su campo en la cabecera es de 16 bits. Otra problemática es el elevado valor de *Round Trip Time* (RTT) y la acumulación de mensajes de confirmación debido a los tiempos de *timeout* y retransmisión en el caso de producirse pérdidas.

EDITORIAL UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Para solventar las problemáticas mencionadas anteriormente, se han propuesto diferentes mecanismos para TCP, así como nuevos protocolos que permitan extraer el máximo rendimiento de las redes LFN.

En este paper se propone otra alternativa mediante un mecanismo de control de congestión que incluye el cálculo del estado del enlace, permitiendo al protocolo adaptar su comportamiento y maximizar el uso del canal de manera casi inmediata. Finalmente, se muestra una prueba de concepto del funcionamiento del protocolo propuesto.

La organización del artículo queda de la siguiente manera. En la Sección II se repasa brevemente el Estado del Arte de los protocolos de nivel de transporte más destacados. En la Sección III se presenta los mecanismos propuestos. En la sección IV se explican una serie de directrices para la implementación dichos mecanismos, así como el Testbed utilizado, las pruebas realizadas y sus resultados. Finalmente, en la sección V se extraen las conclusiones y se presentan las líneas de futuro.

II. Estado del Arte

Este Estado del Arte se focaliza en mostrar los mecanismos de control de congestión de los diferentes protocolos de transporte más destacados para larga distancia [3].

La misión del control de congestión es detectar la congestión de la red antes de que ésta se colapse y actuar en consecuencia [4][5].

A. *Principales causas que producen congestión en la red*

Una de las principales causas de congestión es la incapacidad del *host* destino de procesar toda la información recibida, provocando que se descarten parte de los datos entrantes.

Otra de las causas de congestión es debida a que algún dispositivo intermedio no es capaz de tratar toda la información recibida, por lo que empieza a encolar los datos recibidos, generando retrasos (*delay*). En casos de saturación total existe la posibilidad de que llegue a descartarlos, generando pérdidas.

B. *Estrategias para poder detectar y mitigar la congestión en la red*

Para evitar que la red se sature, es necesario que el protocolo sea capaz de interpretar el estado de la red y actuar en consecuencia para extraer el máximo rendimiento de ella sin que se produzca un deterioro de la comunicación o pérdidas de información.

Por un lado, se plantean técnicas preventivas, como el control de admisión. Se limita el número de usuarios, se monitoriza que el flujo no exceda un límite fijado o se regula el tráfico en el acceso de la red. Para poder aplicar estos mecanismos se debe diseñar la red adecuadamente y tener el control total de ella.

Por otro lado, existen técnicas reactivas que resuelven la congestión una vez ya se ha detectado o cuando está a punto de producirse.

Para su detección existen dos clases:

- Realimentación directa: Los nodos intermedios de la red señalizan a los extremos la existencia de congestión o si hay riesgo de que se produzca. Se indica mediante el marcado de paquetes o el envío de paquetes especiales.
- Realimentación indirecta: Los extremos de la comunicación detectan la congestión basándose en las pérdidas de paquetes, retrasos y la variabilidad en los tiempos de recepción (*jitter*).

C. *Protocolos TCP para larga distancia*

En la actualidad existen diferentes propuestas de protocolos TCP para la transmisión de datos a larga distancia.

Parallel TCP Reno (P-TCP)

P-TCP [6] tiene un funcionamiento similar a TCP Reno y su control de congestión consiste en la transmisión de varios streams de datos en paralelo.

Un inconveniente es que el número de streams puede variar según la red, el tipo de información, etc. Se calcula, aproximadamente, 16 streams para redes de larga distancia, no permitiendo priorización entre tráficos.

Scalable TCP Reno (S-TCP)

S-TCP [7] basa su control de congestión en una mejora de TCP Reno. En lugar de utilizar *Additive Increase*, utiliza un incremento exponencial. Además, aplica un decremento multiplicativo para la reducción de la ventana, siendo un protocolo menos agresivo en caso de pérdidas.

High Speed TCP (HS-TCP)

HS-TCP [6] tiene un comportamiento como TCP Reno en una ventana de congestión pequeña. A partir de que su *Congestion Window* (CW) supere una cantidad de paquetes determinada, un *flag* es activado, modificando su CW en función de una tabla predefinida por el propio protocolo.

El hecho de utilizar una tabla predefinida no permite que sea un protocolo dinámico ni flexible.

H-TCP

H-TCP [8] está basado en HS-TCP. En lugar de utilizar una tabla, utiliza un algoritmo de AIMD heterogéneo que permite adaptarse a las capacidades del canal de forma eficiente.

Basa su comportamiento en las pérdidas producidas y el RTT, mejorando la eficiencia de TCP adaptando sus parámetros dinámicamente.

Es un control de congestión complejo en cuanto a cálculos e implementación.

HSTCP-LP

HSTCP-LP está basado en HS-TCP y TCP-LP [6]. Utiliza solo el ancho de banda residual que no es utilizado por otros flujos. En el caso de haber varios flujos, todos tienen la misma prioridad.

Al tratarse de un protocolo no intrusivo, en redes congestionadas, ve afectado su *throughput*.

D. Otro protocolos para larga distancia

A parte de protocolos basados en TCP, existen otras propuestas como ahora *Stream Control Transmission Protocol* (SCTP) [9] o protocolos basados en UDP que proponen soluciones ante las ineficiencias propias de TCP.

Stream Control Transmission Protocol (SCTP)

SCTP aporta una serie de funcionalidades y mecanismos adicionales de los que TCP carece. Propone un sistema de cabeceras para la gestión de los flujos. Además, este protocolo provee de mayor seguridad en el establecimiento de conexión mediante un proceso de *4-way handshake,* en vez del *3-way handshake* utilizado en TCP, para prevenir ataques de denegación de servicio (DoS).

A pesar de utilizar las mismas variables que TCP, como la CW y de utilizar un control de congestión similar al de TCP (*slow-start*, *congestion avoidance*) con ligeras modificaciones, SCTP basa su funcionamiento en la ventana de recepción (RWND) para evitar que se sature el receptor.

Además, introduce otras características como la posibilidad de la recepción desordenada de paquetes, así como el *multihoming* y el *multistreaming*.

UDP-based protocols

Un conjunto de protocolos basados en UDP apareció tratando de proporcionar un control eficaz de la congestión y funciones de confiabilidad sin ser protocolos orientados a conexión por definición. Entre todos ellos (Tsunami, PA-UDP, SABUL, etc.), *UDP-based Data Transfer* (UDT) es el que presenta una mejor utilización del rendimiento y ofrece una solución prometedora para transferencias de datos pesadas a través de redes de larga distancia, pero también se han identificado algunas deficiencias y dificultades en su implementación real [10][11][12].

III. Especificación de mecanismos

El objetivo principal de los mecanismos propuestos es conseguir el máximo ancho de banda disponible de manera agresiva respecto a otros flujos para el envío de grandes volúmenes de datos *non-real time* sobre redes LFN.

Los mecanismos propuestos se inspiran principalmente en los protocolos SCTP y UDT, siendo planteados para su uso sobre UDP.

Por un lado, se utiliza la estructura de mensajes y cabeceras en SCTP. Principalmente, el mecanismo propuesto se basa en la gestión de streams propuesta por SCTP (la cual su explicación no es objetivo del artículo), así como en el uso del *Selective-ACK* (SACK) para la confirmación de información recibida y solicitud de información perdida; siendo la principal diferencia con SCTP que éste trabaja sobre TCP y el mecanismo propuesto lo hace sobre UDP, cambiando el planteamiento del control de congestión.

Por otro lado, se utiliza un control de congestión basado en el protocolo UDT. El mecanismo basa su fórmula de cálculo de estimación de ancho de banda y control de congestión en la fórmula original de UDT pero modificando el cálculo del incremento del número de paquetes por ráfaga (en base a la eficiencia del envío, haciéndolo más agresivo), así como el cálculo del ancho de banda estimado. Este cálculo de BW es realizado mediante señalización *in-band* durante el envío de mensajes de datos. Todo ello implica una mejora en términos de utilización del rendimiento y flexibilidad.

Este apartado se focaliza en la explicación de los dos mecanismos propuestos; el cálculo del estado de la red y el control de congestión.

A. Estado de la red

Una vez establecida la asociación y negociada la seguridad de la comunicación, se lleva a cabo un proceso de entrenamiento para conocer el estado de la red.

Este cálculo del estado de la red permite al protocolo estimar el ancho de banda máximo de la red, así como el *Round Trip Time* (RTT) y el *Receiving Rate* (RR), o ventana de recepción, de la comunicación.

Funcionamiento

Tal y como se muestra en la *Fig.* 1, este proceso está formado por el envío de 10 ráfagas con el objetivo de tener diferentes muestras representativas a la hora de realizar el cálculo, con bloques ordenados en grupos de 2 (*packet-pair*) a 20 paquetes (*packet-train*) de 9000 bytes (jumboframes), donde se escoge el número de paquetes por bloque dependiendo de la velocidad estimada del enlace en la iteración anterior. Cuantos más paquetes se utilicen, más se reduce la probabilidad de error en la estimación del ancho de banda. Los paquetes a enviar son paquetes de datos, los cuales pueden contener parte de la información de la transferencia o ser paquetes de datos vacíos, según lo requiera la situación de la comunicación.

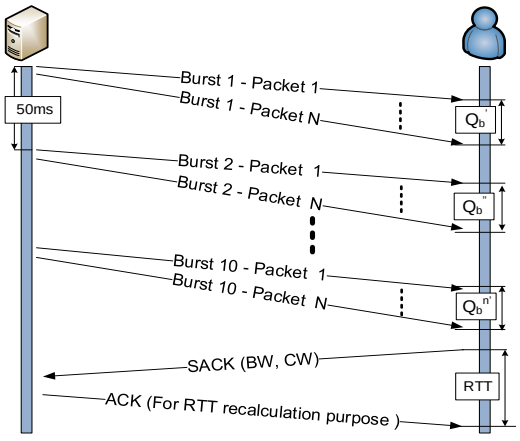


Fig. 1 Estado de la red – Cálculo del BW

Al inicio de esta fase de cálculo del ancho de banda se envían los paquetes de cada bloque de forma consecutiva por parte del emisor hacia el receptor. Por cada bloque, un proceso en el receptor registra el tiempo

de llegada del primer paquete y del último, calculando la diferencia de tiempo y el tamaño total (en bits) de los paquetes recibidos (la suma total del tamaño de cada uno de ellos) para realizar un cálculo del ancho de banda.

Con estos datos, se extrae el ancho de banda (BW), mediante la Ec. 1, donde *b* es el número de bits recibidos y *Qb* es el tiempo entre la llegada del primero y el último de estos bits, y que corresponde al tiempo que ha tardado la red en transportar estos datos. De esta forma se extrae el *bottleneck bandwidth*, es decir, el máximo ancho de banda al que podremos enviar por esa red.

$$BW = \frac{b}{Q_b} \qquad \text{(Ec.1)}$$

El servidor realiza, mediante esta fórmula (Ec.1), estimaciones del BW, una para cada una de las 10 ráfagas de bloques de paquetes, almacenando estos valores. Una vez ha hecho el cálculo 10 veces y dispone de 10 valores de BW, realiza la moda de los valores extraídos para obtener un valor de BW estable y consolidado, evitando calcular la media, la cual se ve afectada en mayor medida por algún cálculo del BW anómalo y falseando el valor real.

Una vez realizado este cálculo, el receptor le comunica al emisor el BW estimado, mediante un mensaje de confirmación selectiva (SACK), así como otros valores de los que ya dispone como la *Congestion Window* (CW) o el *Receiving Rate* (RR) en paquetes por segundo.

Para extraer el *Round Trip Time* (RTT), al enviar el cliente un mensaje de contestación (ACK) como confirmación, el servidor guarda el tiempo transcurrido entre el envío del mensaje de confirmación (SACK) y la recepción del ACK.

Este proceso de cálculo del ancho de banda máximo se realiza de manera *in-band* durante la transmisión de datos mediante los propios mensajes de datos.

B. Control de congestión

Tras el proceso de entrenamiento se descubre el ancho de banda máximo del que puede disponer ese flujo.

Con este valor extraído de la red, se establece un tanto por ciento de ese valor como la velocidad de envío (*Sending Rate* (SR)) inicial. Dependiendo de la agresividad deseada, se establece valor más o menos elevado.

El envío de los datos se realiza mediante ráfagas separadas por un tiempo determinado por el RTT o la mínima resolución temporal que pueda ofrecer el sistema operativo y el hardware sobre el que funcione el proceso. Este tiempo será calculado mediante Ec.2.

$$\mathrm{T_{burst}} = \max(50\mathrm{ms}, \mathrm{RTT}) \qquad \text{(Ec.2)}$$

Tras conocer la velocidad a la que se envían inicialmente los paquetes y determinada la separación entre ráfagas, se define el número de paquetes que se envían en cada ráfaga mediante Ec.3.

$$\#\mathrm{Packets\_per\_burst} = \mathrm{SR} * \mathrm{T_{burst}} \qquad \text{(Ec.3)}$$

En la *Fig.* 2 se muestra el proceso de envío de los datos. El receptor guarda la información que va recibiendo, a la vez que lista los paquetes perdidos, los cuales son pedidos en el siguiente mensaje de confirmación (SACK) de los datos recibidos.

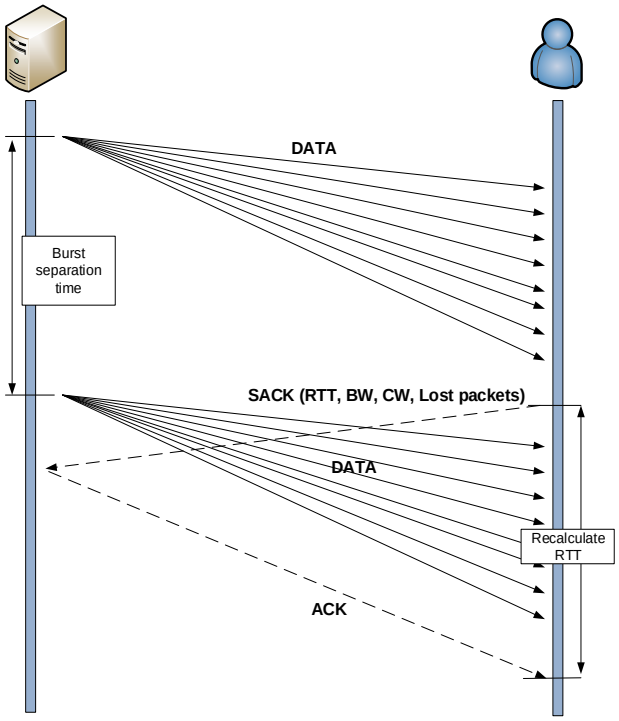


Fig. 2 Envío de datos

El mensaje de confirmación de datos por parte del receptor se envía de manera asíncrona respecto a la ráfaga del emisor al finalizar la ráfaga asociada.

Utilizando estos datos proporcionados por el receptor, el emisor irá modificando el valor la velocidad de envío de paquetes (*Sending Rate* (SR)).

Si al recibir un SACK, éste no indica que se han detectado pérdidas, se realiza el siguiente proceso (Ec.4, Ec.5, Ec.6), calculando el número de paquetes a incrementar (Inc_p) en la siguiente iteración (Ec. 4).

$$Inc_p = \max\left(\frac{1}{\mathrm{MTU}}, 10^{\log(\mathrm{BW}-(\mathrm{SR}*\mathrm{MTU}*8))-C)}\right) \qquad \text{(Ec.4)}$$

Donde el valor de C varía según la eficiencia de la comunicación (Ec.5). Considerando la eficiencia como el cociente entre el ancho de banda utilizado en la iteración anterior respecto al recientemente calculado.

$$C = \begin{cases} 7, & \frac{BW_{IteraciónAnterior}}{BW_{Actual}} < 0.8 \\ \left(\frac{BW_{IteraciónAnterior}}{BW_{Actual}} * 10\right) - 1, & \frac{BW_{IteraciónAnterior}}{BW_{Actual}} \geq 0.8 \end{cases} \qquad \text{(Ec.5)}$$

A diferencia con UDT, el mecanismo propuesto aplica un incremento dinámico en base a la eficiencia del enlace, siendo más agresivo cuando la eficiencia del envío es menor al 80%.

Posteriormente, se calcula el *Sending Rate* (paquetes/ráfaga) de la siguiente ráfaga (Ec.6).

$$\text{Sending Rate (SR)} = \frac{(T_{burst} * SR) + inc_p}{T_{burst}} \quad \text{(Ec.6)}$$

Si por el contrario se ha detectado algún paquete perdido, se aplica la siguiente fórmula (Ec.7):

$$\text{Sending Rate (SR)} = \frac{SR}{1 + 0.125 * \frac{(SR * MTU * 8)}{BW}} \quad \text{(Ec.7)}$$

El objetivo es lograr enviar el mayor número de paquetes en una ráfaga sin saturar el enlace con la mayor eficiencia posible. Esto es posible gracias a la estimación del ancho de banda.

En caso de producirse pérdidas, se realiza una reducción del número de paquetes a enviar en la siguiente ráfaga para evitar que se produzcan más pérdidas, la vez que se busca la máxima eficiencia en el envío.

En el siguiente apartado se muestra el comportamiento y los resultados de estos mecanismos en diferentes situaciones sobre redes LFN durante el envío de datos.

IV. Pruebas y Desarrollo

A partir del análisis y el diseño expuesto en los apartados anteriores, se realiza una primera implementación de los mecanismos presentados con tal de analizar el rendimiento de éstos en redes LFN.

A. *Estructura de implementación del protocolo*

Se debe diferenciar dentro del marco de la comunicación dos roles en lo referente a un envío *Peer to Peer*, emisor y receptor. Ambos nodos finales de la comunicación deben disponer de una instancia de control que gestione los flujos entrantes del protocolo, creando un proceso de recepción específico para cada flujo.

Así pues, una instancia gestiona la transmisión de un único flujo UDP, utilizándose este protocolo como "*Socket Transport*". Esta elección permite el desarrollo sobre una base sólida sin de mecanismos de control de congestión.

Partiendo de estas premisas, se implementan los mecanismos siguiendo una filosofía reactiva en recepción y secuencial en emisión, asumiendo esta última el rol de *master* en la comunicación.

A continuación, se plantea el *testbed* utilizado y las pruebas realizadas para mostrar el rendimiento de los mecanismos en redes LFN.

B. *Testbed*

El *testbed* desplegado para mostrar el comportamiento de los mecanismos LFN se presenta en el escenario de la Fig. 3.

Este consta de dos nodos extremos (receptor y emisor) interconectados a través de un nodo central el cual emula el comportamiento de una red WAN con características LFN. Las conexiones físicas entre dispositivos se realizan mediante pares trenzados CAT-5 a 100 Mbps Full Duplex y latencias de hasta 100ms.

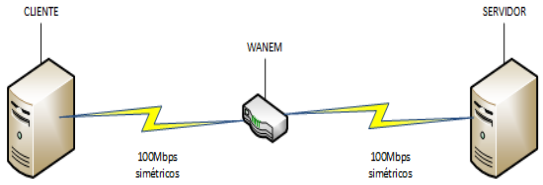


Fig. 3. Diagrama del escenario de pruebas

Para simular diferentes características de red, se utiliza el software *Wanem* [13] en el nodo central. Con tal de generar tráfico adicional entre los nodos finales se utiliza:

- *Iperf* [14] para generar tráfico *UDP*
- *File Transport Protocol* (FTP) para generar tráfico *TCP friendly*

Los objetivos que persiguen demostrar las pruebas a realizar sobre el *testbed* son:

(O1) Eficiencia. Máximo ancho de banda medio alcanzado (Mbps) con diferentes niveles de pérdidas producidas en el enlace y diferentes velocidades de enlace.

(O2) Adaptabilidad. Detección de congestión y modificación del *Sending Rate* para minimizar las pérdidas y maximizar el ancho de banda útil utilizado (Mbps).

(O3) *Friendly Aggressiveness*. Agresividad frente a otros flujos TCP y UDP contemplando el estado de la red.

C. *Pruebas*

Sobre el Testbed (*Fig.* 3) se plantean 3 conjuntos de pruebas:

(P1) Transmisión de un único flujo ante las parametrizaciones descritas en la Tabla I con tal de demostrar la eficiencia del envío ante diferentes velocidades y niveles de pérdidas en el enlace (O1).

(P2) Transmisión de un único flujo compartiendo enlace con flujos agresivos intermitentes no adaptativos (generados con *Iperf*) para demostrar la adaptabilidad del control de congestión (O2).

(P3) Transmisión de un flujo compartiendo enlace con otros protocolos, con tal de demostrar el *Friendly Aggressiveness* frente a estos (O3):

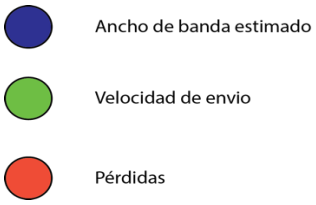
a. UDP (*Iperf*)

b. TCP (FTP)

c. UDP (Mismo control de congestión)

En los siguientes subapartados se exponen, de manera ordenada, los resultados de las pruebas planteadas.

Durante todas las pruebas se envía un total de 1GB. A nivel de gráficas, el color azul indica el ancho de banda estimado, el color verde la velocidad de envío y el color rojo las pérdidas.



Prueba 1

Se plantean diferentes configuraciones (*Tabla I*) con dos objetivos (O1). El primer objetivo (O1.1) es comparar el ancho de banda utilizado respecto al real del enlace (pruebas $P1_1$ y $P1_2$). Y el segundo objetivo (O1.2) es observar la reacción ante la introducción de pérdidas (pruebas $P1_{2\text{-}8}$).

Tabla I
PARAMETRIZACIONES PRUEBA I

Prueba	Velocidad de enlace	Pérdidas
$P1_1$	50 Mbps	0%
$P1_2$	100 Mbps	0%
$P1_3$	100 Mbps	0.001%
$P1_4$	100 Mbps	0.01%
$P1_5$	100 Mbps	0.1%
$P1_6$	100 Mbps	1%
$P1_7$	100 Mbps	3%
$P1_8$	100 Mbps	5%

En la *Fig. 4* y *Fig. 5* se puede observar el resultado de ($P1_1$) y ($P1_2$). En $P1_1$ se aprecia como trabajando a 50Mbps, el ancho de banda estimado por el protocolo es de 49Mbps. En ($P1_2$), a 100Mbps, se estima una velocidad de 96Mbps.

En ambos casos se supera el 95% de utilización de enlace, confirmando parcialmente (O1). Cabe destacar que la velocidad de envío real del protocolo se sitúa en la mayoría de las situaciones por debajo de aquella estimada. Esto es debido al *step* usado al incrementar la velocidad de envío, el definido por el tamaño de los paquetes (*Jumboframe*) que se envían.

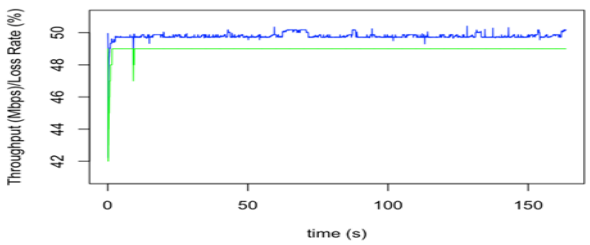


Fig. 4. Resultado de la prueba $P1_1$

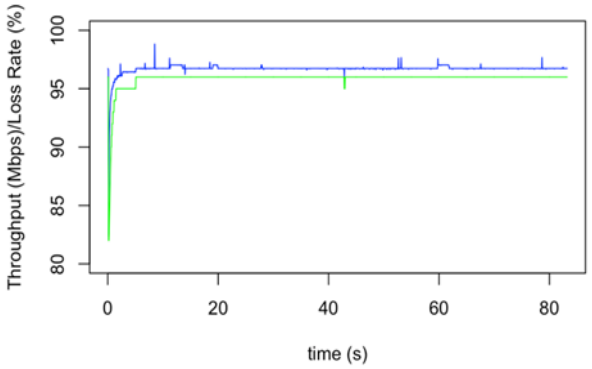


Fig. 5. Resultado de la prueba $P1_2$

Se puede observar como a mayor velocidad de enlace, se reduce la precisión de estimación. Dicha reducción es debida a diversos factores (tamaño de los paquetes, tiempos de ráfaga,…). Esto resalta la importancia del proceso de conocimiento del estado de la red para una parametrización eficiente de las variables dependientes del tipo de enlace.

En las siguientes pruebas se incorporan diferentes niveles de pérdidas. En la prueba ($P1_4$), por ejemplo, se incorpora un porcentaje de pérdidas moderado (0.01%), observándose una disminución de velocidad de transmisión (Fig. 6). La variación máxima de velocidad es de 12Mbps, la cual se recupera rápidamente. No obstante, la velocidad de transmisión media no se ve afectada por las pérdidas moderadas. Esta se aproxima a la velocidad de transmisión máxima estimada, 96 Mbps.

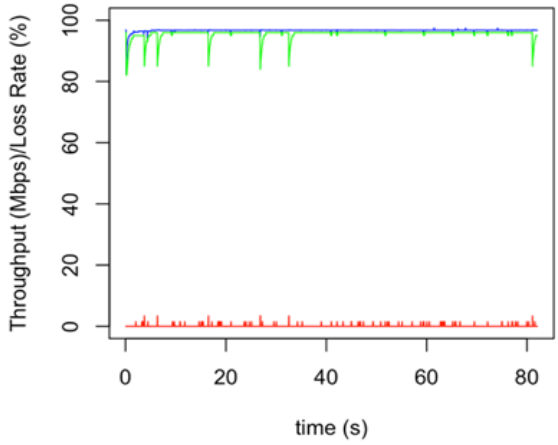


Fig. 6. Resultado de la prueba con pérdidas del 0,01% - $P1_4$

Finalmente, en la *Fig. 7*, se encuentra recogida la velocidad media de transmisión del protocolo ante los diferentes niveles de pérdidas aleatorios planteados en la Tabla I. Cabe destacar que la generación de pérdidas es arbitrária y generada por un software, por lo que, a pesar de modificar el protocolo su comportamiento para evitar pérdidas en el enlace, no se modifica el % de pérdidas producidas. El objetivo principal de la prueba es mostrar la resiliencia ante pérdidas y no el rendimiento ante éstas.

En dicha gráfica se puede observar como la zona de trabajo óptima del protocolo se encuentra en el rango 0%

– 0,1%, donde la velocidad de transmisión no se ve apenas afectada por las pérdidas.

Hasta aproximadamente el 1% de pérdidas no llega a valores inferiores al 50% de *throughput* y en pérdidas de un 5% solamente se ha reducido hasta el 10% del *throughput* total.

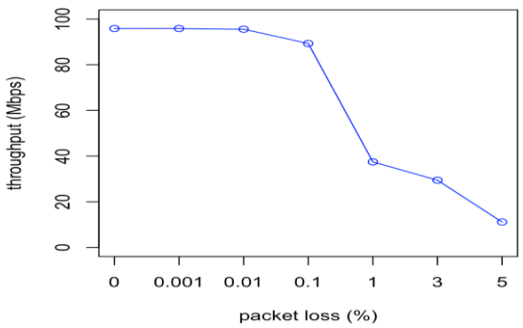


Fig. 7. Comparativa de la velocidad de transmisión media ante diferentes niveles de pérdidas (%)

De esta primera prueba se extrae la capacidad del protocolo para ceñirse al ancho de banda estimado en el proceso de estado de la red. Esto permite alcanzar el máximo ancho de banda disponible de manera rápida con valores de pérdidas aleatorias cercanas al 1%, confirmando el objetivo (O1).

Prueba 2

Mediante la prueba (*P2*), tal y como ha sido expuesto, se pretende verificar (O2). Con tal propósito, se plantea una prueba dividida en diversas etapas, dejando entre etapas un tiempo de recuperación.

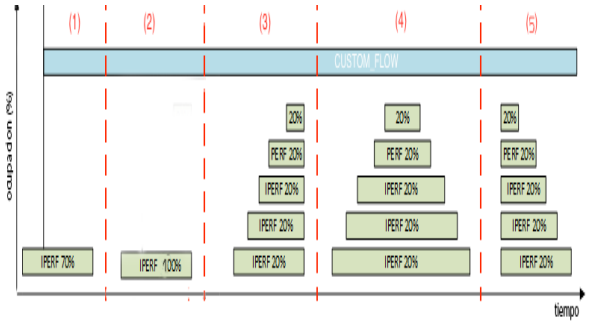


Fig. 8. Secuencia de ejecución dividida en tramos de P2

La secuencia de ejecución de flujos (*Fig. 8*) se describe a continuación:

(P2.1) Se inicia la prueba con una ocupación inicial de enlace del 70%. Se inicia un flujo poco después.

(P2.2) Se añade un tráfico interferente transmitiendo al 100%

(P2.3) Se introduce e incrementa de manera progresiva un flujo interferente en *steps* del 20% de ocupación hasta llegar al 100%

(P2.4) Se incrementa de manera progresiva un flujo interferente en *steps* del 20% de ocupación hasta llegar al 100% y se decrementa hasta llegar al 0%.

(P2.5) Se decrementa de manera progresiva el tráfico interferente de 100% a 0% de ocupación, en steps del 20%.

En la *Fig. 9* se pueden observar los resultados de (*P2*), identificando los tramos definidos anteriormente.

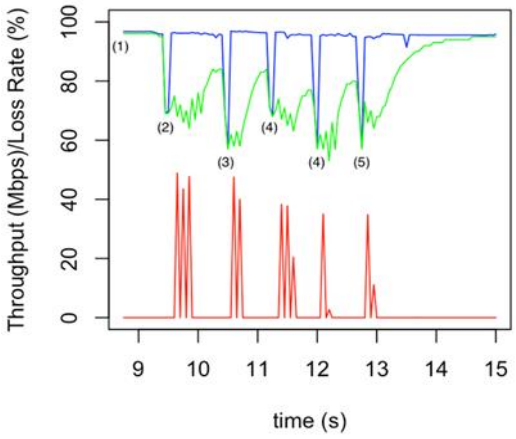


Fig. 9. Resultados de (P2)

Observando los resultados, la reactividad a las pérdidas se puede descomponer en tres fases (entre los segundos 10 y 12), observables en la *Fig. 10*.

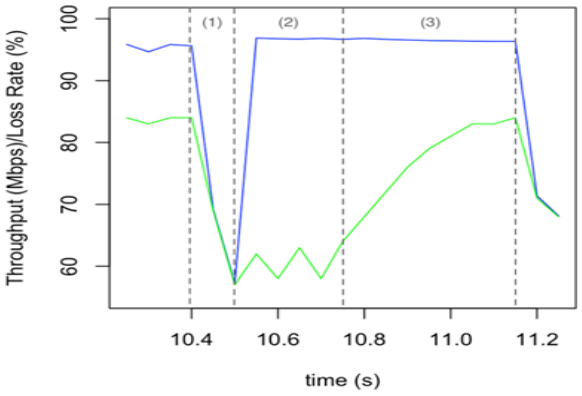


Fig. 10. Descomposición en fases de la reacción a pérdidas

En un primer lugar, se disminuye la velocidad de envío inicial hasta que la velocidad de envío agregada entre el flujo interferente y el flujo generado deja de provocar congestión (*Fig. 10* - (1)).

El control de congestión detecta nuevamente que el ancho de banda disponible es el máximo del canal. En ese punto (segundo 10,5), se entra en una segunda fase en la que se experimentan pérdidas generadas por el período de congestión anterior (*Fig. 10* - (2)). Por otro lado, en esa misma fase, dado que la estimación del enlace ya no detecta congestión, se incrementa la ventana de envío.

Estos dos aspectos, generan una variación en la velocidad de envío alrededor del valor en que se finaliza la primera fase de pérdidas. En ciertas situaciones,

cuando estas son considerablemente elevadas, la velocidad de envío decrementa proporcionalmente a estas. Finalmente, una vez se han tratado las pérdidas, se entra en recuperación (*Fig. 10* - (3)).

Se observa que los mecanismos reaccionan tanto a la congestión como a las pérdidas. La congestión provoca una disminución de la velocidad de envío, la cual se produce de manera continuada hasta descongestionar el enlace. Las pérdidas provocan disminuciones puntuales que previenen que el mecanismo entre en recuperación cuando el canal aún no está preparado. Éste también es capaz de recuperar la velocidad de envío en cuanto el canal lo permite en tiempos inferiores a medio segundo, demostrando la adaptabilidad del control de congestión.

Prueba 3

Con tal de demostrar (O3) se plantean 3 comparativas. Para ello, se muestra el comportamiento de un flujo dotado con los mecanismos diseñados al compartir el enlace con otros protocolos de transporte.

En un primer lugar (O3.1), se comprueba cómo se comporta ante un flujo agresivo como es un flujo UDP sin sistema de control de congestión. Los resultados se pueden observar a continuación.

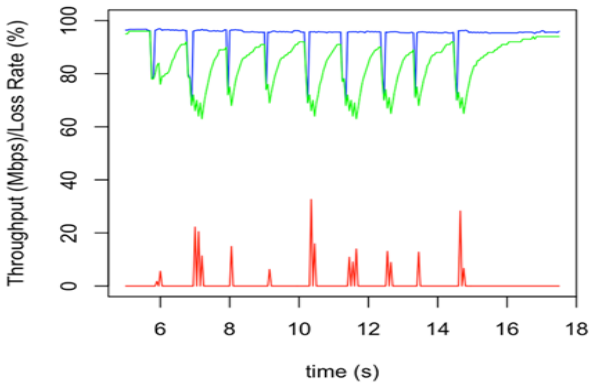


Fig. 11. Rendimiento en comparativa entre UDP

En la *Fig. 11* se puede observar el rendimiento de un flujo ante la presencia de un flujo UDP constante a 25Mbps. En estas circunstancias, el control de congestión no es capaz de ocupar todo el canal dado que el flujo con el que lo comparte es de elevada agresividad. Dicha situación genera pérdidas considerables que hacen que, aunque en promedio la velocidad de envío media sea de aproximadamente 75Mbps (el ancho de banda libre del canal), el gran número de retransmisiones a realizar implican una disminución significativa de la velocidad real de transmisión de datos. Esto es debido a que el flujo UDP se muestra invariable ante las pérdidas por saturación del enlace.

El comportamiento observado ante flujos UDP agresivos es coherente con el planteamiento del protocolo (obtener el mayor ancho de banda posible en cada momento, siendo agresivo al compartir el enlace).

Uno de los puntos a estudiar es una mayor estabilización ante flujos que no implementan control de congestión y que, por lo tanto, no varían su velocidad de envío a lo largo de la transmisión.

En contraste a dicha situación, se encuentran los resultados mostrados en la *Fig. 12* y *Fig. 13* representando el comportamiento del protocolo al compartir el enlace con un flujo *TCP friendly* (O3.2).

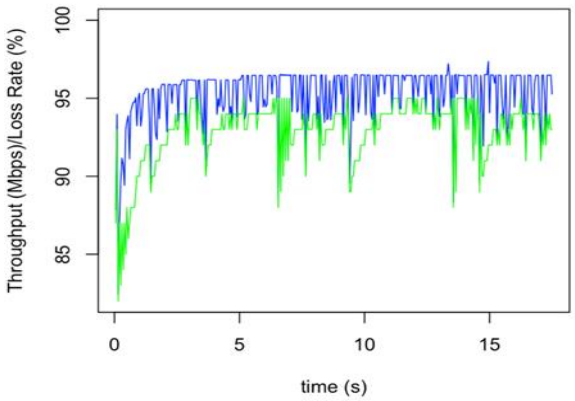


Fig. 12. Rendimiento MBTAP en comparativa TCP Friendly vs UDP-modified (TCP primero)

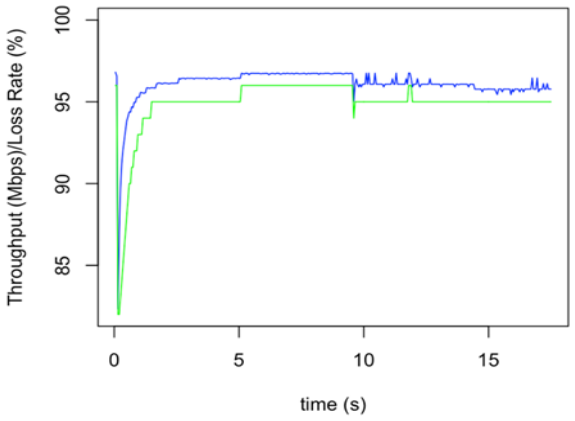


Fig. 13. Rendimiento en comparativa TCP Friendly vs UDP-modified (UDP-modified primero)

Tanto en la *Fig. 12*, se inicia un flujo *TCP Friendly* cuando el flujo de UDP modificado ya se encuentra en transmisión, como en la *Fig. 13*, se inicia un flujo UDP modificado cuando el *TCP Friendly* ya se encuentra en transmisión, se puede apreciar como la presencia de un flujo *TCP Friendly* en el canal no comporta una disminución de la velocidad de transmisión.

De los resultados mostrados, se puede concluir que el control de congestión se encuentra en el punto de agresividad planteado en su diseño. Suficientemente agresivo para utilizar la mayor parte del enlace en detrimento de otros flujos sin llegar al nivel de agresividad de UDP.

Por último, la Fig. 14 y la Fig. 15 muestran el comportamiento al transmitir simultáneamente dos flujos con el mismo mecanismo de control de congestión (O3.3). En ellas se observa cómo existe una contienda por el ancho de banda del canal (100Mbps), logrando una repartición del 50% para cada uno pero produciéndose pérdidas.

Ante dicha situación se podría esperar que ambos flujos se estabilizarán en velocidades de envío cercanas a 50Mbps. No obstante, la velocidad en la que el flujo se

estabiliza depende de diversos factores, los cuales fluctúan durante la transmisión. Dado que el estado de la red que considera el mecanismo es independiente para cada uno de los flujos, el comportamiento de cada uno de estos es indeterminado para el otro flujo. Esto hace que ambos flujos compitan por el ancho de banda disponible sin tener en cuenta la presencia de otros flujos, con lo que se genera congestión dada la agresividad del protocolo. Así pues, la implementación actual provoca una transmisión poco estable al encontrar diversos flujos que utilizan el mismo mecanismo a través de un mismo enlace.

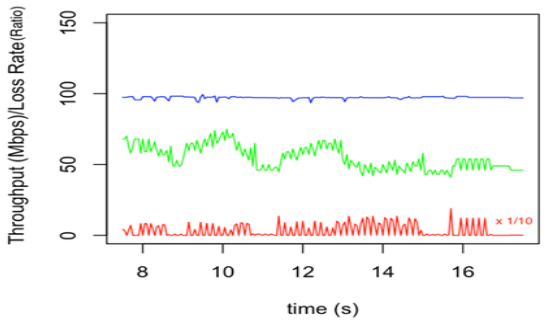


Fig. 14. Transmisión simultánea entre dos flujos con el control de congestión propuesto (A)

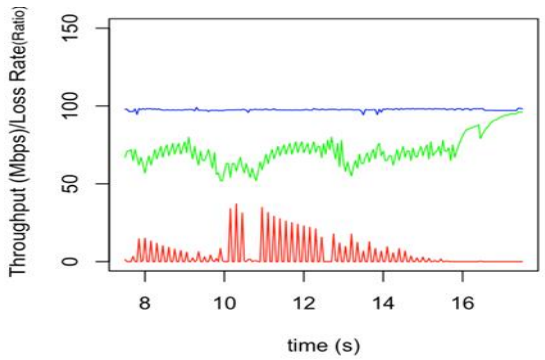


Fig. 15. Transmisión simultánea entre dos flujos con el control de congestión propuesto (B)

V. Conclusiones

El diseño de los protocolos de transporte base como son TCP y UDP no fue concebido para su uso en redes LFN. Esto es debido a que su alto ancho de banda y elevado retardo provoca ineficiencias para este tipo de comunicaciones.

Este artículo repasa el estado del arte de propuestas sobre TCP, UDP e incluso nuevos protocolos como SCTP que ofrecen soluciones eficientes, pero no definitivas.

En este sentido, se presentan dos mecanismos. El primero es el cálculo de estado de la red, el cual se realiza al inicio de la comunicación, así como durante la misma de manera *in-band* mediante el uso de los mensajes de datos. Este proceso se compone del envío de varias ráfagas de mensajes que permiten calcular el ancho de banda máximo del enlace, además de extraer otra información como el *Receiving Rate* (RR) y el *Round Trip Time* (RTT). Esto permite al segundo mecanismo, el control de congestión, iniciar la comunicación utilizando todo el ancho de banda disponible, así como adaptar el envío durante la comunicación.

Estos mecanismos combinados tienen como objetivo obtener una alta eficiencia y adaptabilidad sobre las redes LFN, así como un comportamiento agresivo frente a otros flujos.

Se ha demostrado mediante diferentes pruebas sobre un testbed de emulación WAN que estos mecanismos permiten obtener un alto ancho de banda, cercano al 96% de su capacidad total, de manera eficiente, incluso con valores de pérdidas aleatorias cercanas al 0,1%.

Además, en una de las pruebas se ha evaluado y verificado su capacidad de adaptación ante variaciones de ancho de banda disponible durante la comunicación, introduciendo diferentes flujos concurrentes en diferentes momentos de la transmisión.

Finalmente, con el fin de evaluar también el nivel de agresividad de estos mecanismos frente a otros flujos, se han realizado varias pruebas que prueban este comportamiento agresivo frente a estos (TCP, UDP y otro flujo que también implementa el mecanismo propuesto).

De cara al futuro, está previsto diseñar un protocolo de transporte que incluya estos dos mecanismos y realizar pruebas de funcionalidad integral de éste. Además, se deberían evaluar los mecanismos propuestos en entornos WAN reales y proponer alguna solución de mejora del *fairness* entre varias sesiones utilizando el mecanismo de congestión propuesto.

Referencias

[1] Jacobson, V., LBL, Braden, R., ISI, "TCP Extensions for Long-Delay Paths" October 1988

[2] Jacobson, V., Braden, B., Borman, D., Satyanarayanan, M., Kistler, J.J., Mummert, L.B. and Ebling, M.R., 1992. RFC 1323: TCP extensions for high performance.

[3] Bullot, H.; Les Cottrell, R., "Evaluation of Advanced TCP Stacks on Fast Long-Distance Production Networks" Journal of Grid Computing, 2003, Volume 1, Number 4, Page 345

[4] Hanson, R. H., Lespagnol, A., Mazraani, T. Y., Milburn, B. J., White, J. B., & Dabir, S. C. (1997). "Traffic management and congestion control for packet-based networks." U.S. Patent No. 5,633,861. Washington, DC: U.S. Patent and Trademark Office.

[5] Iren, S., Amer, P. D., & Conrad, P. T. (1999). The transport layer: tutorial and survey. ACM Computing Surveys (CSUR), 31(4), 360-404.

[6] Bashir, K., "Comparative study on advance TCP stacks and their performance analysis," 8th International Multitopic Conference, 2004. Proceedings of INMIC 2004., 2004, pp. 256-263.

[7] Kelly, T., "Scalable TCP: improving performance in highspeed wide area networks". SIGCOMM Comput. Commun. Rev. 33, 2 (April 2003), 83-91.

[8] Leith, D.; Shorten, R., "H-TCP: TCP for high-speed and long-distance networks"

[9] Nagamalai, D., Jae-Kwang Lee, (2004) "Performance of SCTP over high speed wide area networks," Cybernetics and Intelligent Systems. IEEE Conference on , vol.2, no., pp.890,895, 20046.

[10] Gu, Y., Grossman., R. L., (2007) "UDT: UDP-based data transfer for high-speed wide area networks." Computer Networks 51, no. 7: 1777-1799.

[11] Eckart, B., He, X., & Wu, Q. (2008, April). Performance adaptive UDP for high-speed bulk data transfer over

dedicated links. In Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on (pp. 1-10). IEEE.

[12] Gu, Y., R. L. Grossman. (2007) "UDT: UDP-based data transfer for high-speed wide area networks." Computer Networks 51, no. 7: 1777-1799.

[13] WANem, The Wire Area Network emulator. Version 3.0 "wanem.sourceforge.net" [Fecha de consulta: 27 abril 2017]

[14] iPerf3 - The ultimate speed test tool for TCP, UDP and SCTP https://iperf.fr [Fecha consulta: 27 abril de 2017]

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6601

Caracterización experimental del comportamiento de Network Coding para comunicaciones multicast

Pablo Garrido, Ramón Agüero
Departamento de Ingeniría de Comunicación,
Universidad de Cantabria
39005, Santander
{pgarrido,ramon}@tlmat.unican.es

***Resumen*—Las comunicaciones multicast, caracaterizadas por la existencia de una única fuente, que transmite la misma información a múltiples destinos, están llamadas a ser un ingrediente relevante en las redes de siguiente generación. Este trabajo se centra en el uso del esquema de codificación de red Random Linear Network Coding (RLNC) para ofrecer una mayor escalabilidad en ese tipo de comunicaciones, frente a otros esquemas más tradicionales, incrementando la fiabilidad extremo a extremo. Los resultados teóricos ponen de manifiesto la gran ventaja que supone el utilizar técnicas de codificación, frente al uso de retransmisiones selectivas para recuperar la información perdida. Además, se muestra la viabilidad de la solución propuesta, desplegándola sobre una plataforma experimental compuesta por múltiples dispositivos de bajo coste, Raspberry-Pi's.**

***Palabras Clave*—Network Coding, Multicast, Random Linear Network Coding, Implementación**

I. INTRODUCTION

Las comunicaciones multicast están cobrando gran importancia en los últimos años, especialmente para servicios de transmisión de contenidos en directo (*streaming*), tanto de vídeo o de audio. Sin embargo, ofrecer un servicio fiable y escalable no es trivial. Entre otras razones, establecer mecanismos para que el nodo origen conozca el estado de los múltiples receptores genera una sobrecarga importante a medida que el número de destinos aumenta.

Han surgido diversas propuestas para ofrecer servicios multicast. Entra ellas destaca el uso, por parte de los nodos destino, de reconocimientos selectivos negativos [1]. Esta es la base del protocolo NACK-Oriented Reliable Protocol (NORM), propuesto por un grupo de trabajo del IETF [2]. Sin embargo, este tipo de soluciones genera una alta sobrecarga, debido principalmente a la información transmitida por los múltiples receptores. Esta sobrecarga, en muchos casos, hace que las propuestas no sean escalables, por lo que se restringen a un número de usuarios reducido. El protocolo NORM utiliza un esquema de reconocimientos probabilístico, permitiendo así una mayor escalabilidad [3].

Por otro lado, se ha propuesto aprovechar las posibilidades de las técnicas de codificación fuente sobre redes multicast [4]. En esta aproximación, destacan las soluciones basadas en códigos LT [5] o Raptor [6]. Ambos se caracterizan por generar un número ilimitado de paquetes codificados que, con muy alta probabilidad, permiten recuperar toda la información, recuperándose ante eventuales pérdidas.

Compartiendo algunas de las características de las técnicas de codificación fuente, han ganado popularidad las soluciones de codificación de red, Network Coding (NC), propuestas inicialmente por Ashlwede *et al.* en [7]. Entre las diferentes propuestas que se han llevado a cabo, destaca el esquema Random Linear Network Coding (RLNC), propuesto por Ho *et al.* en [8]. A diferencia de la codificación fuente, los nodos intermedios de la red pueden participar de manera más activa en la comunicación, descartando o generando paquetes recodificados, lo que reduce el impacto de las pérdidas en canales multisalto [9], aprovechando además la capacidad de escucha oportunista de las redes inalámbricas [10]. Los autores de [11], [12] ya propusieron aprovechar el esquema RLNC en redes multicast, aunque en ambos casos los resultados se obtienen mediante estudios basados en simulación y ambas soluciones se basan en el uso de reconocimientos.

En este trabajo se presenta un esquema NC para redes multicast que no hace uso de reconocimientos y que permite ofrecer un servicio con una calidad adecuada. Primero se analizará la propuesta desde un punto de vista analítico, comparándola con un esquema basado en Automatic Repeat Request (ARQ), en concreto el protocolo NORM. Además, el trabajo presenta la implementación de la solución propuesta sobre una plataforma de dispositivos de bajo coste, Raspberry-Pis, sobre la que se analizarán las

prestaciones de las dos alternativas, a través de una extensa campaña de medidas.

El resto del trabajo se estructura como sigue: la Sección II describe brevemente el funcionamiento de RLNC y del protocolo NORM. Además, se desarrollan las expresiones que permitirán comparar los rendimientos teóricos de ambos esquemas. La Sección III presenta los principales resultados obtenidos. En una primera parte se realiza una comparativa teórica entre el esquema implementado y una solución basada en ARQ para posteriormente describir la plataforma experimenta desplegada y describir los resultados obtenidos tras una extensa campaña de medidas llevadas a cabo sobre la misma. Finalmente, en la Sección IV se enumeran las principales conclusiones que pueden extraerse del trabajo, indicando una serie de líneas de investigación que quedan abiertas, tanto sobre el estudio del esquema RLNC como en la mejora y extensión de la plataforma.

II. PRELIMINARES

En esta sección se expondrá brevemente la operación básica del esquema RLNC, así como el funcionamiento del protocolo NORM. Además, se introducirán expresiones que permitirán analizar teóricamente el comportamiento de ambas soluciones, que serán empleadas posteriormente a la hora de comparar ambos esquemas.

A. *Random Linear Coding (RLNC)*

El esquema RLNC fue originalmente presentado en [8] y en este trabajo se hace uso de un esquema similar al utilizado en [13]. El nodo origen divide la información a transmitir en generaciones. Cada generación contiene un número paquetes, k, y cada paquete contiene L bytes, que coincide con la MTU de la tecnología de red que se esté usando. El nodo origen transmite paquetes codificados generados como una combinación aleatoria de paquetes que pertenezcan a la misma generación:

$$p' = \sum_{i=0}^{k} c_i \cdot p_i \tag{1}$$

donde c_i se corresponde con los coeficientes seleccionados de forma aleatoria en un cuerpo finito de Galois, $GF(2^q)$. Estos coeficientes se pueden representar como un vector, $\bar{c}$, que será incluido en una cabecera, sin codificar, en cada paquete p' transmitido.

En el nodo destino se mantiene una matriz de decodificación D de tamaño $k \times k$. Por cada paquete codificado recibido se extrae de la cabecera el vector de coeficientes, que se inserta en D (en la fila i, de acuerdo al rango actual de la matriz) y se comprueba si dicho vector es linealmente independiente de los anteriormente recibidos (el rango de la matriz se vería incrementado). Si no fuera así el paquete recibido se descartaría, dado que no proporciona información novedosa. Una vez que el destino ha recibido k paquetes codificados linealmente independientes, es decir la matriz de decodificación tiene rango k, el destino está en disposición de recuperar los paquetes originales.

Para calcular el número de paquetes que el destino debe recibir para poder decodificar una generación se debe conocer primero la probabilidad de que un paquete codificado recibido sea linealmente independiente de los anteriormente recibidos. Esta probabilidad depende de la cantidad de información recibida hasta el momento o, lo que es lo mismo, el rango de la matriz de decodificación y viene dada por la siguiente expresión:

$$\text{Prob}_{r^+}^{\text{RLNC}} = 1 - \frac{(2^q)^r}{(2^q)^k} \tag{2}$$

El número medio de paquetes codificados que debería recibir el nodo destino para poder decodificar una generación viene dado por:

$$\overline{\#\text{TX}} = \sum_{i=0}^{k-1} \frac{1}{\text{Prob}_{r^+}^{\text{RLNC}}} = k + \alpha \tag{3}$$

donde α es una constante que no depende del tamaño de la generación pero sí del cuerpo de Galois elegido, como se demuestra en [14] y, que en el caso binario ($GF(2)$), $\alpha \approx 1{,}6$, para ir aproximándose a cero a medida que el tamaño del cuerpo crece. Hay que destacar que bajo esquemas de codificación RLNC no es importante qué paquetes han sido recibidos, sino recibir suficientes paquetes de información para decodificar. En este aspecto se encuentra la principal ventaja de utilizar esquemas de codificación como RLNC sobre redes multicast, donde cada destino debe recibir aproximadamente $k+\alpha$ paquetes codificados (k linealmente independientes) para poder recuperar la información, sin importar exactamente cuales, ya que cada uno de ellos transporta la misma información. Se evita así la necesidad de transmitir notificaciones sobre qué paquetes en concreto falta a cada nodo destino.

En el esquema implementado el nodo origen transmite por cada generación un número de paquetes extra, $N = k\,(1+r)$, donde r es un factor de redundancia que se deberá configurar según las condiciones particulares de cada caso (canal y grado de servicio). La probabilidad de que un receptor cualesquiera pueda decodificar una generación tras haber recibir N' paquetes codificados viene dada por la siguiente expresión, presentada por Trullols en [15]:

$$\begin{aligned}\xi_q(k,N) = \xi_q^0 \Bigg(& \begin{bmatrix} N \\ N-k \end{bmatrix}_{2^q} + \\ & + \sum_{i=1}^{N-k} (-1)^i \binom{N}{i} \begin{bmatrix} N-i \\ N-k-i \end{bmatrix}_{2^q} \Bigg)\end{aligned} \tag{4}$$

donde $\begin{bmatrix} m \\ n \end{bmatrix}_q$ son los coeficiente q-binomiales (o Gauss) [15] y ξ_q^0 es la probabilidad de decodificar una generación tras recibir exactamente k paquetes codificados:

$$\xi_q^0 = \xi_q(k,k) = \frac{(2^q)^{k^2}}{((2^q)^k - 1)^k} \prod_{j=1}^{k} \left(1 - \frac{1}{(2^q)^j}\right) \tag{5}$$

Se asume un canal entre el nodo origen y el nodo destino, i, con una Frame Error Rate (FER) conocida, FER_i y que las pérdidas siguen una distribución uniforme. Bajo estas condiciones, y siguiendo la expresión de Trullols (Eq. (4)), se obtiene la probabilidad de decodificar una generación tras transmitir N paquetes por el nodo origen:

$$\text{Prob}_{\text{dec}} = \sum_{i=k}^{N} \binom{N}{i} \text{FER}_i^j \times (1 - \text{FER}_i)^{N-i} \times \xi_q(k, N) \tag{6}$$

Finalamente, la probabilidad de que un nodo destino, i, reciba toda la información (por ejemplo, un fichero), compuesta por M generaciones, es:

$$\text{Prob}_{\text{Succes}} = (\text{Prob}_{\text{dec}})^M \tag{7}$$

esta probabilidad es independiente del número de dispositivos que participen en la red multicast. Se debe escoger un valor de redundancia adecuado a las peores condiciones, teniendo en cuenta que un exceso de redundancia perjudica el rendimiento de la red, ya que origina un mayor número de transmisiones.

B. NACK-oriented Reliable Protocol (NORM)

El protocolo NORM está diseñado para ofrecer una transmisión fiable para uno o más destinos sobre una red IP multicast. El objetivo es ofrecer una solución escalable y robusta sobre redes heterogéneas. Para ello se basa en el uso de Negative ACKnowledgment (NACK)s selectivos por parte de los destinos, que se utilizan para recuperar los paquetes perdidos.

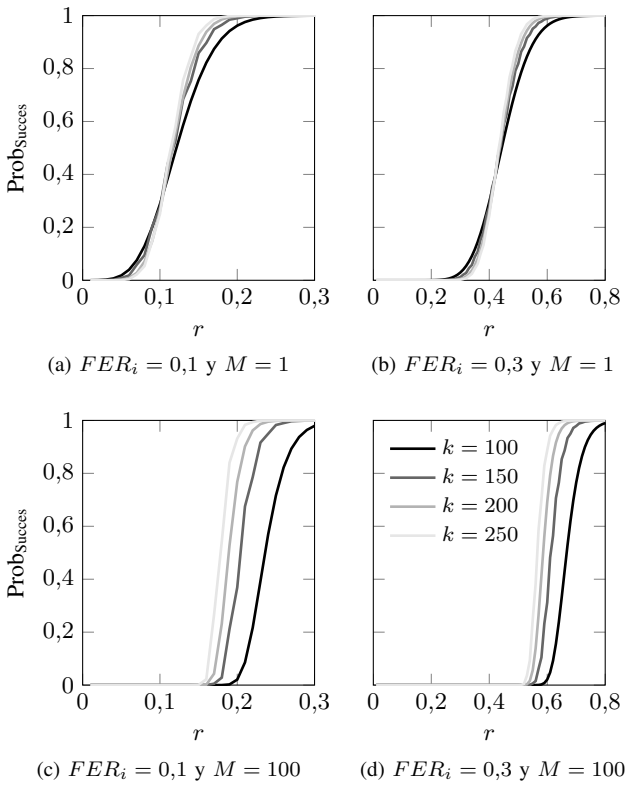
La principal limitación es el volumen de tráfico generado por las peticiones de información no recibida por los destinos. Esta sobrecarga incrementa linealmente con el número de dispositivos, generando un número elevado de transmisiones hacia el nodo origen. Para reducir este tráfico NORM utiliza un mecanismo de reconocimientos probabilístico [3], que le permite incrementar la escalabilidad manteniendo un nivel adecuado de fiabilidad. Para más detalles sobre el protocolo, el lector puede acudir a [2].

Se asume que el nodo origen quiere transmitir un fichero de tamaño, $T = M \times k$, y que la calidad del canal entre el nodo origen y los destinos es conocida e igual para todos, $\overline{FER}$. Bajo estas circunstancias es fácil establecer que la probabilidad de que un paquete llegue a todos los destinos (R) es $(1 - FER)^R$. Por tanto, la probabilidad de que se hayan perdido X paquetes se puede obtener mediante una distribución binomial, con probabilidad $p = 1 - (1 - FER)^R$:

$$\begin{aligned}\text{Prob}_{\text{lost}}(X, T) = \binom{L}{X} (1 - (1 - FER)^R)^X \times \\ \times ((1 - FER)^R)^{(T-X)}\end{aligned} \tag{8}$$

El número medio de paquetes que se necesitan retransmitir bajo un esquema NORM es, por tanto:

$$\overline{\text{TX}_{\text{NORM}}} = T \times (1 - (1 - FER)^R)^X \tag{9}$$



(a) $FER_i = 0{,}1$ y $M = 1$ (b) $FER_i = 0{,}3$ y $M = 1$

(c) $FER_i = 0{,}1$ y $M = 100$ (d) $FER_i = 0{,}3$ y $M = 100$

Figura 1: Probabilidad de éxito en función de la redundancia en el sistema y para diferentes tamaños de generación. Se compara el resultado para diferentes probabilidades de error en el canal.

III. RESULTADOS

En esta sección se comparará incialmente el rendimiento del esquema RLNC con el del protocolo NORM, utilizando las expresiones teóricas que se han presentado en la Sección II. Posteriormente se describirá la plataforma que se ha desplegado, mediante el uso de Raspberry-Pi's, para evaluar el comportamiento de ambas soluciones sobre equipos reales. Se finalizará describiendo los resultados obtenidos tras una extensa campaña de medidas.

A. Análisis teórico

Un primer análisis del esquema RLNC propuesto se muestra en la Figura 1, que representa la probabilidad de éxito, definida como la probabilidad de que un receptor cualquiera haya podido decodificar todas las generaciones. Se comparan los valores obtenidos para diferentes configuraciones, modificando tanto el tamaño de la generación como la probabilidad de error en el canal ($FER_i = 0{,}1, 0{,}3$). Evidentemente, la redundancia necesaria vendrá marcada fuertemente por la probabilidad de error en el canal. Además, cuanto mayor sea el número de generaciones a recuperar, mayor será también la redundancia, tal y como se puede ver en la expresión 7. Aumentar el tamaño de las generaciones, para transmitir así la misma cantidad de información, reduciendo el número de generaciones necesarias, disminuiría la redundancia necesaria. Sin embargo, es importante tener en cuenta asimismo la sobrecarga

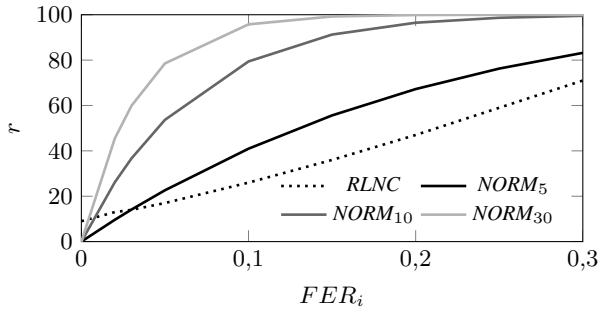


Figura 2: Redundancia que se debería utilizar bajo el esquema RLNC para obtener una probabilidad de 99,99 % de éxito en comparativa con un esquema simplificado de NORM

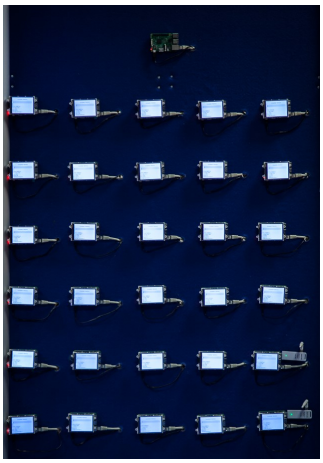


Figura 3: Imagen de la plataforma desplegada. Transmisor en la parte superior y 30 receptores en la matriz de 6x5

generada por el vector de coeficientes que, según se analizó en [13], no debería superar un tamaño de $k = 255$ para un escenario en el que la Maximum Transfer Unit (MTU) sea de 1500 bytes, puesto que este vector tendría un tamaño de $\frac{k \cdot q}{8}$ bytes.

En la Figura 2 se comparan el esquema implementado RLNC con una versión simplificada de NORM, tal y como se describe en la Sección B,(Eq. 9). Bajo la leyenda RLNC se representa la redundancia con la que se obtendría una probabilidad de éxito del 99,99 %, utilizando la expresión (Eq. 7). En ambos caso la información a transmitir es la misma, 100 generaciones (M) con 100 paquetes cada generación (k), lo que hace un total de 100×100 paquetes (L). En el caso de utilizar un esquema RLNC, esta redundancia sería la misma, independientemente del número de dispositivos, siempre que los canales presenten una calidad mejor que FER_i. En cambio, esto no ocurre para el caso del esquema NORM, donde el número de receptores claramente afecta al número de retransmisiones. Como se puede ver en la Figura 2, la redundancia bajo un esquema de retransmisiones selectivas como NORM es hasta 4,5 veces mayor que un esquema de codificación RLNC.

B. *Plataforma*

Con intención de estudiar la aplicabilidad del esquema RLNC en entorno reales, se ha desplegado una plataforma compuesta por dispositivos de bajo coste, Raspberry-Pis. En la Figura 3 se puede ver el despliegue de la plataforma. En concreto, se han dispuesto 31 Raspberry-Pis sobre un tablero, 30 de ellos desplegados según una malla, que harán las veces de receptores multicast. Todos los dispositivos cuentan con un interfaz inalámbrico (802.11) y otro Ethernet, que se utiliza para realizar todas las operaciones de gestión del experimento: comunicarse con cada uno de los dispositivos, inicializarlos, recoger los datos tras las medidas, etc. El interfaz inalámbrico es el único que se utilizará durante los experimentos.

La red está configurada en modo infraestructura, de manera que el nodo superior hace las funciones de punto de acceso del resto de dispositivos, además de tomar el papel de transmisor de la red multicast. Esta configuración permite tener mayor control sobre el tráfico que se envía por la red. El resto de dispositivos, situados según una malla de 6×5, son los nodos receptores que, además, cuentan con una pantalla que muestra las estadísticas de cada experimento/medida. Todos los dispositivos son Raspberry-Pi 3, que cuentan con interfaz 802.11, excepto los dos situados en la parte inferior derecha, que son Raspberry-Pi 2, que utilizan un dongle TP-Link TLWN722N.

El esquema RLNC implementado utiliza la librería KODO [16] tal como está detallado en la Sección II. En concreto, el nodo transmisor, dispositivo en la parte superior del panel, transmite $N = k \times r$ paquetes codificados por cada generación en modo broadcast. Los receptores, resto de dispositivos de la malla, reciben los paquetes que no se hayan perdido debido a interferencias y las condiciones adversas del canal inalámbrico. Los receptores procederán a recuperar la generación siempre que sea posible. En caso de recibir paquetes que pertenecen a la siguiente generación, la generación queda identificada en la cabecera de cada paquete codificado, marcará la transmisión como incompleta, es decir, no podrá recibir el fichero en su totalidad, pero intentará decodificar la siguiente generación.

Hay varios trabajos que desplegaron con anterioridad plataformas con dispositivos de bajo coste para el estudio de las prestaciones que ofrecen esquemas de Network Coding [17], [18]. En ambos se utilizan Raspberry-Pis como dispositivos de bajo coste y se utiliza la librería KODO para las tareas codificación y decodificación. En [17] se estudia las ventajas de utilizar esquemas de Network Coding en redes con cooperación, es decir, donde nodos intermedios de la red cooperan entre sí para recuperar la información que no han recibido otros nodos de la red. Mientras que en [18] también se analiza las prestaciones del esquema RLNC para comunicaciones multicast inalámbricas. Sin embargo, no presentan un estudio analítico ni comparan con un esquema alternativo para comunicaciones multicast como es NORM.

Además de analizar el comportamiento bajo condiciones adversas, cada dispositivo cuenta con un archivo de

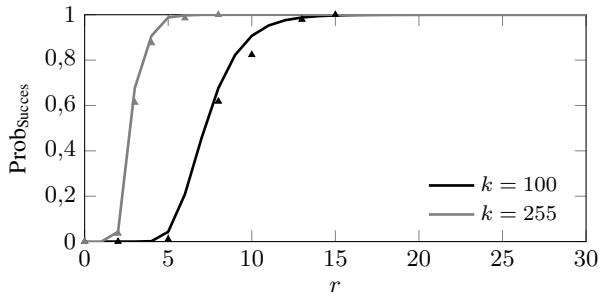
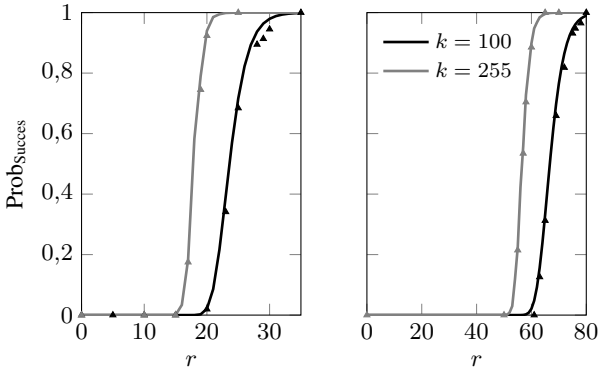


Figura 4: Probabilidad que un dispositivo reciba todas las generaciones. Las líneas se corresponden con el valor teórico, mientras que los marcadores reflejan los valores observados sobre la plataforma de medida.



(a) $FER = 0{,}1$ (b) $FER = 0{,}3$

Figura 5: Probabilidad que un dispositivo reciba todas las generaciones con canales con error. Las líneas se corresponde con el valor teórico, mientras que los marcadores reflejan los valores observados sobre la plataforma de medida

configuración donde se puede establecer fijar una probabilidad de error 'sintética', por la que se descartarían, en el dispositivo i, paquetes de manera aleatoria, con una probabilidad FER_i.

Todos los resultados que se muestran a continuación se obtienen tras promediar los valores obtenidos tras 100 experimentos. La cantidad de información transmitida en cada experimento es de 100 generaciones.

En la Figura 4 se compara la probabilidad de éxito teórica frente a la que se obtiene en las medidas. Se ha utilizado para ello, como probabilidad de error en los resultados teóricos (FER_i^j en (Eq. 6)), la media observada en el conjunto de dispositivos. Como se puede ver, hay una pequeña diferencia entre los dos tipos de resultado, lo que puede justificarse por la gran variabilidad del medio físico, en el que, además, las pérdidas se caracterizan por no seguir una distribución uniforme y habitualmente aparecen a 'ráfagas', por ejemplo tras la ejecución de algún proceso de búsqueda activa por parte de los dispositivos activos en el área de influencia de la plataforma.

Como se mencionó anteriormente, los nodos receptores cuentan con un archivo de configuración, que permite establecer una probabilidad de error sintética en cada nodo. En la Figura 5 se representa la probabilidad de éxito cuando todos los nodos están configurados con la misma FER (0,1 ó 0,3). El sistema tiene un comportamiento muy similar, independientemente de la calidad de canal, aunque se ve claramente que la redundancia aumenta a medida que la calidad de los canales inalámbricos es peor.

IV. CONCLUSIONES

En este trabajo se ha presentado un sistema basado en el esquema RLNC para ofrecer un servicio de conexión fiable y escalable para dar servicios multicast en redes inalámbricas. Primero se han planteado expresiones analíticas que modelan el comportamiento de esta solución, así como del correspondiente protocolo NORM. Se puso de manifiesto que, con un esquema basado en reconocimientos selectivos, la sobrecarga total aumenta con el número de receptores en la red. Los resultados muestran que la sobrecarga puede ser hasta 4,5 veces mayor, si se compara con la solución propuesta.

Mediante el despliegue de una plataforma de dispositivos de bajo coste, Raspberry-Pi's, se ha podido analizar la viabilidad del esquema propuesto, a través de una extensa campaña de medidas, cuyos resultados muestran un comportamiento muy similar al adelantado por el modelo teórico.

Gracias a la realización de este trabajo surgen varias líneas de trabajo que se esperan continuar en el futuro:

- Análisis de esquemas de reconocimiento probabilísticos para mejorar la fiabilidad del sistema. Aunque el uso de retransmisiones implica un incremento de la sobrecarga, con en el esquema RLNC solo se necesitan reconocimientos por cada generación, lo que ya reduciría considerablemente la sobrecarga.
- Transmisión de vídeo sobre la plataforma. La retransmisión de contenido multimedia (*streaming*) es, sin duda, una de las aplicaciones más relevantes para las redes multicast. Se utilizará la plataforma desplegada para reproducir vídeo transmitido por el nodo transmisor en las pantallas de los receptores. En estas condiciones las consideraciones son diferentes, ya que un tamaño de generación elevado daría lugar a un retardo inaceptable en la recepción mientras que la pérdida de una generación no es tan relevante, ya que lo que determina la calidad del servicio es la continuidad del vídeo en el receptor.
- Redes multi-salto. También se analizará el comportamiento del esquema RLNC cuando se utilice sobre topologías multi-salto. Como ya se ha mencionado, RLNC ofrece la posibilidad de que los nodos intermedios descarten o generen paquetes recodificados. En varios trabajos previos [19], [20] ya se analizaron estas soluciones, en un entorno de simulación, por lo que resultaría interesante replicar dichos experimentos aprovechando la plataforma desplegada, para comprobar su viabilidad sobre dispositivos reales.

ACKNOWLEDGEMENTS

Los autores querían expresar su agradecimiento al Gobierno de España por su financiación en el proyecto "*Aprovisionamiento Dinámico de Conectividad en Escenarios inalámbricos 5G de alta Densidad* **ADVICE** (TEC2015-71329-C2-1-R).

REFERENCIAS

[1] N. Seddigh, B. Nandy, and J. Salim, "System and method for a negative acknowledgement-based transmission control protocol," Apr. 25 2006, uS Patent 7,035,214. [Online]. Available: https://www.google.com/patents/US7035214

[2] B. Adamson, C. Bormann, M. Handley, and J. Macker, "Negative-acknowledgment (nack)-oriented reliable multicast (norm) protocol," Tech. Rep., 2004.

[3] J. Nonnenmacher and E. W. Biersack, "Optimal multicast feedback," in *INFOCOM '98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3, Mar 1998, pp. 964–971 vol.3.

[4] D. Vukobratovic, V. Stankovic, D. Sejdinovic, L. Stankovic, and Z. Xiong, "Scalable video multicast using expanding window fountain codes," *IEEE Transactions on Multimedia*, vol. 11, no. 6, pp. 1094–1104, Oct 2009.

[5] M. Luby, "LT codes," in *Proc. of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002, pp. 271–280.

[6] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[7] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, Jul 2000.

[8] T. Ho, R. Koetter, M. Medard, D. Karger, and M. Effros, "The benefits of coding over routing in a randomized setting," in *Proc of the IEEE International Symposium on Information Theory*, 2003, p. 442.

[9] P. Pahlevani, D. E. Lucani, M. V. Pedersen, and F. H. P. Fitzek, "PlayNCool: Opportunistic network coding for local optimization of routing in wireless mesh networks," in *Proc. of the IEEE Globecom Workshops (GC Wkshps)*, Dec 2013, pp. 812–817.

[10] S. Chachulski, M. Jennings, S. Katti, and D. Katabi, "Trading structure for randomness in wireless opportunistic routing," in *Proc. of the ACM conference on Applications, techonologies, architectures and protocols for computer communications*, vol. 37, no. 4. ACM, Aug. 2007, pp. 169–180.

[11] X. Xiao, L. M. Yang, W. P. Wang, and S. Zhang, "A wireless broadcasting retransmission approach based on network coding," in *Proc. of the 4th IEEE International Conference on Circuits and Systems for Communications*, May 2008, pp. 782–786.

[12] D. Nguyen, T. Tran, T. Nguyen, and B. Bose, "Wireless broadcast using network coding," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 914–925, Feb 2009.

[13] D. Gómez, E. Rodríguez, R. Agüero, and L. Muñoz, "Reliable communications over lossy wireless channels by means of the combination of UDP and random linear coding," in *Proc. of the IEEE Symposium on Computers and Communications (ISCC)*, June 2014, pp. 1–6.

[14] P. Garrido, D. E. Lucani, and R. Aguero, "A markov chain model for the decoding probability of sparse network coding," *IEEE Transactions on Communications*, vol. PP, no. 99, pp. 1–1, 2017.

[15] O. Trullols-Cruces, J. Barcelo-Ordinas, and M. Fiore, "Exact decoding probability under random linear network coding," *IEEE Communications Letters*, vol. 15, no. 1, pp. 67–69, January 2011.

[16] M. V. Pedersen, J. Heide, and F. Fitzek, *Kodo: An Open and Research Oriented Network Coding Library*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 145–152.

[17] H. Khamfroush, D. E. Lucani, J. Barros, and P. Pahlevani, "Network-coded cooperation over time-varying channels," *IEEE Transactions on Communications*, vol. 62, no. 12, pp. 4413–4425, Dec 2014.

[18] A. Paramanathan, P. Pahlevani, S. Thorsteinsson, M. Hundeboll, D. E. Lucani, and F. H. P. Fitzek, "Sharing the pi: Testbed description and performance evaluation of network coding on the raspberry pi," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–5.

[19] P. Garrido, D. E. Lucani, and R. Aguero, "How to tune sparse network coding over wireless links," *Proc. of the IEEE International Wireless Communications and Networking Conference*, 2017.

[20] D. Gomez, P. Garrido, E. Rodriguez, R. Aguero, and L. Munoz, "Enhanced opportunistic random linear source/network coding with cross-layer techniques over wireless mesh networks," in *Proc. of the IFIP Wireless Days Conference (WD)*, Nov 2014, pp. 1–4.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6610

Plataforma extremo-a-extremo compatible con el estándar HbbTV 2.0 para la TV híbrida y multi-dispositivo

Dani Marfil, Fernando Boronat, Mario Montagud, Pau Salvador.
Departamento de Comunicaciones, Immersive Interactive Media (IIM) R&D Group,
Universitat Politècnica de València - Campus de Gandia
C/ Paranimf, 1 46730 Grau de Gandia (Valencia).
{damarre@dcom., fboronat@dcom., mamontor@, pasallla@epsg.}upv.es

***Resumen*- En este artículo se presenta una plataforma extremo-a-extremo para la generación, entrega y sincronización de contenidos híbridos (*broadcast* y *broadband*) relacionados, tanto en un mismo como en múltiples dispositivos (ej. escenario multi-pantalla). La plataforma es compatible con la versión más reciente del estándar *Hybrid Broadcast Broadband TV* (2.0.1) y, además, incluye soluciones eficientes para aspectos clave que no están especificados en dicho estándar, pero que son necesarios para la implantación satisfactoria de servicios de TV híbridos y multi-dispositivo, como: mecanismos de señalización para descubrir, asociar y describir los contenidos (híbridos) adicionales relacionados, mecanismos de interacción y negociación entre los dispositivos disponibles y soluciones de sincronización multimedia adaptativa (incluyendo protocolos, algoritmos y técnicas de ajuste) para alinear en el tiempo, de manera precisa, la presentación de los contenidos disponibles. La plataforma se ha evaluado objetiva y subjetivamente para el caso de uso de la TV multi-vista y multi-dispositivo, obteniendo resultados satisfactorios y prometedores.**

***Palabras Clave*- HbbTV, Sincronización Híbrida, IDES, Sincronización Multimedia, Escenario Multi-Dispositivo, TV Multi-Vista, DVB, DASH, QoE**

I. INTRODUCCIÓN

En la actualidad, una gran variedad de tecnologías de distribución de contenidos, de dispositivos de consumo, así como de contenidos multimedia, está al alcance de los consumidores. En cuanto a las tecnologías de distribución, los contenidos multimedia se pueden enviar bien a través de redes de radiodifusión, en adelante, redes *broadcast*, como, por ejemplo, DVB (*Digital Video Broadcasting*), o bien a través de redes IP, en adelante, redes *broadband*, como, por ejemplo, Internet.

Por una parte, las tecnologías *broadcast* pueden proporcionar el mismo contenido multimedia a un gran número de consumidores. Por otra parte, las tecnologías *broadband* pueden proporcionar servicios bidireccionales, interactivos y adaptativos, en base a las preferencias y recursos de los consumidores. Sin embargo, suelen presentar un rendimiento más pobre en términos de escalabilidad, estabilidad y latencia, en comparación con las tecnologías *broadcast*. En este contexto, el contenido puede ser entregado empleando diversas variantes de descarga y de transmisión (*streaming*) de contenidos, siendo esta última opción la que ha ganado mayor popularidad en los últimos años. Respecto a tecnologías o servicios de *streaming*, pueden distinguirse dos alternativas principales: gestionados y no gestionados [1]. Los servicios gestionados operan típicamente en entornos controlados como, por ejemplo, IPTV (*Internet Protocol TV*), y se basan principalmente en *push-based streaming*, empleando los protocolos RTP/RTCP (*Real-time Transport Protocol / RTP Control Protocol*) [2]. Dichos protocolos son especialmente apropiados para servicios interactivos, en los que el retardo no debe superar ciertos umbrales. Los servicios no gestionados pueden operar en entornos de área extensa y están típicamente basados en *pull-based streaming*, utilizando HTTP (*HyperText Transfer Protocol*) como mecanismo de descarga adaptativa, siendo típicamente conocidos como HAS (*HTTP-based Adaptive Streaming*). Las principales ventajas que ofrecen son la adaptabilidad, escalabilidad, fiabilidad, ubicuidad y

eficiencia en el coste. En este contexto, diferentes empresas y organismos de estandarización han especificado su propia solución HAS, entre las cuales destacan: *HTTP Live Streaming* (HLS) de Apple; *MPEG Dynamic Adaptive Streaming over HTTP* (DASH) de ISO/IEC y MPEG [3]; y *Smooth Streaming* de Microsoft (MSS). Las soluciones HAS están siendo mejoradas continuamente y ampliamente adoptadas para la distribución de contenidos multimedia en redes *broadband*. Como prueba de evidencia, DASH ha sido adoptado por el estándar *Hybrid Broadcast Broadband TV* (HbbTV) [4] y está siendo adoptado por servicios y plataformas muy populares.

De algún modo, las tecnologías *broadcast* y *broadband* se han convertido en rivales en el mercado actual de distribución y consumo de contenidos multimedia. Sin embargo, la inter-operabilidad, coordinación y convergencia entre ambas tecnologías, en combinación con la amplia disponibilidad de dispositivos multi-conectados, abre la puerta a nuevas posibilidades de innovación y oferta de servicios multimedia enriquecidos. Algunos ejemplos de estos servicios son: el consumo en paralelo de diferentes flujos de vídeo (ej. el modo *Picture-In-Picture* -PiP- o el modo mosaico), provisión de escalabilidad espacial, temporal o de color [5], *tiled-streaming* (ej. distribución de vídeo UHD -*Ultra High Definition*-, donde diferentes áreas espaciales del mismo vídeo se entregan a través de diferentes flujos) [6], selección personalizada de flujos de audio (ej. sustitución del audio del contenido *broadcast* por uno proveniente de una emisora de radio *on-line*), etc.

Debido al potencial que ofrece la coordinación y convergencia entre las tecnologías *broadcast* y *broadband*, se persigue disponer de un ecosistema multimedia híbrido en el que ambas tecnologías se complementen y aumenten su valor unidas. Una prueba de evidencia es el estándar HbbTV [4], que proporciona mecanismos para armonizar la entrega y consumo interactivo de contenidos *broadcast* y *broadband* a través de TV conectadas y dispositivos secundarios con conectividad IP.

Sin embargo, la distribución y consumo de contenidos multimedia a través de tecnologías híbridas todavía se enfrenta a numerosos retos, como: la configuración y adaptación de dispositivos a esta reciente tecnología, cuya última especificación [4] aún no ha sido comercializada; la especificación e implementación de soluciones para la señalización y descubrimiento de los servicios multimedia híbridos disponibles y las aplicaciones relacionadas; el diseño e implementación de mecanismos de descubrimiento e interacción entre los dispositivos involucrados en la sesión multimedia; el diseño e implementación de soluciones de sincronización multimedia adaptativa y precisa para alinear en el tiempo el consumo de los contenidos híbridos, tanto en escenarios mono- como multi-dispositivo, etc.

Como principal contribución de este trabajo, se presenta una plataforma compatible con el estándar HbbTV 2.0.1 que incluye soluciones eficientes para los retos anteriormente identificados. Además, la plataforma se ha evaluado, tanto objetiva como subjetivamente, para el caso de uso de TV multi-vista y multi-dispositivo que, tal y como se argumenta en las próximas secciones, es un caso de uso bastante relevante y con potencial comercial.

El artículo se estructura de la siguiente manera: en la sección 2, se revisan las soluciones de sincronización híbrida e IDES (*Inter Device Media Synchronization*) más relevantes. En la sección 3, se presenta la plataforma desarrollada, describiendo sus componentes y funcionalidades. En la sección 4, se presentan algunos resultados de las evaluaciones objetivas y subjetivas realizadas. Finalmente, en la sección 5, se presentan las conclusiones y algunas ideas para trabajo futuro.

II. Estado del arte

A lo largo de los años, se han propuesto diferentes soluciones para la sincronización de contenidos multimedia, utilizando distintas tecnologías de transmisión, en diferentes entornos de red y aplicaciones. Asimismo, numerosos estudios han analizado los avances con respecto a la sincronización multimedia. Una clasificación de soluciones de *sincronización inter-media* e *Inter-Destination Media Synchronization* (IDMS) puede encontrarse en [7], mientras que en [8] se proporciona un repaso histórico sobre la sincronización multimedia, teniendo en cuenta los avances tecnológicos, modelos teóricos y estudios sobre la percepción humana. El estudio en [9] proporciona una revisión de los estándares más recientes para la *sincronización inter-media, híbrida e IDMS*. Además, en [10] se analiza cómo las referencias de reloj y las marcas de tiempo se utilizan en diferentes estándares MPEG y DVB.

Esta sección se centra principalmente en la revisión de los trabajos más relevantes en el contexto de la sincronización híbrida. Además, también se describen brevemente las pruebas de concepto más relevantes implementadas para IDES. Las soluciones de sincronización basadas en técnicas propietarias (como *watermarking* o *fingerprinting*) no se consideran, debido a sus múltiples inconvenientes (baja precisión, sobrecarga, sensibilidad al ruido, escalabilidad baja...), tal y como se describe en [9], además de no ser apropiadas en el contexto de este trabajo.

En [11] se propone un algoritmo para sincronización inter-media entre flujos multimedia generados por el mismo proveedor de contenidos. Dicho algoritmo se basa en controlar la reproducción del audio y del vídeo teniendo en cuenta el valor de los campos PCR (*Program Clock Reference*) y PTS (*Presentation Timestamp*) asociados, incluidos en los flujos MPEG2-TS. El uso de los campos PCR y PTS para sincronizar diferentes flujos también se analiza en [12]. Sin embargo, en este caso los flujos son generados por distintas fuentes que comparten una fuente de reloj común (ej., utilizando *Network Time Protocol*, NTP)

para insertar valores PTS en los flujos MPEG2-TS de manera sincronizada.

En [13] también se aborda la transmisión híbrida *broadcast*/*broadband* de contenidos audiovisuales utilizando MPEG2-TS. En dicho trabajo, se identifica el hecho de que los codificadores *broadcast* y *broadband* no compartan sus relojes como una barrera para conseguir la sincronización híbrida. Como respuesta, se propone unificar la transmisión *broadcast* y *broadband* a través de la utilización del protocolo IP en ambas e insertando marcas de tiempo comunes.

En [14], el concepto de contenidos híbridos se analiza desde un punto de vista diferente: se investiga cómo segmentar un único flujo y programar la transmisión de los distintos segmentos a través de todas las redes disponibles, con tal de garantizar una reproducción apropiada en la parte del consumidor, con el mínimo número de interrupciones.

En [15] también se investiga sobre el consumo sincronizado de contenidos híbridos. En particular, dicho trabajo se centra en la sincronización de un flujo de contenido audiovisual con formato MPEG2-TS transmitido a través de un canal IPTV, empleando los protocolos RTP/RTCP, y un flujo de audio con formato MP3, transmitido vía *Internet Radio*, empleando también los protocolos RTP/RTCP. El objetivo de la sincronización se consigue asegurándose que todos los proveedores de contenido involucrados insertan marcas de tiempo provenientes de una fuente de reloj común, mediante el uso de NTP, y, posteriormente, ejecutando dos procesos principales: sincronización inicial y sincronización continua.

En [16] se propone una solución diferente para la sincronización híbrida. Se basa en el uso de relojes globales, no requiere la existencia de un canal de retorno y no requiere comunicación entre las redes y/o tecnologías utilizadas. Se investigan dos escenarios. Por un lado, la sincronización entre un flujo de audio transmitido vía *broadcast* FM y un flujo audiovisual MPEG2-TS transmitido vía *broadband* se consigue mediante el uso de un reloj UTC (*Universal Time Clock*) común e insertando marcas de tiempo de dicho reloj en las estructuras RDS (*Radio Data System*) del flujo FM y en la tabla *Time and Date Table* (TDT) del flujo MPEG2-TS. Por otro lado, también se consigue la sincronización entre un MPEG2-TS transmitido vía *broadcast* y otro vía *broadband*. En dicho caso, se logra sincronizar ambos flujos mediante la inserción de marcas de tiempo en las tablas TDT de cada uno de ellos.

En [18] se muestra que se puede conseguir sincronización a nivel de trama entre contenidos híbridos en un único dispositivo, empleando para ello flujos MPEG2-TS locales y un flujo externo MPEG-DASH. Para ello, utilizan la solución basada en PCR/PTS en combinación con una solución propuesta por la ETSI [19] para proporcionar líneas de tiempo absolutas en los flujos MPEG2-TS. Esta última solución permite insertar líneas de tiempo extrínsecas y absolutas dentro de un flujo MPEG2-TS, consiguiendo de esta forma un reloj común para diferentes flujos. Así, se superan los problemas que surgen cuando únicamente se confía en la solución basada en PCR/PTS (como en [11] y [12]). En dicho trabajo, se implementa un banco de pruebas utilizando el *framework* multimedia GStreamer [20] (también el empleado en la plataforma propuesta).

El trabajo en [21] es una evolución del trabajo en [16], y presenta una solución para la transmisión de contenidos híbridos (DVB y DASH), basada en la solución propuesta por la ETSI [19]. En dicho trabajo se extiende la plataforma GPAC [22] (también utilizada en la plataforma presentada) para: un mejor soporte de DASH; poder identificar la localización del contenido *broadband* (DASH) adicional; así como incluir una descripción del contenido que está siendo transmitido. Asimismo, el trabajo en [5], enumera y describe casos de uso relevantes relacionados con la transmisión de contenidos híbridos y presenta un banco de pruebas, también basado en GPAC, para implementar uno de ellos. En dicho trabajo se utiliza un mecanismo más reciente para conseguir sincronización híbrida, bajo el nombre de TEMI (*Timing External Media Information*) [17], propuesto por MPEG y DVB como una mejora (*amendment*) a la especificación ISO/IEC 13818-1. Este mecanismo cuenta con las siguientes características: 1) inserción de una línea temporal extrínseca, absoluta y estable en el flujo MPEG2-TS; 2) inserción de la localización (vía URL) del contenido *broadband* relacionado en el flujo MPEG2-TS; 3) notificación del momento de disponibilidad del contenido *broadband*. Los dos primeros tipos de metadatos se conocen comúnmente como *TEMI timeline* y *location descriptor*, respectivamente. En particular, en [5] se añade soporte para la inserción y extracción de marcas TEMI en el multiplexor y demultiplexor MPEG2-TS, así como para la codificación *High Efficiency Video Coding* (HEVC) y su extensión escalable multi-nivel o multi-capa (*Layered*-HEVC). Mediante este esquema de codificación y transmisión híbrida, la capa base HEVC se envía vía *broadcast* (DVB) y las capas de mejora vía *broadband* (DASH) siendo recibidas bajo demanda por los consumidores.

Nuestro trabajo también se basa en el uso de TEMI para conseguir sincronización híbrida, por las ventajas que ofrece sobre el mecanismo propuesto por la ETSI [19], tal y como se describe en [9], y por ser el mecanismo adoptado por el estándar HbbTV.

Todos los trabajos mencionados hasta este punto han abordado la sincronización híbrida en un único dispositivo, pero la sincronización IDES también se está convirtiendo en un tema de especial relevancia, debido a la masiva proliferación y uso de dispositivos secundarios, tales como *smartphones* y *tablets*. En este contexto, el trabajo en [23] implementa el mecanismo propuesto por la ETSI basado en líneas temporales absolutas [19] como una solución para proporcionar sincronización híbrida en escenarios TV multi-pantalla, utilizando dispositivos secundarios. Las pruebas de

evaluación muestran que la precisión obtenida en la sincronización entre un flujo de audio *multicast* y un flujo de vídeo DVB-T es suficiente para conseguir *lip-sync* (precisión del orden de 80 ms). En [24], se combinan un conjunto de componentes tecnológicos (mecanismo definido por la ETSI [19], protocolo IDES, mecanismos de descubrimiento y solución para *tiled-streaming*) con el objetivo de proporcionar experiencias multi-pantalla inmersivas. En concreto, se desarrollan aplicaciones HbbTV (previas a la especificación de la versión 2.0) para sincronizar contenidos DVB-T en un terminal híbrido con contenidos DASH (*tiled-streaming*) en dispositivos secundarios. Sin embargo, el contenido *broadcast* se "simula" en dicho trabajo con contenido local, sin realizar transmisión y recepción DVB-T.

III. Plataforma compatible con HbbTV 2.0

La plataforma extremo-a-extremo para la distribución y consumo sincronizado de contenidos multimedia híbridos desarrollada en este trabajo se divide en dos partes: la de proveedor(es) de contenidos y la del usuario/consumidor final. La Fig. 1 proporciona una visión general de la plataforma.

La plataforma es válida para entornos mono- y multi-dispositivo, en los que existe un dispositivo principal, en adelante MS (*Main Screen*) y uno (o varios) dispositivos secundarios, en adelante CS (*Companion Screen*), integrados en el mismo dispositivo que el MS o bien en dispositivos secundarios (ver Fig. 1). El MS consume contenidos MPEG2-TS recibidos vía *broadcast* (DVB), sobre un evento específico (p.ej., un concierto), en una TV principal conectada (terminal híbrido). Este flujo *broadcast*, además, contiene información insertada (metadatos) que permite que los CS activos puedan consumir contenidos multimedia adicionales relacionados (p.ej., vídeos relacionados con el contenido *broadcast*) distribuidos vía *broadband* (DASH).

La plataforma se ha desarrollado empleando, en su mayor parte, el *framework* GStreamer [20], tanto en dispositivos basados en Linux como en Android.

A. Parte de Proveedor(es) de Contenidos

En la parte del proveedor de contenidos (zona izquierda de la Fig. 1) la plataforma incluye las funcionalidades necesarias para la codificación, preparación, segmentación, almacenamiento, modulación y transmisión de los contenidos multimedia. Asimismo, esta parte también incluye funcionalidades para la generación, inserción y almacenamiento de información o metadatos (ej. líneas de tiempo absolutas, descripción y localización de los contenidos multimedia disponibles, etc.) necesarios para conseguir sincronización híbrida.

La plataforma permite la entrega de contenidos multimedia vía *broadcast*, empleando DVB-T, y vía *broadband*, empleando DASH, que es la tecnología adoptada por HbbTV para la distribución de contenidos *broadband*. Además, también soporta otras tecnologías como HLS o *Real Time Streaming Protocol* (RTSP) + RTP/RTCP.

Con respecto a *broadcast*, la plataforma permite la la configuración de la codificación de los contenidos, la inserción de los descriptores TEMI (empleando herramientas de GPAC [22]) y la modulación y transmisión de los contenidos generados, utilizando una tarjeta PCI Dektec DTA-2111 y su software asociado *StreamXpress*. La tasa de inserción de los descriptores TEMI es configurable (aunque típicamente se añade el *TEMI timeline descriptor* una vez por cada trama).

Con respecto a *broadband*, la plataforma soporta todos los requisitos necesarios para ofrecer contenidos DASH, empleando una herramienta desarrollada por nuestro grupo en un trabajo previo [25]. En concreto dichos requisitos son: 1) codificación en diferentes calidades (*bitrates*, resoluciones, patrones de GoP - *Group of Pictures*-…); 2) segmentación, con un tamaño de segmentos (*chunks*) configurables; 3) generación del *Media Presentation Description* (MPD); y 4) almacenamiento del contenido DASH y del MPD en un servidor web convencional. Además, se han realizado procesos similares para añadir soporte para HLS, mientras que también se ha desarrollado un servidor RTSP + RTP/RTCP, empleando para ello librerías y plugins disponibles en el *framework* GStreamer [20].

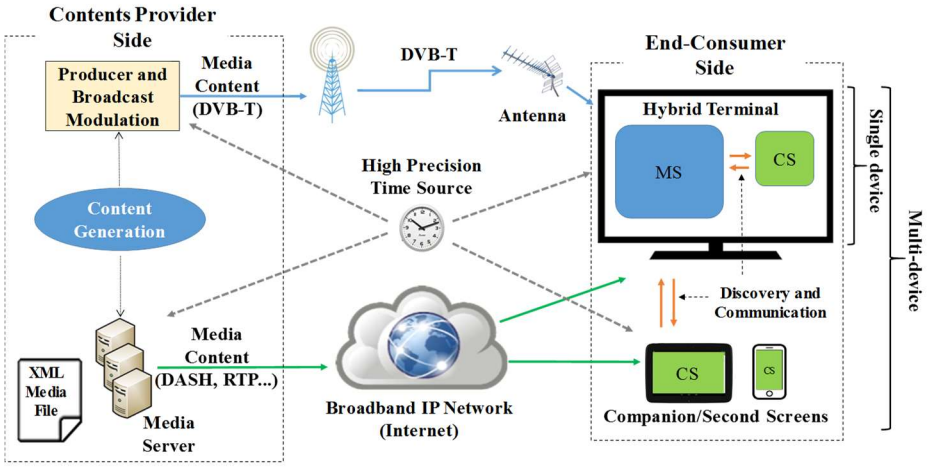


Fig. 1 Visión general de la plataforma desarrollada y del escenario evaluado

Aparte de la generación y preparación de los contenidos multimedia híbridos, así como de la inserción de descriptores TEMI, también es imprescindible la inclusión de mecanismos adicionales para indicar la disponibilidad de dichos contenidos multimedia relacionados y para proporcionar los metadatos relevantes sobre los mismos (ej. su relación con el contenido *broadcast*, la naturaleza del contenido, las URL para acceder a los mismos…). De esta forma los receptores pueden conocer su existencia y disponer de la información necesaria para decidir si quieren o no reproducirlos y, en caso afirmativo, preparar los reproductores adecuados. Para el descubrimiento de los contenidos y la notificación de su existencia, se emplea el *TEMI location descriptor*. Como mecanismo de enlazado y descripción de contenido, se emplea un fichero XML, con los nodos y atributos necesarios. Este fichero XML se almacena en un servidor web (ej. el mismo servidor que almacena el contenido *broadband*), al cual se accede vía HTTP utilizando la URL insertada en el *TEMI location descriptor*. Dependiendo de la naturaleza de los contenidos híbridos disponibles, se incluyen metadatos relevantes específicos, como por ejemplo la tecnología de distribución (DASH, HLS, RTP…), los tipos de contenido (audio, video, web), su formato de encapsulación (MP4, MPEG2-TS…), su codificación (H264, MP3…) y otros metadatos relevantes (ej. una breve descripción del contenido, idioma del audio, etc.). Además, el fichero XML también incluye otro tipo de información relevante, como la fecha de su última modificación, así como la tecnología utilizada para obtener una fuente de reloj global (ej. *Precision Time Protocol* -PTP-, NTP…), junto con la URL del servidor utilizado. Esta fuente de reloj global se utilizará para proporcionar una noción coherente del tiempo en la sesión multimedia, ya sea en la inserción de marcas temporales TEMI o en su interpretación/intercambio en la parte del consumidor final, posibilitando así el objetivo de la sincronización híbrida. Este fichero XML es descargado y analizado por el terminal híbrido en la parte del usuario final.

B. Parte del Usuario/Consumidor Final

En la parte del usuario/consumidor final (zona derecha de la Fig. 1), la plataforma incluye los módulos necesarios para implementar las siguientes funcionalidades:

- Sintonización, recepción y reproducción de canales DVB-T. Estas funcionalidades están integradas en el módulo MS.
- Descubrimiento de contenidos adicionales relacionados, accesibles vía *broadband*, accediendo e interpretando la información contenida el fichero XML mencionado con anterioridad. Esta funcionalidad está integrada en el MS.
- Descubrimiento de un MS activo por parte de un CS y la asociación entre ellos. Por un lado, se ha adoptado una implementación *open-source* del protocolo DIAL (*Discovery And Launch Protocol*) [26], adoptado por HbbTV para descubrimiento automático de los dispositivos MS y CS involucrados. En particular, dicha implementación se ha integrado en el terminal híbrido, siendo activada al lanzar el MS y quedando a la espera de nuevas conexiones de CS. Por otro lado, las funcionalidades que ofrece el protocolo *Simple Service Discovery Protocol* (SSDP) [27] se han desarrollado desde cero. El proceso que corresponde al uso de SSDP es invocado cada vez que un CS quiere descubrir un dispositivo con un servidor DIAL activo. Tras el descubrimiento, ambos módulos/dispositivos (MS y CS) se asocian y se comunican a través de un canal bidireccional (vía *websockets*). Este canal de comunicación permite el intercambio de información relevante para conseguir el objetivo de la sincronización híbrida.
- Identificación, selección, recepción, procesado y reproducción (adaptativa) de los contenidos *broadband* relacionados. Estas funcionalidades están integradas en el módulo CS, que puede ser ejecutado en el terminal híbrido (mismo dispositivo que el MS) o en uno o varios dispositivos independientes, pero conectados a la misma red local (ver Fig. 1). La identificación y selección del contenido adicional *broadband* se logra gracias a la información disponible en el fichero XML, cuya dirección está insertada en el *TEMI location descriptor*, a su vez insertado en el flujo MPEG2-TS del contenido *broadcast*. El fichero XML es analizado por el MS, que extrae la información de interés necesaria para los CS y la envía a través del canal de comunicación establecido entre ambos. La recepción, procesado y reproducción de los contenidos adicionales *broadband* se consigue generando y enlazando los módulos de GStreamer necesarios para implementar un reproductor multimedia que soporte la naturaleza del contenido.
- Sincronización multimedia. La *sincronización intra-media* e *inter-media* para/entre los elementos multimedia (audio, video…) de cada flujo multiplexado recibido está soportada de manera nativa en GStreamer. La sincronización híbrida entre diferentes flujos en un mismo dispositivo (es decir, entre los contenidos reproducidos por el MS y el CS, ambos en el terminal híbrido), se consigue extrayendo las marcas temporales TEMI y comparando el estado de los procesos de reproducción de los flujos involucrados, utilizando para ello la información de las líneas de tiempo relativas (es decir, los valores de los campos PCR/PTS) y absolutas TEMI, y ajustando el estado de los procesos de reproducción de los flujos requeridos si la asincronía entre ellos excede un determinado umbral (configurable). Las líneas de tiempo se intercambian entre MS y CS a través del

canal bidireccional establecido previamente. Respecto a las técnicas de ajuste del proceso de reproducción, se pueden adoptar dos alternativas. Por una parte, se puede realizar técnicas de ajuste agresivas, como son saltos y pausas. Esta alternativa es apropiada cuando la asincronía supera un cierto umbral superior (U1) o cada vez que se active la reproducción de un nuevo contenido (o flujo) *broadband*. Por otro lado, la adopción de técnicas de reproducción multimedia adaptativa (*Adaptive Media Playout*, AMP) es mucho más conveniente cuando el nivel de asincronía que debe corregirse supera un umbral U2, pero no supera el umbral U1 (U1>U2). AMP consiste en ajustar de manera suavizada (acelerando o reduciendo) la tasa de reproducción para corregir situaciones de asincronía. Permite alcanzar una mayor precisión en la sincronización y evitar discontinuidades en los procesos de reproducción, que pueden resultar molestas para la percepción humana (mala calidad de experiencia, QoE) [28]. Por su parte, IDES entre flujos reproducidos por el MS y por otros CS en diferentes dispositivos independientes se puede alcanzar de manera similar, aunque, en este caso, las funcionalidades de MS y CS se integran en diferentes dispositivos.

IV. Evaluación

El funcionamiento de la plataforma, para cada una de sus funcionalidades, ha sido evaluado, tanto objetiva como subjetivamente, obteniendo resultados muy satisfactorios. El escenario de evaluación, la metodología seguida y algunos de los resultados obtenidos se presentan a continuación.

A. Escenario evaluado

La plataforma se ha evaluado para el caso de uso de TV multi-vista y multi-dispositivo. Este caso de uso permite al proveedor de contenidos, para un determinado evento (ej. un evento deportivo, un concierto musical…), ofrecer un flujo *broadcast* (DVB) con las escenas seleccionadas por el realizador, complementadas con contenidos adicionales capturados por otras cámaras, ofreciendo un punto de vista diferente e información adicional del evento, que son distribuidos vía *broadband* (ej. DASH, HLS, RTP…). Los contenidos provenientes de estas cámaras adicionales pueden ser reproducidos por el CS, bien en el mismo terminal híbrido (ej. en modo PiP o mosaico, junto al contenido principal) o bien en dispositivos secundarios, de manera sincronizada. De acuerdo con un estudio realizado recientemente por nuestro grupo [29], se trata de un caso de uso que despierta un gran interés en los consumidores y que tiene mucho potencial comercial.

Concretamente, el escenario evaluado en este trabajo consiste en cuatro cámaras con distintas vistas de un concierto[1]. Las escenas seleccionadas por el realizador son distribuidas vía DVB-T y son reproducidas por el MS en el terminal híbrido. Por otro lado, los contenidos capturados por la misma cámara junto con los capturados por tres cámaras adicionales, se preparan y distribuyen vía DASH, pudiendo ser reproducidos por el CS en el mismo terminal híbrido o en dispositivos secundarios. La aplicación HbbTV desarrollada para el CS incluye un menú que permite la selección y/o cambio dinámico de la cámara a reproducir. Además, se ha configurado un umbral de ±80ms, de manera que no se realizan ajustes de reproducción si la asincronía calculada no lo supera.

La Fig. 2 muestra el aspecto del escenario multi-vista y multi-pantalla que se ha implementado y evaluado, el cual permite una mayor inmersión y personalización en las experiencias de consumo de contenidos de TV.

La Fig. 3, muestra un ejemplo del fichero XML utilizado para notificar la existencia de los contenidos híbridos relacionados en este escenario. Aunque en el escenario evaluado todos los contenidos *broadband* se distribuyen vía DASH, la figura muestra cómo se señalizarían en el caso en el que se distribuyeran utilizando otras tecnologías *broadband*, como HLS y RTP.

Fig. 2 Escenario multi-vista y multi-dispositivo evaluado

```
<Media Content Description File>
    <CAM id=1 protocol="dash" media_type="AV" media_format=
    "h264/aac" metadata="TV Camera/english" uri=
    "http://192.16.0.10/dash/cam1/cam1.mpd" />
    <CAM id=2 protocol="hls" media_type="AV" media_format=
    "h264/aac" metadata="Close-Up Camera/english" uri=
    "http://192.16.0.10/hls/cam2/cam2.m3u8" />
    <CAM id=3 protocol="rtp" media_type="AV" media_format=
    "h264/aac" metadata="First Row Camera/english" uri=
    "rtsp://224.0.0.10:5001/show" />
    <URL id=1 protocol="http" media_type="website"
    media_format="html" metadata= "url_event/english" >
    http://iim.webs.upv.es />
    <LASTUPDATE format="dd/mm/yyy-hh:mm:ss">10/11/2017-
15:45:30 />
    <CLOCK protocol= "ntp" media_type="time"
    format="64_bit_ntp_time" url= "ntp.upv.es" value=
    390909493091340" />
</ Media Content Description File >
```

Fig. 3: Ejemplo del fichero XML para el escenario evaluado

[1] Youtube es el propietario de los contenidos multimedia utilizados en el escenario de evaluación de este trabajo. Se trata de un concierto, grabado con cuatro cámaras (y con mismo audio), de la cantante Madilyn Bailey en Los Ángeles, 2015.

Vídeos demostrativos mostrando las capacidades y funcionalidades de la plataforma están disponibles en http://iim.webs.upv.es/prototypes.html

B. *Evaluación objetiva*

La plataforma permite la reproducción de contenidos *broadband* relacionados a través de los CS, tanto en el terminal híbrido como en dispositivos secundarios. Cuando se reproducen en el mismo terminal híbrido, los contenidos pueden presentarse en modo mosaico o PiP. Aunque los ecos del audio y la sincronización entre imagen y audio (*lip-sync*) puede percibirse, resulta difícil discernir visualmente el nivel de asincronía existente mediante estas configuraciones. Una manera sencilla de comprobar el nivel de sincronización consiste en reproducir el mismo contenido tanto en el MS como en el CS, pero recortando las imágenes recibidas por cada uno en su mitad derecha e izquierda y colocarlas una junto a la otra. Si la reproducción de ambos procesos se percibe como un único vídeo (ver Fig. 4 y vídeos disponibles en el enlace proporcionado) significa que la precisión conseguida es muy satisfactoria.

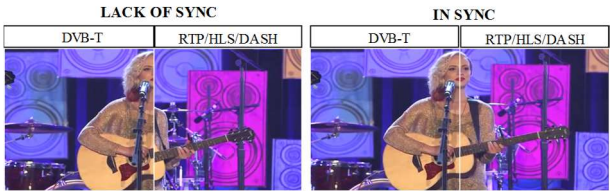


Fig. 4: Comprobación visual de la sincronización

Otro método utilizado para comprobar visualmente la asincronía entre los flujos multimedia consumidos consiste en superponer un texto indicando el número de trama del vídeo, insertado durante su proceso de codificación. Mediante este método, se puede tomar fotos (ver Fig. 2) o grabar vídeos del escenario evaluado (ver enlace proporcionado) e ir pausando el vídeo en instantes específicos para comprobar el desfase entre tramas existente en cada momento.

Además, se ha utilizado otro método más sistemático y transparente para los usuarios para medir las asincronías. Consiste en comparar (a nivel de código) las líneas temporales e instantes de reproducción de cada flujo e ir registrando las asincronías calculadas. Como ejemplo, la Fig. 5 muestra la evolución de la asincronía entre el MS y el CS (en dispositivos diferentes) cuando reproducen un flujo DVB-T y uno DASH, respectivamente. Puede observarse que, tras un proceso inicial de sincronización cuando se lanza el CS, la asincronía se mantiene dentro de límites aceptables (inferiores a ±80ms). Cuando se cambia a una cámara/vista diferente en el CS, se recupera la sincronización tras un periodo de ajuste inicial. La Fig. 6 confirma que el nivel de asincronía se mantiene dentro de los límites permitidos, exceptuando situaciones esporádicas (ej. al lanzar el CS, al cambiar de cámara, en situaciones de congestión de red o del dispositivo…). Por tanto, el rendimiento de la plataforma en cuanto a precisión de sincronización es satisfactorio.

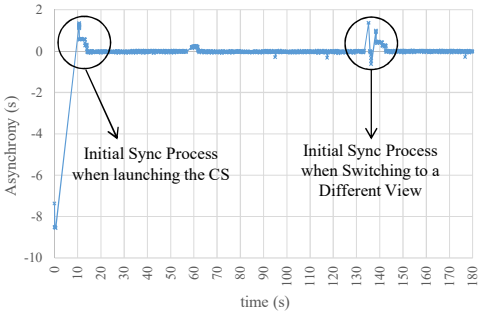


Fig. 5: Evolución de la asincronía entre el MS y el CS

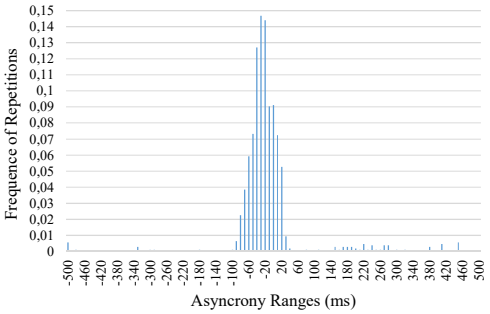


Fig. 6: Distribución de los valores de asincronía

C. *Evaluación subjetiva*

El rendimiento de la plataforma (en cuanto a sincronización, estabilidad y fluidez), para el caso de uso presentado, junto con la usabilidad de la misma, la percepción de los niveles de retardos y sincronización y el interés por parte de los consumidores en este tipo de plataformas y servicios, también han sido evaluados subjetivamente. Concretamente, 24 usuarios participaron en el estudio, los cuales, tras una breve introducción, probaron la plataforma utilizando dos tipos de dispositivos (*tablets* y *smartphones*).

En primer lugar, se presentó a los usuarios la visualización del mismo contenido tanto en el MS como en el CS, con el objetivo de determinar la precisión en la sincronización alcanzada. Para ello se mostraron 5 casos presentados de forma aleatoria, donde se forzaron asincronías fijas en el CS de 0, ±1 y ±3 segundos. Tras la visualización de cada caso, los usuarios utilizaron la métrica MOS (*Mean Opinion Score*) para evaluar el nivel de sincronización percibido, en la que un valor de 1 se corresponde con un nivel de sincronización muy malo y un 5 con un nivel excelente. Los resultados que se muestran en la Tabla 1 confirman que la QoE fue muy buena cuando no se forzó asincronía, aunque empeoró significativamente al aumentarla. Se presentan los valores obtenidos junto con el intervalo de confianza (IC) del 95% para cada caso visualizado.

Tabla 1: MOS ± IC del 95% para cada caso

Caso	***-3"***	***-1"***	***0"***	***+1"***	***+3"***
MOS ±IC 95%	*1.75* ±0.34	*2.17* ±0.43	*4.63* ±0.21	*2.46* ±0.45	*1.75* ±0.34

En segundo lugar, se permitió a los participantes utilizar libremente la plataforma, tras lo cual rellenaron un cuestionario. Los resultados obtenidos demuestran que, en general, los usuarios están muy interesados en este tipo de servicios multimedia híbridos (94%). Se mostraron muy satisfechos con la usabilidad de la plataforma (100%), con la utilidad de la misma (83%) y con su fluidez, estabilidad y nivel de sincronización conseguido (96%, aproximadamente, en todas ellas).

V. Conclusiones y trabajo futuro

Este artículo ha presentado una plataforma extremo-a-extremo compatible con la versión más reciente del estándar HbbTV (2.0.1) para la distribución y reproducción sincronizada de contenidos híbridos (*broadcast* y *broadband*) relacionados, tanto en un único dispositivo (ej. TV conectadas) como en múltiples dispositivos (ej. escenarios multi-pantalla). La plataforma ha sido evaluada para el caso de uso de TV multi-vista y multi-dispositivo, obteniendo resultados prometedores en cuanto a su rendimiento, usabilidad e interés despertados, reflejando así su potencial en el actual paradigma de consumo de contenidos multimedia.

Como trabajo futuro, se pretende minimizar los retardos, mejorar la precisión de sincronización y optimizar el rendimiento para mejorar la QoE. Además, con el fin de tener una plataforma basada completamente en GStreamer, se desarrollará un módulo insertor de descriptores TEMI con dicho *framework*. Asimismo, se extenderá la plataforma con el fin de incluir soporte para diferentes sistemas como Windows o iOS; además de para otros tipos de formatos y contenidos multimedia (ej. UHD, omnidireccionales, etc.), así como contenido en vivo. Finalmente, la plataforma se implementará y probará en sistemas reales de TV.

Agradecimientos

El trabajo ha sido financiando por el MINECO y FEDER con ref. TEC2013-45492-R. Agradecer a Samsung Electronics Iberia S.A. la prestación de los dispositivos para el desarrollo y evaluación de la plataforma.

Referencias

[1] A. Begen, et al, "Watching Video over the Web: Part 1: Streaming Protocols", IEEE Internet Computing, 15(2), pp. 54-63, April 2011.

[2] H. Schulzrinne, et al, "RTP: A Transport Protocol for Real-Time Applications", IETF Standard, RFC 3550, July 2003.

[3] Information Technology – Dynamic Adaptive Streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats, ISO/IEC 23009-1, 2014. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=65274

[4] HbbTV 2.0.1 Specification, HbbTV Association Resource Library, https://www.hbbtv.org/resource-library, July 2016.

[5] J. Le Feuvre, et al, "A Test Bed for Hybrid Broadcast Broadband Services". MediaSync Workshop 2015, Brussels (Belgium), June 2015.

[6] R. van Brandenburg, et al, "Immersive second-screen experiences using hybrid media synchronization", MediaSync Workshop 2013, Nantes (France), October 2013.

[7] F. Boronat et al, "Multimedia group and inter-stream synchronization techniques: A comparative study", Information Systems, 34(1), pp. 108–131, March 2009.

[8] Z. Huang, et al, "Evolution of Temporal Multimedia Synchronization Principles: A Historical Viewpoint", ACM TOMCAPP, Vol. 9, No. 1s, Article 34, October 2013.

[9] M. Oskar van Deventer, et al, "Standards for Multi-stream and Multi-device Media Synchronization", IEEE Communications Magazine, 54(3), pp. 16-21, March 2016.

[10] L. Beloqui, et al, "Understanding Timelines Within MPEG Standards", IEEE Communications Surveys & Tutorials, 18(1), pp. 368-400, Firstquarter 2016.

[11] L. Ehley, B. Furht, M. Ilyas, "Evaluation of multimedia synchronization techniques", International Conference on Multimedia Computing and Systems, pp. 514-519, May 1994.

[12] M. Armstrong, et al, "Enabling and Enriching Broadcast Services by Combining IP and Broadcast Delivery", BBC Research White Paper WHP 185, September 2010.

[13] S. Aoki, et al, "A New Transport Scheme for Hybrid Delivery of Content over Broadcast and Broadband". IEEE International Symposioum on Broadband Multimedia Systems and Broadcasting (BMSB) 2011, Fraunhofer (Germany), June 2011.

[14] K. Evensen, et al, "Improving the Performance of Quality-Adaptive Video Streaming over Multiple Heterogeneous Access Networks", ACM MMSYS 2011, pp. 57-68, San Jose, California (USA), February 2011.

[15] L. Beloqui, et al, "Interactive Multi-source Media Synchronisation for HbbTV". MediaSync Workshop 2012, Berlin (Germany), October 2012.

[16] C. Concolato, et al, "Synchronized Delivery of Multimedia Content over Uncoordinated Broadcast Broadband Networks". ACM MMSYS 2012, Chapel Hill, North Carolina (USA), February 2012.

[17] ISO/IEC 13818-1:2013/PDAM 6 Delivery of Timeline for External Data, 2013, http://mpeg.chiariglione.org/standards/mpeg-2/systems/textisoiec-13818-12013pdam-6-delivery-timeline-external-data

[18] A. Veenhuizen, et al, "Frame accurate media synchronization of heterogeneous media sources in an HBB context", MediaSync Workshop 2012, Berlin (Germany), October 2012.

[19] ETSI TS 102 823 v1.1.1 Digital Video Broadcasting (DVB); Specification for the carriage of synchronized auxiliary data in DVB transport streams, 2005. http://www.etsi.org/deliver/etsi_ts/102800_102899/102823/01.01.01_60/ts_102823v010101p.pdf

[20] GStreamer: open source multimedia framework. https://GStreamer.freedesktop.org/

[21] J. Le Feuvre, C. Concolato, "Hybrid Broadcast Services using MPEG DASH", MediaSync Workshop 2013, Nantes (France), October 2013.

[22] GPAC https://gpac.wp.mines-telecom.fr/

[23] C. Howson et al, "Second Screen TV Synchronization", IEEE International Conference on Consumer Electronics, Berlin (Germany), January 2011.

[24] R. van Brandenburg, et al, "Immersive second-screen experiences using hybrid media synchronization", MediaSync Workshop 2013, Nantes (France), October 2013.

[25] D. Gómez, et al, "End-to-End DASH Platform including a Network-based and Client-based Adaptive Quality Switching Module", ACM MMSYS 2016, Klagenfurt (Austria), May 2016.

[26] DIAL (DIscovery And Launch) protocol specification, Version 1.7.2, 2015. http://www.dial-multiscreen.org/dial-protocol-specification/DIAL-2ndScreenProtocol-1.7.2.pdf

[27] Simple Service Discovery Protocol SSDP, Internet Draft http://quimby.gnus.org/internet-drafts/draft-cai-ssdp-v1-03.txt

[28] M. Montagud et al, "How to perform AMP? Cubic adjustments for improving the QoE", Computer Communications, Volume 103, pp 61-63, ISSN 0140-3664, May 2017.

[29] F. Boronat, et al, "Preferencias, necesidades y expectativas de los usuarios españoles en escenarios multimedia híbridos broadcast/broadband", VI Congreso de TV Digital Interactiva (TVDi), Mallorca (Spain), October 2015.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6449

PRoFIT: modelo forense-IoT con integración de requisitos de privacidad

Ana Nieto, Ruben Rios, Javier Lopez
Network, Information and Computer Security (NICS) Lab,
Lenguajes y Ciencias de la Computación,
Universidad de Málaga
Campus Teatinos s/n, 29071 Málaga (España)
E-mail:{nieto,ruben,jlm}@lcc.uma.es

***Resumen*—La *Internet de las cosas* (IoT) complica sobremanera la extracción de evidencias electrónicas que pueden servir de base para una investigación forense. En entornos altamente cambiantes y con una densidad de dispositivos tan elevada, es muy difícil entender completamente el contexto de la ofensa. Es por ello que la cooperación de los individuos, aún no estando directamente implicados en la ofensa, puede ser muy relevante para el analista forense. En este artículo se propone un nuevo modelo para la IoT-Forensics, que pretende sentar las bases para la cooperación voluntaria de los individuos en las investigaciones de delitos telemáticos. Para ello, el modelo integra requisitos de privacidad de la norma ISO/IEC 29100:2011 durante todo el ciclo de vida de la investigación.**

***Palabras Clave*—forense, IoT, privacidad, seguridad, testigos digitales.**

I. INTRODUCCIÓN

La informática forense tradicional se basa en una serie de procesos bien establecidos que tienen como objetivo preservar la evidencia electrónica. Existen varios modelos muy similares entre sí, que describen conductas a seguir precisas y bien definidas, pero son exageradamente estáticos [1]. Las evidencias - discos o dispositivos de los que se extraerán evidencias electrónicas - se recaban por medio de una orden judicial al principio del proceso. Las evidencias electrónicas se obtienen conforme a su volatilidad, siendo posible que algunas de éstas se obtengan en la propia escena (p.ej. volcados de memoria). Se implementan cadenas de custodia por medio de documentos que firman los responsables autorizados a gestionar las evidencias. Esto es así con el fin de preservar la integridad de la prueba, pero no deja de ser un proceso poco flexible, concebido para escenarios muy estáticos.

Sin embargo, estamos viviendo un boom de dispositivos sin parangón. Actualmente los analistas forenses se encuentran con el problema de que hay un sinfín de dispositivos para los que aún no existen herramientas o procesos bien definidos para regular su tratamiento forense [2]. No sólo los dispositivos, sino también las infraestructuras/plataformas intermedias que se usan para la comunicación entre los objetos, plantean grandes desafíos forenses. Precisamente, en 2013 surge el concepto de IoT-Forensics [3] para destacar los problemas que el paradigma IoT introduce en el ámbito forense. Tanto a nivel de adquisición de información como a nivel de cómo se gestiona esa información [4], o la problemática de la avalancha de datos que deben procesarse y correlacionarse.

Pero en IoT-Forensics el problema va más allá; el usuario y sus dispositivos están demasiado conectados como para continuar obviando que el espinoso problema de la privacidad debe ser finalmente abordado. Precisamente, en este artículo definimos el modelo PRoFIT (*Privacy-aware IoT-Forensic*) que integra propiedades de privacidad conforme a la norma ISO/IEC 29100:2011 en las fases de un modelo forense adaptado para la IoT. Este modelo, más dinámico que sus predecesores facilitaría la cooperación voluntaria de los dispositivos del entorno, promoviendo enfoques como por ejemplo la *testificación digital* [5]. En dichos enfoques el dispositivo del usuario forma parte activa en la gestión de evidencias electrónicas, permitiendo adquirir información sobre *ofensas telemáticas* que de otra forma pasarían desapercibidas.

Este artículo se estructura como sigue. La sección II describe la base del modelo forense y los principios de privacidad bajo los que se definen el modelo PRoFIT, así como los trabajos relacionados con este área. La sección III describe el modelo PRoFIT. La sección IV propone un escenario de ejemplo para facilitar la comprensión del modelo, en la que un cliente solicita el inicio de una investigación empleando una herramienta PRoFIT-compliant y hay varios dispositivos que pueden actuar como testigos digitales del hecho. Finalmente se detallan las conclusiones.

II. ANTECEDENTES

Proponer un modelo forense con características de privacidad significa encontrar un balance entre dos posturas tradicionalmente enfrentadas pero que deben vivir en simbiosis dado que el usuario, y sus datos, juegan un papel central en la IoT. El objetivo de esta sección es sentar las bases para la comprensión del modelo propuesto, mostrando por una parte las fases de un modelo forense (Sección A) y, por otra, los requisitos de privacidad (Sección B).

A. Modelos Forenses

Los modelos forenses están destinados a preservar la evidencia electrónica durante todo su ciclo de vida. En esta sección presentamos algunos de estos modelos. En concreto nos centramos en modelos que describen detalladamente las fases a realizar por el analista. A pesar de que el modelo propuesto toma como partida las fases de un modelo tradicional, comentamos también modelos específicos para IoT con el fin de contar con una visión amplia de las soluciones relacionadas con el área de estudio.

1) Tradicionales: En 2001 se propone el modelo DFRW como un esfuerzo conjunto por parte de investigadores, empresas y entidades legales (conocidas como LEAs por sus siglas en inglés, *Legal Enforcement Agency*), que ha sido refinado posteriormente. En [1] se puede consultar una revisión del DFRW y otros modelos posteriores basados en éste (p.ej. ADFM, IDIP, etc.). A su vez, en dicho trabajo, se propone el *Enhanced Systematic Digital Forensic Investigation Model* (ESDFIM), que propone unas fases bastante intuitivas y generales. Estas fases, que detallamos a continuación, son fácilmente adaptables a entornos más restringidos y a la vez multidisciplinares como la IoT.

Durante la fase de (1) *preparación*, se realizan las acciones necesarias previas a la investigación (p.ej. análisis de la legalidad vigente, solicitud de órdenes de registro, preparación de las herramientas para recabar información). La fase de (2) *adquisición y preservación* marca el inicio del ciclo de vida de la evidencia (identificación, obtención de las evidencias volátiles y no volátiles, etiquetado y empaquetado, transporte, etc.). Cabe destacar que el modelo considera evidencias físicas tangibles en este punto. Durante la fase de (3) *examinación y análisis* el examinador forense y los expertos en la materia examinan y analizan *el contenido* de los dispositivos digitales que fueron obtenidos legalmente y convenientemente preservados. En los modelos forenses habitualmente las siguientes fases son para la generación de informes y admisibilidad [1], sin embargo, este modelo sugiere la fase de (4) *compartición de información*, que es la habilidad de intercambiar datos relativos a una investigación entre varios países, organizaciones, personas autorizadas y tecnologías. La siguiente fase es la de (5) *presentación*. Durante dicha fase se procede a presentar a las autoridades competentes los resultados de la investigación. La admisibilidad de la evidencia electrónica depende de cómo ésta se presente durante esta fase. Finalmente la fase de (6) *revisión* compete a la evaluación del proceso de investigación de cara a su mejora. Incluye los procesos para devolver la evidencia una vez procesada (p.ej. un PC) a su dueño.

Por otra parte, la multinorma ISO/IEC 27050:2016 define las fases: identificación, preservación, recolección, procesamiento, revisión, análisis y producción. Sin embargo, escogemos como base para este trabajo el modelo ESDFIM, dado que las fases anteriores pueden mapearse fácilmente a las definidas por la norma y, además, el modelo ESDFIM define la fase *compartición de información*, no contemplada en todos los modelos ni en la norma, que es de especial interés en escenarios de la IoT por permitir la colaboración entre diferentes entidades.

2) Especificos IoT: Los modelos IoT-Forensics que definen fases para la investigación son muy escasos. Como tales, definen fases [6] y [7], siendo este último el más exhaustivo. De hecho, [7] compara su framework con otras soluciones para IoT-Forensics no reflejadas aquí ([8] y [9]), ya que no definen exactamente fases. En concreto, [6] define las fases: planificación (autorización y obtención de órdenes judiciales), IoT, adquisición de datos (identificación de dispositivo, zona, triage, adquisición de datos de las plataformas de acumulación, datos estructurados y no estructurados), cadena de custodia, análisis en laboratorio, resultado, prueba y defensa, consecución y almacenamiento. Dentro del modelo [7], las fases son más generales: proceso proactivo (definición del escenario IoT, identificación de las fuentes de la evidencia, planificación de la detección de incidencias, recolección potencial de evidencias, preservación digital, almacenamiento de evidencias potenciales), IoT-forense (*Cloud forensics*, *network forensics*, *device level forensics*), proceso reactivo (inicialización, adquisitivo, investigativo), proceso concurrente (obtener autorización, documentación, cadena de custodia, investigación física).

Estos modelos aún no consideran la posibilidad de que la adquisición que están planificando afecte a la admisibilidad de la evidencia por no respetar derechos éticos fundamentales y de privacidad. Además, cuentan con el problema añadido de que dependen de órdenes de registro en la primera fase, no considerando entornos IoT con alta densidad de dispositivos que estén dispuestos a cooperar. Por ejemplo, los *testigos digitales* [5] proponen la cooperación entre dispositivos con capacidades de seguridad para desplegar *cadenas de custodia digital* en la IoT. Sin embargo, ningún modelo forense sienta las bases que permita dicha cooperación. A consecuencia de esto, los ataques contra los dispositivos del entorno se seguirán produciendo sin que queden evidencias sobre el suceso y, peor aún, sin que las víctimas lleguen a saberlo.

En resumen, tanto los modelos tradicionales como los modelos propuestos hasta ahora para entornos IoT no están concebidos para considerar la cooperación voluntaria de los actores del entorno, que puedan actuar por ejemplo como *testigos digitales*. Este es un enfoque que aquí pretendemos abordar, y para ello necesitamos precisamente considerar los requisitos de privacidad.

B. Principios de Privacidad

Existen diversas leyes que tienen como objetivo establecer límites a la recolección, procesamiento y difusión de información de carácter personal al que nos vemos sometidos cuando interactuamos con otras entidades o servicios. Estas leyes tiene el objetivo de proteger la privacidad de los usuarios mediante una serie de normas y buenas prácticas.

En 1974 una ley estadounidense establece lo que se conoce como prácticas justas de información (FIPs, *Fair Information Practices*). Esta ley establece una serie de principios que más tarde han sido recogidos y adaptados por diversas guías [10], directivas [11] y estándares, como la ISO/IEC 29100:2011 [12], que considera hasta 11 principios de privacidad, que se detallan a continuación.

Estos principios o prácticas tienen como objetivo devolver al usuario el control sobre sus datos personales. Para ello, el usuario debe dar su consentimiento expreso a la recolección y procesamiento de sus datos personales (P1: *consent and choice*). El sistema que quiere recoger y/o procesar datos del usuario debe informarle del propósito específico para tal recolección y éste debe ser legítimo (P2: *purpose legitimacy and specification*). Además, el sistema debe limitarse a solicitar aquellos datos estrictamente necesarios para el fin especificado (P3: *collection limitation*). Es necesario minimizar la cantidad de datos que son enviados y procesados por el sistema (P4: *data minimization*). El sistema no debe utilizar la información recabada para más fines que los especificados originalmente y no debe almacenarla una vez haya servido su propósito (P5: *use, retention and disclosure limitation*). La información aportada por el data subject debe ser precisa, veraz y actual. Esto es especialmente relevante si la información no procede directamente del usuario en cuestión (P6: *accuracy and quality*). El usuario debe ser consciente en todo momento de las políticas, procedimientos y prácticas de aplicación del sistema (P7: *openess, transparency and notice*). Además, debe tener la posibilidad de acceder a los datos recolectados sobre su persona así como proponer correcciones (P8: *individual participation and access*). Por otra parte, el sistema es responsable de seguir las prácticas o principios de privacidad establecidos y, en caso de no cumplirlos, el data subject puede solicitar compensaciones (P9: *accountability*). Por ello, el sistema debe poner en práctica los mecanismos necesarios para proteger la información personal de los usuarios de accesos no autorizados, pérdidas o manipulaciones (P10: *information security controls*) y medidas que permitan auditar o verificar que se cumple con las medidas de privacidad (P11: *compliance*).

Aunque la puesta en marcha de estas prácticas no garantiza completamente un uso indebido de la información personal de los usuarios, establece una base consistente sobre la que trabajar en pos de preservar la privacidad y ayuda a los usuarios a establecer cierto nivel de confianza con aquellas entidades que solicitan sus datos.

III. EL MODELO PROFIT

El modelo PRoFIT se basa en la cooperación de los dispositivos del entorno. Esto requiere redefinir las fases de los modelos de referencia usados hasta ahora. En particular, a continuación redefinimos las fases 1-3 del modelo ESDFIM (Sección A), definiendo las siguientes fases para PRoFIT (Fig. 1): (1) preparación, (2) recolección basada en el contexto, (3) análisis de datos y correlación, (4) compartición de la información, (5) presentación y (6) revisión.

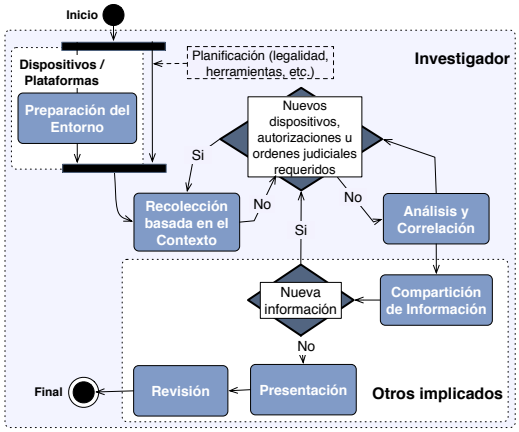


Fig. 1. Fases del modelo PRoFIT

En concreto, la fase 1 es híbrida en nuestro modelo. Esta fase se divide en dos flujos, uno para el investigador y otro para los dispositivos y/o plataformas IoT. Las fases 2-6 forman parte de la investigación forense, siendo el investigador su actor principal, aunque los dispositivos y los usuarios están involucrados para la implementación de los permisos de privacidad. Las fases 4-6 heredan los principios del modelo original (ESDFIM) mencionados en la Sección II.A. Estas fases se modifican para atender a los requisitos de privacidad que deben ser considerados. En concreto, la Tabla I resume los requisitos de privacidad considerados para cada una de las fases del modelo PRoFIT. Cabe destacar que, durante todo el proceso debe garantizarse que se cumple con los principios de privacidad. De ahí que P11 sea un principio trasversal a todas las fases.

Tabla I
REQUISITOS DE PRIVACIDAD EN LAS FASES PROFIT

Fases PRoFIT	**ISO/IEC 29100**					
Preparación del entorno	P1	P2	P4	P7		P11
Recopilación basada en el contexto	P1	P2	P3	P6	P8	
Análisis de datos y correlación	P9	P10				
Compartición de información	P1	P2	P10			
Presentación	P4	P6				
Revisión	P5	P7				

P1. Consentimiento y elección, P2. Legitimidad y especificación del propósito, P3. Recopilación limitada, P4. Minimización de datos, P5. Limitación de uso, retención y divulgación, P6. Calidad y precisión, P7. Transparencia y aviso, P8. Participación y acceso, P9. Responsabilidad, P10. Controles de seguridad de los datos, P11. Conformidad

A. Contexto de Investigación

IoT-Forensics [3] es un nuevo paradigma donde va a ser prácticamente imposible aplicar técnicas forenses eficazmente sin la información adicional de dispositivos que, a priori, tal vez no tengan nada que ver con un caso, pero que estaban en la escena de la ofensa. Para hablar de privacidad en IoT-Forensics necesitamos hacerlo en base al contexto (Fig. 2).

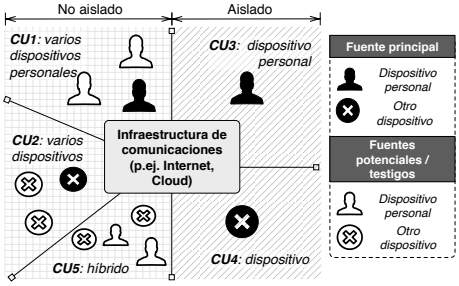


Fig. 2. Casos de uso

Simplificando el problema a un único dispositivo de entrada, que puede ser personal (p.ej. teléfono, implante) o no serlo (p.ej. dispositivos de carácter más general, como un PC del trabajo), la investigación puede consistir en obtener los datos de dicho dispositivo solamente, o requerir información de otros dispositivos - personales o no - que tengan relevancia. Distinguimos tres perfiles de dispositivo (con independencia de que sean personales) para la investigación:

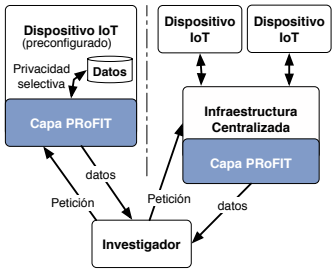
- Ofendido/víctima. Interesado directo en un caso que requiere aplicar mecanismos forenses (p.ej. demandante). Puede solicitar que se inicie una investigación sobre sus datos (c.f. caso de uso en Sección IV)
- Sospechoso. Dispositivo que puede contener evidencias electrónicas inculpatorias o bien exculpatorias y pertenece a un sospechoso.
- Testigo. Dispositivo implicado (directamente o indirectamente) en un caso por poder aportar evidencias electrónicas relevantes para la investigación, pero que a priori no es ni ofendido ni sospechoso.

B. Fase híbrida: preparación

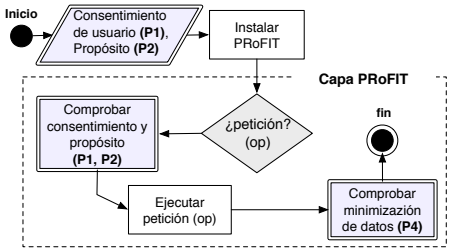
La fase de preparación define dos flujos: uno para el investigador que realiza la planificación tradicional (p.ej. preparación de herramientas forenses y otras operaciones que el investigador considere oportunas antes de la siguiente fase), y otro para la preparación del entorno IoT (dispositivos y/o plataformas).

Nos centramos en la *preparación del entorno IoT*; momento en el que los dispositivos pueden ser preconfigurados atendiendo a criterios de privacidad acordados con el usuario, p.ej. usando el software PRoFIT (Fig. 3(a)). Esto significa que cada dispositivo/plataforma recabará sólo la mínima información necesaria.

Cabe resaltar que este enfoque tiene diferentes interpretaciones y grados; puede ser tan restrictivo o laxo como se desee. Desde configurar un terminal para eliminar



(a) Preconfiguración del entorno IoT



(b) Ciclo de vida del software PRoFIT

Fig. 3. Preparación del entorno

toda la información concerniente a la privacidad de un usuario (p.ej. configurar que se borre el rastro de las aplicaciones de la memoria) - son técnicas anti-forenses, que pueden llevar a que la solución sea totalmente privada pero inservible para un análisis forense - hasta almacenar de toda la información pública a su alcance (p.ej. el usuario mantiene datos compartidos con otros usuarios que han pre-definido una relación abierta y consienten en ese nivel de privacidad abierta). Este último caso permite recabar toda la información posible, por lo que en ese caso la privacidad es obviada.

La Fig.3(b) simplifica la actividad esperada de un software compatible con PRoFIT instalado en un dispositivo IoT o en una plataforma intermedia. Conforme a los requisitos de privacidad, es necesario el consentimiento expreso del usuario (p.ej. el propietario del dispositivo o el administrador de la plataforma) sobre la instalación de este software (P1 y P2). Una vez instalado, el software controlará que las operaciones que se soliciten desde el propio sistema (p.ej. almacenar evidencia de forma local) o desde terceros (p.ej. solicitar evidencias) no interfieran con el consentimiento del usuario sobre los niveles de privacidad esperados. Una vez comprobada la petición y los derechos de ejecución, la petición se ejecuta y los resultados (p.ej. datos que deben ser enviados a un tercero o bien datos que deben ser almacenados en el dispositivo) se procesan de acuerdo al principio de minimización (P4).

C. Fases del investigador

La *recolección basada en el contexto* (fase 2) y el *análisis y correlación* (fase 3) corresponden al proceso de investigación. Ambas fases se muestran en la Fig.4(a).

Basándonos en los tres perfiles de dispositivo involucrados definidos (Sección A - ofendido, sospechoso y ofensor), durante la fase 3, el investigador recaba los datos

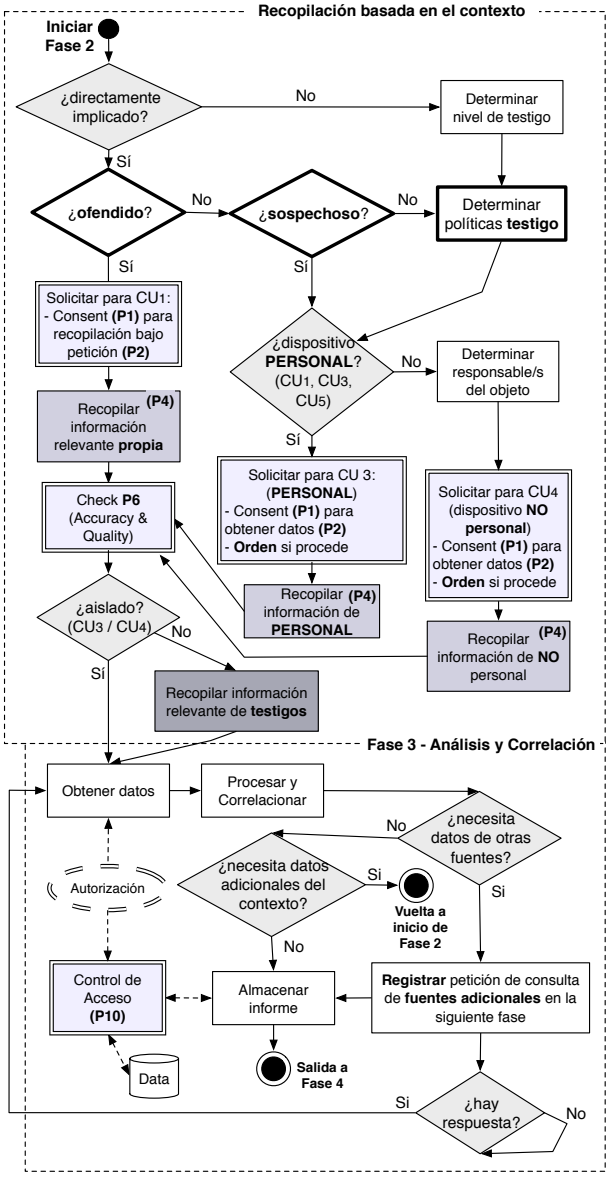
conforme al modelo PRoFIT considerando la privacidad de las fuentes. Destacamos que este enfoque se dirige a recabar las evidencias electrónicas de dispositivos que seguirán *bajo el control de sus propietarios o responsables durante todo el proceso*. De otra forma, se aplicarían los cauces tradicionales. En esta fase es donde se realiza la diferenciación entre los casos de uso (CU) a fin de establecer políticas de seguridad y privacidad de grano fino conforme al contexto.

Si el dispositivo a analizar es el *ofendido*, asumimos que se conoce la identidad del propietario del dispositivo personal, o la del responsable del dispositivo en caso de ser un objeto no personal (p.ej. porque hayan presentado una denuncia). En el consentimiento que el solicitante debe firmar (P1) tiene que constar el propósito para el que se usarán los datos (P2) - en este caso, constará que la investigación fue a petición del propietario/responsable. El investigador además debe comprobar que el dispositivo pertenece al usuario o que el responsable tiene autorización para solicitar la investigación. Entonces se realiza la recopilación de las evidencias electrónicas del dispositivo/plataforma que sean relevantes, empleando herramientas forenses, realizando pruebas para asegurar la integridad de la evidencia electrónica y documentando el proceso (P6).

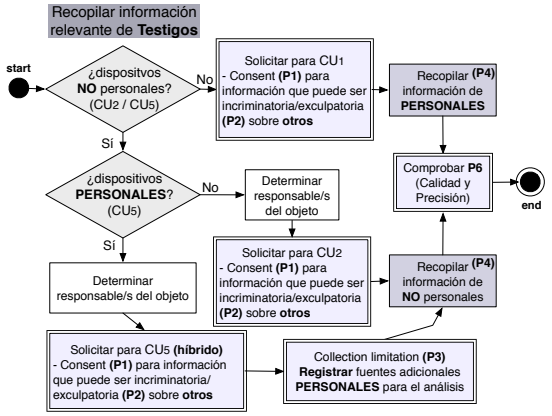
Si la investigación lo requiere, se comprueba entonces si el dispositivo estaba aislado, o si había otros dispositivos en su entorno. Por ejemplo, para determinar si hubo un posible contagio (p.ej. malware), o bien por si otro de los dispositivos tiene más datos que ayuden a la investigación del caso (p.ej. testigos digitales [5]). La Fig. 4(b) aporta información sobre los procesos a seguir para solicitar la colaboración de los posibles testigos, diferenciando también entre los diferentes contextos para determinar los responsables. Respecto a estos pasos, destacar que, considerando la propiedad P3, no se inician los procesos para el análisis de nuevos dispositivos personales si no es necesario (p.ej. si con la información recabada ya se puede resolver el caso).

Por otra parte, si el dispositivo a analizar pertenece a un *sospechoso*, los permisos solicitados serán diferentes. En este caso probablemente el investigador requiera poseer una orden judicial (*warrant*). A diferencia del ofendido o los posibles testigos, el sospechoso no tendrá un interés especial en la colaboración - a excepción de si su dispositivo contiene evidencias exculpatorias. Tanto en este caso como si el objeto es un *testigo*, deben determinarse los responsables del objeto, en caso de no ser un dispositivo personal (p.ej. porque se encuentre un dispositivo sospechoso y no se sepa quién tuvo acceso a este, o porque el testigo forme parte de una organización).

Por último, si el dispositivo va a ser analizado como *testigo*, incluimos una cláusula que indica que los datos recopilados no pueden ser usados para inculpar al testigo. Si un testigo puede ser un potencial ofensor, el investigador debe cuidarse de los datos que se recaban siguiendo este procedimiento, o no recabarlos. Cabe destacar, que si un sospechoso aporta sus evidencias electrónicas (i) se expone



(a) Preconfiguración del entorno IoT



(b) Recabar información relevante de testigos

Fig. 4. Fases - Investigador

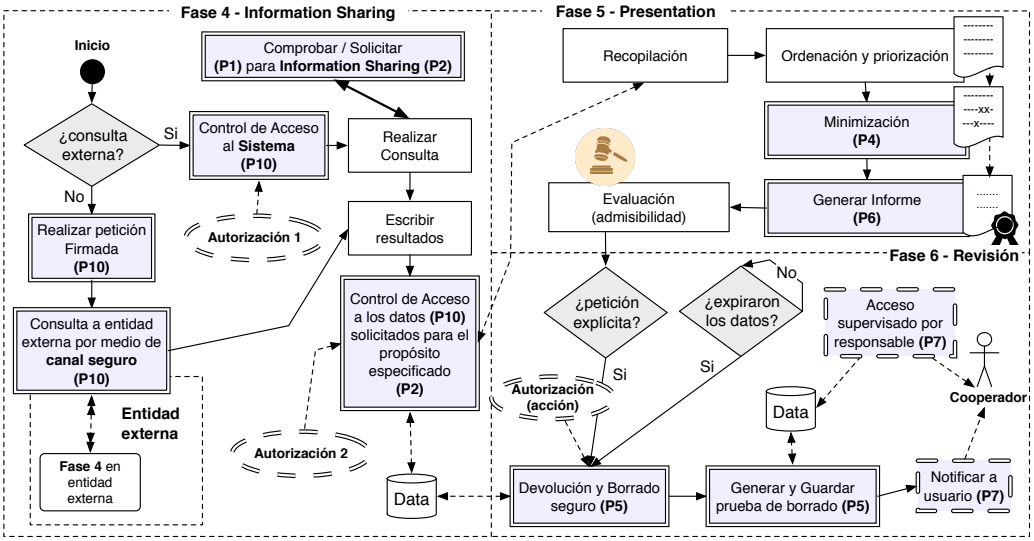


Fig. 5. Fases - Otros

a ser conocido y (ii) no invalida las evidencias electrónicas aportadas por otros, por lo que para el sospechoso resulta más sencillo eliminar sus evidencias electrónicas que ampararse en la cobertura de actuar como testigo.

Los datos recabados por este procedimiento pasan a la fase de *análisis de datos y correlación* (fase 3). Durante esta fase de la investigación se accede a las evidencias electrónicas almacenadas para su procesamiento y correlación. Como ya indicaba la Fig. 1, si durante esta fase se identifica que deben recabarse nuevas evidencias electrónicas se vuelve a la fase anterior. Así mismo, si el investigador considera que debe consultar otros datos - consultas externas - que pueden ser relevantes para el análisis, se pasa a la siguiente fase para obtener la información que sirva de retroalimentación. En ese caso, una vez que se tienen los resultados deben volver a consultarse los permisos de acceso a los datos por si durante el tiempo de la consulta - que podría tardar días o más tiempo - dichos permisos caducaron o fueron retirados.

Cabe destacar que los permisos P1 y P2 - que afectan al usuario - no se observan en esta fase, bajo la premisa de que los datos aquí tratados fueron proporcionados bajo los consentimientos proporcionados en la fase anterior.

D. Fases con participantes externos y de distinto perfil

Las tres últimas fases involucran distintos tipos de participantes. Estas tres últimas fases heredan los principios del modelo ESDFIM, aunque el modelo PRoFIT las redefine para su adaptación a la IoT y para integrar los requisitos de privacidad listados en la Tabla I. Comprende las fases de compartición de información (fase 4), presentación (fase 5) y revisión (fase 6).

La *compartición de información* involucra a usuarios externos (p.ej. otras entidades legales) con autorización para acceder a los datos. En este caso, consideramos que el propietario del objeto al que pertenecen las evidencias electrónicas tal vez no hubiese dado permisos para compartir esta información - o un resumen de ésta - con otras entidades externas, por lo que se comprueba el consentimiento dado y el propósito indicado (P1,P2). El flujo mostrado en la Fig. 5 sirve tanto para consultas sobre datos de diferentes investigaciones llevadas a cabo por la misma agencia, hasta consultas realizadas a entidades externas. Se consideran dos criterios de autorización; uno por el uso de los servicios de acceso a los datos y otra para poder realizar cambios sobre los datos.

La fase de *presentación* tiene como objeto generar el informe forense de forma que sea entendible por actores involucrados en el caso que no tienen que ser necesariamente expertos en la materia (p.ej. abogados, jueces) (Fig. 5). Considerando que durante las fases anteriores diversa información ha podido ser generada, durante esta fase se garantizará que la calidad y precisión de los datos son suficientes como para esclarecer el caso (P6). Se evitará dar más detalle del estrictamente necesario así como la aparición de datos sobre terceras partes no involucradas directamente en el caso (P4). Por lo tanto, se ordenan los resultados obtenidos y la minimización se aplica si hay datos redundantes, o bien datos que finalmente no son relevantes para el informe final. Dicho informe se genera teniendo en cuenta los requisitos de calidad (p.ej. legibilidad, entendimiento, información relevante al caso), y por último se procede a su evaluación, consistente en su admisibilidad para el caso.

Finalmente, el objetivo de la fase de *revisión* es eliminar de forma segura las evidencias electrónicas pasado un tiempo que no puede ser inferior a la duración de un caso. Entonces, se procede a eliminar el material de las bases de datos (P5) y a notificar al usuario de su eliminación. Este paso de notificación, así como el acceso supervisado para que el implicado compruebe que no constan sus datos (P7), son opcionales pero necesarios. Durante esta fase también se devuelven las posibles evidencias físicas (p.ej. PC) de haberse combinado el enfoque PRoFIT con los procedimientos tradicionales.

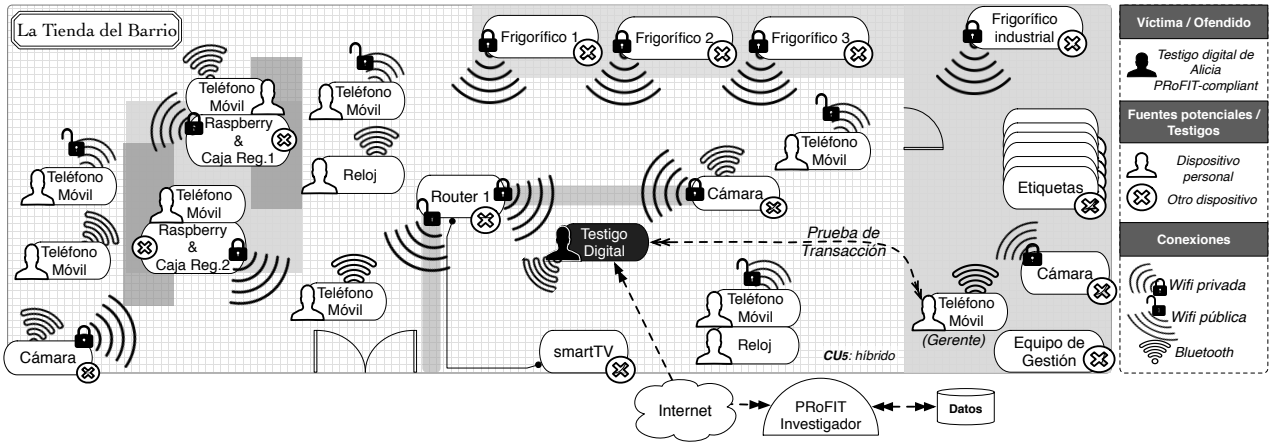


Fig. 6. Caso de uso 1: tienda de comestibles

IV. CASOS DE USO - TESTIFICACIÓN DIGITAL

En esta sección se propone dos casos de uso que se enfocan dentro del contexto de *testificación digital* [5]. Un testigo digital es un dispositivo IoT con características de seguridad probadas capaz de recolectar y salvaguardar evidencias digitales de su entorno. Suponemos que los testigos digitales que se muestran en estos casos de uso son *PRoFIT-compliant*. Esto significa que el software preinstalado en los testigos digitales, cuyos requisitos básicos se definen en [5], se encuentran adaptados para implementar las fases del modelo PRoFIT descritas en la sección III.

El primer caso de uso muestra un escenario de infección por malware en el que se pone de manifiesto la flexibilidad del modelo PRoFIT, mientras que el segundo caso de uso muestra un escenario en el que es necesaria una orden de registro. En este último caso, las opciones de privacidad son prácticamente nulas, no obstante también se siguen las fases del modelo PRoFIT. Mostramos así que el modelo propuesto permite balancear las características de privacidad según el contexto, y que los casos más tradicionales/restringidos basados en órdenes judiciales siguen respetándose.

A. Malware social - La tienda del barrio

Alicia tiene un teléfono móvil con software PRoFIT instalado (fase 1). Supongamos que Alicia entra en una tienda en la que hay varios dispositivos IoT, tanto personales como no personales (Fig. 6).

Durante su estancia en la tienda, el teléfono móvil de Alicia detecta un ataque que procede de un dispositivo del entorno. Tras detectar el ataque, el software PRoFIT decide almacenar la información relativa al mismo. Además, hace un hash a las evidencias recabadas y alerta a Alicia. Parece que algún dispositivo del entorno está infectado e intenta propagar un gusano aprovechando una vulnerabilidad en la aplicación *deals4U*, que utiliza la tecnología Bluetooth para escanear otros dispositivos del entorno y así conocer las ofertas del día y el número de unidades de alimentos disponibles (p.ej. en los frigoríficos).

Tras ser notificada de la ofensa, Alicia, que decide que este incidente debe ser reportado lo antes posible, envía las evidencias almacenadas al sistema PRoFIT, solicitando así el inicio de una investigación forense (fase 2). Entonces, el sistema asigna un investigador PRoFIT para el caso, que analiza los datos proporcionados (fase 3) y confirma que se trata de un ataque lanzado de forma local. Sin embargo, no tiene evidencias suficientes para poder llevar a cabo la investigación y sugiere al agente PRoFIT instalado en el dispositivo de Alicia recabar nuevas evidencias de otros dispositivos que quieran colaborar (vuelta a la fase 2).

Siguiendo la metodología (Sección III), el agente local PRoFIT pregunta primero a los dispositivos no personales, buscando a su responsable, en este caso el gerente de la tienda (derecha Fig. 6), para obtener la autorización. El responsable accede a colaborar y autoriza que sus dispositivos envíen información al investigador PRoFIT, usando el agente PRoFIT de Alicia como pasarela. Esta información se cifra y firma, y el agente tiene que emitir una prueba de que estos datos fueron enviados al agente PRoFIT remoto. El gerente de la tienda puede usar esta prueba para solicitar al investigador PRoFIT tanto (i) una comprobación de los datos que ha proporcionado, como (ii) retractarse y solicitar la eliminación de su declaración.

A la luz de las nuevas evidencias aportadas por el gerente de la tienda (fase 3), los resultados de la investigación apuntan a que el malware está latente en una de las Raspberry Pi de las cajas registradoras y que llegó a través del router, según los logs de este último. A partir de aquí la investigación continúa con el objetivo de llegar al origen del problema. Para ello, Alicia consiente que los datos proporcionados puedan ser compartidos con terceras partes (fase 4).

Tras unos días, una versión mejorada del malware causa daños en otros dispositivos IoT. Afortunadamente, el sistema PRoFIT guardó información sobre los inicios del ataque. La correlación con otras pruebas de un sistema externo permite determinar la procedencia del malware y se detiene a un sospechoso. Entonces, algunos de los datos aportados por Alicia y otros dispositivos se utilizan para elaborar el informe final (fase 5), que finalmente es admitido a juicio. Tras producirse la sentencia y transcurrido

un tiempo, los datos de los cooperadores son eliminados del sistema PRoFIT (fase 6).

Aunque este es un escenario hipotético y el malware, así como la aplicación *deals4U*, son ficticios, no es descabellado que ataques de este tipo pudieran producirse (o se estén produciendo) sin que el usuario lo perciba [13].

B. Registro en un almacén

El agente de policía Juan tiene un dispositivo que es testigo digital con capacidad de custodia, es decir, un *custodio digital*. Este tipo de testigo digital tiene privilegios en cuanto a que pertenece a un agente de la ley.

Juan tiene que realizar un registro en un almacén en el que hay varios dispositivos IoT (p.ej. cámaras, sensores y actuadores, etc.). Se sospecha que alguno de los dispositivos guarda evidencias electrónicas que pueden ser claves para resolver una investigación.

Para realizar el registro eficazmente, en el custodio se almacena una orden de registro firmada y es preconfigurado para recabar evidencias relevantes para el caso, conforme a los permisos y propósitos detallados en la orden judicial (fase 1). Nótese que, en este caso, la primera fase ejecuta ambos flujos de preparación: la tradicional - adaptada para automatizar la recogida de evidencias electrónicas - y del dispositivo IoT (Fig. 1).

Durante el registro, Juan es el especialista encargado de almacenar las evidencias digitales volátiles usando su custodio digital. Para ello, su dispositivo escanea la red del almacén y guarda el estado de las conexiones. También recibe los volcados de memoria y otros datos que Juan decide almacenar en el dispositivo. Todos estos pasos se hacen obviando las solicitudes y consentimientos de usuario porque se tiene una orden judicial para llevar a cabo los procedimientos que está realizando cabo Juan.

Una vez en el laboratorio, durante el análisis (fase 3) los datos recabados se procesan y se extraen las evidencias electrónicas relevantes para la investigación. En este caso particular, no se requieren consultas a bases de datos externas (fase 4). Los informes finales se redactan (fase 5), las evidencias son aceptadas para su vista y, transcurrido un tiempo, los objetos recabados durante el registro, de los que se extrajeron las evidencias, se devuelven a su dueño (fase 6).

V. CONCLUSIONES Y TRABAJO FUTURO

En este artículo definimos el modelo PRoFIT para la investigación forense en entornos IoT. A diferencia de otros enfoques, este modelo integra requisitos de privacidad (ISO/IEC 29100:2011) como parte de su metodología. El objetivo es promover la cooperación controlada de dispositivos IoT - personales o no - en investigaciones forenses. Para facilitar la comprensión del modelo se desarrollan dos casos de uso - análisis de la propagación de malware en una tienda de comestibles y recabación de evidencias electrónicas en base a una orden de registro.

Como trabajo futuro, queremos extender el trabajo actual para definir cómo usar este modelo en el contexto de la testificación digital, como un mecanismo de mitigación para los problemas de privacidad que pueden encontrarse en este tipo de contextos.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la Junta de Andalucía a través del proyecto FISICCO (TIC-07223), y por el Ministerio de Economía y Competitividad a través de los proyectos SMOG (TIN2016-79095-C2-1-R) e IoTest (TIN2015-72634-EXP).

REFERENCIAS

[1] K. Kyei, P. Zavarsky, D. Lindskog, and R. Ruhl, "A review and comparative study of digital forensic investigation models," in *International Conference on Digital Forensics and Cyber Crime*. Springer, 2012, pp. 314–327.

[2] S. Watson and A. Dehghantanha, "Digital forensics: the missing piece of the internet of things promise," *Computer Fraud & Security*, vol. 2016, no. 6, pp. 5–8, 2016.

[3] E. Oriwoh, D. Jazani, G. Epiphaniou, and P. Sant, "Internet of things forensics: Challenges and approaches," in *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on*. IEEE, 2013, pp. 608–615.

[4] T. Geller, "In privacy law, it's the us vs. the world," *Communications of the ACM*, vol. 59, no. 2, pp. 21–23, 2016.

[5] A. Nieto, R. Roman, and J. Lopez, "Digital witness: Safeguarding digital evidence by using secure architectures in personal device," *IEEE Network*, In Press.

[6] S. Perumal, N. M. Norwawi, and V. Raman, "Internet of things (iot) digital forensic investigation model: Top-down forensic approach methodology," in *Digital Information Processing and Communications (ICDIPC), 2015 Fifth International Conference on*. IEEE, 2015, pp. 19–23.

[7] V. R. Kebande and I. Ray, "A generic digital forensic investigation framework for internet of things (iot)," in *Future Internet of Things and Cloud (FiCloud), 2016 IEEE 4th International Conference on*. IEEE, 2016, pp. 356–362.

[8] E. Oriwoh and P. Sant, "The forensics edge management system: A concept and design," in *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*. IEEE, 2013, pp. 544–550.

[9] S. Zawoad and R. Hasan, "Faiot: Towards building a forensics aware eco system for the internet of things," in *Services Computing (SCC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 279–284.

[10] Organisation for Economic Co-Operation and Development (OECD), "The OECD Privacy Framework," 2013, [Last Access: 02/2017]. [Online]. Available: http://www.oecd.org/internet/ieconomy/privacy-guidelines.htm

[11] The European Parliament and the Council of the European Union, "Regulation (eu) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," 2016, [Last Access: 02/2017]. [Online]. Available: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[12] *ISO/IEC 29100:2011 - Information technology - Security techniques - Privacy framework*, JTC 1/SC 27 Std., 2011. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45123

[13] S. Peng, S. Yu, and A. Yang, "Smartphone malware and its propagation modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 925–941, 2014.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6577

3GPP NB-IoT, tecnología y herramientas de medida

Almudena Díaz Zayas, Pedro Merino Gómez
Departamento de Lenguajes y Ciencias de la Computación,
Universidad de Málaga, Andalucía Tech,
Edificio de Investigación Ada Byron, Málaga, España, 29071
almudiaz@lcc.uma.es, pedro@lcc.uma.es

F. Javier Rivas Tocado
Keysight Technologies
Málaga, Spain
Email: javi_rivas@keysight.com

***Resumen*—La primera versión de los estándares 3GPP Narrow Band-IoT (NB-IoT) se finalizó en Junio de 2016 como parte de la Release 13. NB-IoT es una nueva tecnología de acceso radio que puede coexistir con los despliegues actuales de GSM, UMTS y LTE. De hecho, la especificación de NB-IoT se ha integrado en los estándares LTE. NB-IoT va un paso más allá que la especificación de MTC (Machine Type Communication), enfocándose en dispositivos de extremadamente bajo coste, despliegues masivos y tasas de transmisión reducidas con un ancho de banda de solamente 200 kHz (de ahí su nombre). En este artículo realizaremos una revisión en detalle de las especificaciones 3GPP destacando las modificaciones necesarias sobre los despliegues LTE tradicionales para proporcionar conectividad a dispositivos de usuario (UEs) Cat-NB1. Asimismo, introduciremos una perspectiva novedosa sobre las ventajas de usar de nuevas herramientas y soluciones para el análisis y medida de comunicaciones NB-IoT, como el UXM E7515A de Keysight Technologies, proporcionando también lecciones aprendidas en el incipiente uso de esta nueva tecnología.**

***Palabras Clave*—LTE, Narrowband-IoT, Machine Type Communication, entorno de pruebas, Internet de las cosas**

I. INTRODUCCIÓN

Como resultado del study item TR 45.820 [1] de 3GPP, titulado 'Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)', se ha propuesto una nueva tecnología radio llamada NB-IoT. NB-IoT se está integrando en los estándares LTE y se ha estandarizado una nueva categoría de dispositivos (UEs) in [2], llamada Cat-NB1. Las funcionalidades proporcionadas por NB-IoT se han introducido previamente en [4][5] y están orientadas a cubrir el segmento inferior de las comunicaciones máquina a máquina, también conocidas como Machine Type Communication (MTC): cobertura extendida, soporte de un número masivo de dispositivos de baja tasa de transmisión, alta latencia permitida, dispositivos de ultra-bajo coste, bajo consumo de potencia y arquitectura de red optimizada.

Como se muestra en la Tabla I, los requerimientos de NB-IoT son los mismos que para eMTC (enhanced MTC) [3], pero con el foco en escenarios masivos MTC en el segmento de bajo coste. NB-IoT proporciona un coste por dispositivo incluso inferior que eMTC, y una mayor extensión de cobertura con un presupuesto de pérdidas de propagación para el enlace (MCL) de 164dB (al menos en despliegue aislado). La duración de la batería se ha buscado que pueda alcanzar los 10 años con una batería de 5 Vatios Hora. NB-IoT es compatible con los modos mejorados de recepción discontinua (eDRX) introducidos en la Release 13 de 3GPP para reducir el consumo de potencia. El número de dispositivos soportado por celda es de unos 55.000.

Al contrario que eMTC, que únicamente puede ser desplegada en modo in-band, utilizando bloques de recursos dentro de una portadora LTE normal, NB-IoT puede desplegarse también en los bloques de recursos sin usar dentro de la banda de guarda (guard-band) de la portadora LTE, o en modo aislado (standalone) para despliegues en espectro dedicado. NB-IoT resulta particularmente adecuado para el reaprovechamiento de los canales GSM.

Una prueba de concepto inicial se llevó a cabo por la Universidad de Aalto y Ericcson Research en Finlandia. Estos test iniciales tuvieron lugar en modo standalone y usando una arquitectura Cloud RAN (C-RAN). La capacidad de NB-IoT de soportar altas latencias lo hace adaptarse muy bien a los caos de uso que involucran despliegues basados en C-RAN.

En este artículo describimos la evolución de los estándares 3GPP relacionados con MTC/NB-IoT y detallamos los cambios introducidos en LTE para proporcionar soporte a NB-IoT, el último estándar de 3GPP para IoT. El artículo está organizado como sigue. En la sección II, proporcionamos el contexto necesario sobre los estándares 3GPP para MTC/IoT para entender el papel de NB-IoT en la era IoT. La sección III describe las modificaciones introducidas en la red de acceso radio. La transformación

	eMTC (Enhanced MTC) Release 13	NB-IoT (Release 13)
Coverage Enhancement	15 dB gain better compared to GPRS	20 dB gain compared to GPRS
Reduced complexity	Similar to EGPRS modem cost	Ultra-low cost (~1$)
Power efficiency	10 year battery life	10 year battery life
Latency	10 sec	10 sec
Capacity	∼10000 devices per cell	∼50000 decives per cell
Coexistence	LTE in-band	GSM standalone, LTE in-band, LTE guard-band

Cuadro I
EMTC VS NB-IOT FEATURES

arquitectural sufrida en el EPC se explica en la sección IV. En la sección V realizamos una aproximación novedosa a NB-IoT desde un punto de vista de prueba y medida, adelantando unas primeras lecciones aprendidas. Finalmente, en la sección VI, proporcionamos una comparación adicional entre MTC y NB-IoT, e introducimos mejoras adicionales de Releases posteriores.

II. CONTEXTO

3GPP (Third Generation Partnership Project) es el organismo de estandarización que especifica los sistemas de comunicaciones móviles LTE/LTE-Advanced, así como 3G UTRA y 2G GSM. Los estándares 3GPP están estructurados en Releases. La definición de la tecnología móvil conocida como Long Term Evolution (LTE) se inició en 2005 y las primeras especificaciones 3GPP se introdujeron en la Release 8 en Diciembre de 2007. El término 'user equipment' (UE), se ha usado tradicionalmente por 3GPP para referirse a los dispositivos celulares utilizados por los suscriptores para acceder a los servicios de red móvil. Un UE puede ser un Smartphone o un dispositivo empotrado contenido en un equipo de comunicaciones máquina a máquina (M2M). Para soportar múltiples tipos de UE con distintas capacidades, 3GPP define categorías diferentes en 3GPP 36.306. Las categorías difieren en las máximas tasas de transmisión soportadas en los enlaces ascendente y descendente, que también están asociadas, por ejemplo, con el soporte de multiplexación espacial (transmisión MIMO). También se utilizan las categorías en los eNodeB (estacione base LTE) para determinar las condiciones bajo las cuales tendrá lugar la comunicación con el UE. Como se elaborará en más detalle a continuación, el concepto de categoría de UE tiene mucha importancia en IoT. La Tabla II proporciona un resumen de las categorías de terminales de usuario orientadas a escenarios MTC/IoT.

La versión inicial de los estándares LTE Machine Type Communication (MTC) se introdujo en la release 8 y estaba basada en la Categoría 1. La Categoría 1 era la de menor capacidad con una tasa de transmisión máxima de 10Mbs en el enlace descendente y de 5 Mbps en el enlace ascendente. Por otra parte la Categoría 5, la de mayor capacidad en Release 8, soportaba tasas de transmisión de 300Mbps en el enlace descendente y 75 Mbps en el ascendente. Aunque los dispositivos con Categoría 1 no soportan transmisión MIMO, aun así, incorporan 2 antenas receptoras y soportan todas las opciones de ancho de canal de RF desde 1.4 a 20 MHz. Además, esta categoría no cumple con los requerimientos de duración de batería, coste y rango necesarios para IoT. La primera Release de MTC estaba enfocada en la optimización de los mecanismos de tarificación, direccionamiento, ubicación fija, poca movilidad y baja actividad de los terminales, manejo de un gran volumen de suscripciones y de datos de usuario en la red, gestión de problemas en servicios M2M y en aspectos de seguridad

En la Release 12 se introdujo una nueva categoría de UE, la Categoría 0, que proporciona reducciones de coste de aproximadamente el 50 % comparada con la Categoría 1. También se introdujo un nuevo modo de ahorro de potencia en esta Release, en 3GPP TS 24.301 y 23.682. El principal propósito de esta nueva funcionalidad es reducir el consumo de energía mientras que el dispositivo no está transmitiendo ni recibiendo. Otras soluciones para MTC se discutieron en Release 12 como parte de 3GPP 36.888. La especificación de MTC UE Cat0 incorpora otras mejoras como la transmisión half-duplex, una única cadena de RF y la reducción de la tasa pico de transmisión. La reducción de coste usando un ancho de banda reducido y las mejoras de cobertura se pospusieron a la Release 13.

Continuando el trabajo normativo iniciado en Release 12 para mejorar la adecuación de LTE al prometedor mercado IoT, el objetivo clave de eMTC (frecuentemente referenciado como LTE-M) en Release 13 es la definición de una nueva categoría de UE de baja complejidad que soporte un ancho de banda reducido, baja potencia de transmisión, menor soporte de modos de transmisión descendentes, duración de batería ultra-larga mediante reducción de consumo y operación con cobertura extendida. A los dispositivos de este tipo se les asigno la categoría Cat-M1 (anteriormente conocida como CAT-M). En la Release 13 MTC consigue una reducción adicional del 50 % del coste mediante la restricción del ancho de banda a 1.5MHz, lo cual tiene un fuerte impacto en el diseño del receptor, reduciendo la complejidad del procesamiento de banda base. Esta Release también introduce mejoras de cobertura para ampliar el alcance en 15 dB, permitiendo a los operadores alcanzar dispositivos MTC en condiciones de cobertura pobre como contadores colocados en sótanos. Estas dos mejoras tienen un impacto importante en el diseño de los canales físicos y lógicos que tienen que soportar modos de bajo ancho de banda y cobertura mejorada. Otra mejora clave en Release 13 es la introducción de mejoras en recepción discontinua (eDRX) [6] que está basada en el uso de temporizadores más largos de DRX (recepción discontinua) para alcanzar reducciones adicionales de consumo de potencia. El concepto de DRX

	Release-8	Release-12	Release-13	Release-13
	Cat. 1	Cat. 0	Cat. M1	Cat. NB1
Donwlink peak	10 Mbps	1 Mbps	1 Mbps	200 kbps
Uplink peak rate	5 Mbps	1 Mbps	1 Mbps	144 kbps
Number of antennas	2	1	1	1
Duplex mode	Full duplex	Half duplex	Half duplex	Half duplex
UE receive bandwidth	20 MHz	20 Mhz	1.4 MHz	200 kHz
UE transmit power	23 dBm	23 dBm	20 dBm	23 dBm
Complexity	100 %	50 %	20 %	15 %
Use case	Voice services for emergency in elevators,Smart Grid Management, Kids/Elderly/Pet tracking	Cat0 is the interim solution prior Cat-M. Cat 0 is used for replacing Cat1, but cannot replace voice use cases.	Environment monitoring, Vehicle tracking	Smart metering, smart buildings, home automation
Availability	Available	Available	2017	2017

Cuadro II
EVOLUTION OF UE CATEGORIES FOR MTC/IOT

consiste en la monitorización de los canales descendentes durante períodos de tiempo limitados. Durante el resto del tiempo el dispositivo se duerme para reducir el consumo de potencia.

En Mayo de 2014, Huawei y Vodafone propusieron un study item para NB-M2M al grupo 3GPP GERAN que rápidamente consiguió un fuerte soporte y creciente atención de otros operadores líderes. En Octubre del mismo año, QC envió una nueva propuesta de narrowband IoT llamada NB-OFDM. En Mayo de 2015, ambas tecnologías se fusionaron en NB-CIoT (NarrowBand Cellular IoT).

Mientras tanto, Ericcson aceleró su investigación en narrowband IoT y propuso NB-LTE (Narrowband LTE) en agosto de 2015. En Septiembre de 2015, el 3GPP aceptó la inclusión de ambas tecnologías como un Work Item en Release 13 [1]. NB-CIoT es un enfoque desde cero (clean slate) promovido por Huawei. La principal diferencia entre NB-LTE y NB-CIoT viene de cuanto de las redes LTE actuales se puede reusar para IoT. NB-CIoT requiere nuevos chipsets y no es compatible hacia atrás con ninguna red LTE anterior a Release 13. NB-LTE, por contraste, podría ser completamente integrada en las redes LTE, funcionando en las bandas LTE actuales sin necesidad de una red solapada. En Noviembre de 2015, 3GPP acordó que ambas iniciativas evolucionaran hacia un único estándar llamado Narrowband IoT (NB-IoT). NB-IoT especifica un nuevo acceso radio para IoT celular, basado en buena medida en una variante no compatible hacia atrás de LTE, que se orienta a una cobertura mejorada en interiores, soporte para un número masivo de dispositivos de baja tasa de transmisión, escenarios poco sensibles al retardo, dispositivos de ultra bajo coste, bajo consumo de potencia y una arquitectura de red optimizada. Así, NB-IoT se ha convertido en la principal solución de 3GPP para redes de área extensa y bajo consumo LPWAN (Low Power Wide Area Network), reemplazando a las propuestas NB-LTE y NB-CIoT (Cellular IoT).

En Junio de 2016 3GPP completó la estandarización de NB-IoT, la nueva tecnología radio de banda estrecha desarrollada para la Internet de las cosas (IoT). Los UEs NB-IoT se denominan como Cat-NB1 (también referenciados en la literatura como Cat-M2). La reducción de su complejidad, comparada con Cat.1 es de hasta el 90 %. Habiendo terminado la estandarización de NB-IoT en Junio de 2016, llevó solamente 9 meses estandarizar esta nueva tecnología después de la fase de estudio, lo que demuestra la importancia de IoT para el 3GPP.

El trabajo llevado a cabo en el grupo de GSM/EDGE dentro de 3GPP para cobertura extendida de GSM para IoT (EC-GSM-IoT), se ha integrado en estos dos proyectos: eMTC (LTE-M) y NB-IoT.

En paralelo con estos procesos de estandarización llevados a cabo en 3GPP, hay otras iniciativas que operan de forma no licenciada. Este es el caso de LoRaWAN, Sigfox y OnRamp Wireless, Wihtless -N & -P, etc. La mayoría de estas redes hace uso de las bandas de frecuencia no licencias ISM para uso industrial, científico y médico. Testas tecnologías están actualmente disponibles, han sido desplegadas y cumplen con los cuatro factores para LPWAN (largo alcance, muy baja potencia, baja tasa de transmisión y bajo coste). Algunas están basadas en protocolos apoyados por alianzas industriales como la LoRaWAN Alliance y Wightless SIG, otros están basados en protocolos propietarios y otros son estándares en progreso. Además de las diferencias tecnológicas, hay que destacar que algunas de estas tecnologías están sujetas a nuevos modelos de negocio. En un plano económico, una ventaja adicional de NB-IoT respecto de estas tecnologías, puede venir del mayor consenso entre operadores, la integración en las redes celulares y de la resultante economía de escala.

III. RED DE ACCESO RADIO NB-IOT

NB-IoT proporciona acceso radio a servicios de red usando una capa física [8] optimizada para muy bajo consumo y coste. El ancho de banda del canal completo es de 180 kHz, la separación entre subportadoras puede ser de 15kHz o de 3.75 kHz (solamente en el uplink), el esquema de modulación más alto es QPSK, hay soporte para operación FDD y half-duplex. El enlace descendente de NB-IoT se basa en OFDM y el esquema de transmisión usa un único bloque de recursos físicos (PRB). El enlace

ascendente se basa en single-carrier FDMA. Para la transmisión ascendente, hay dos modos de operación posibles, transmisión de tono único (single tone) y multi-tono (multi-tone). En tono único, está permitido tanto 3.75Hz como 15KHz de espaciado entre subportadoras, mientras que en multi-tono solamente se puede usar espaciado de 15KHz. La transmisión multi-tono permite agrupar conjuntos de 3,6 o 12 subportadoras. De forma adicional, la duración mínima de las unidades de recursos usadas en la planificación depende del número de subportadoras asignadas y del modo de operación, yendo desde 1 ms en transmisión multi-tono con 12 subportadoras, a 32ms en transmisión de tono único de 3.75kHz.

Hay ciertas funciones de protocolos E-UTRA (Evolved Universal Terrestrial Radio Access) que soportan todos los UEs Rel-8 que sin embargo no son usadas por los UEs NB-IoT: movilidad entre distintas tecnologías radio (inter-RAT), traspasos, reporte de medidas, funciones de alerta públicas, tasas de datos garantizadas (GBR), grupos cerrados de usuarios (CSG), soporte de femtoceldas (HeNBs), de nodos intermedios (relaying), agregación de portadoras, conectividad dual, NAICS (Network Assisted Interference Cancellation and Suppression), comunicaciones de difusión y multipunto MBMS, servicios de tiempo real, evitación de interferencia para coexistencia interna, interoperabilidad con WLAN asistida por la red, comunicación y descubrimiento entre dispositivos (sidelink), reducción del número de drive tests (MDT), llamadas de emergencia y conmutación a llamadas de circuitos (CS fallback).

En NB-IoT, se puede soportar el posicionamiento usando la arquitectura existente de Servicios de Localización (LCS) mediante medidas únicamente en el eNB. Se introducen cinco nuevos canales:

- Narrowband Physical Broadcast
- Narrowband Physical Downlink Shared Channel (NPDSCH). Transporta el DLSCH (Downlink Shared Channel) y el PCH (Paging Channel) para UEs NB-IoT.
- Narrowband Physical Downlink Control Channel (NPDCCH). Informa al UE NB-IoT de la asignación de recursos del PCH y del DL-SCH, y transporta el 'scheduling grant' asignado al UE NB-IoT para transmisión en el enlace ascendente.
- Narrowband Physical Uplink Shared Channel (NPUSCH). Transporta el UL-SCH (Uplink Shared Channel) y la confirmación ACK/NACKs Hybrid ARQ en respuesta a la transmisión descendente hacia el dispositivo UE NB-IoT.
- Narrowband Physical Random Access Channel (NPRACH). Transporta el preámbulo de acceso aleatorio enviado por el UE NB-IoT.

Respecto de la capa de control de acceso al medio (MAC), NB-IoT introduce cambios para reducir el consumo de potencia y para hacer el funcionamiento del planificador más flexible y simple. Debido a los menores requisitos de tasa de transmisión, se utiliza un único proceso HARQ. Esto permite eliminar el identificador de HARQ de las asignaciones del planificador y así usar un menor número de bis para una mayor eficiencia y robustez. Adicionalmente, las retransmisiones en el enlace ascendente (UL) dejan de ser síncronas sino siempre adaptativas y asíncronas tanto en UL como en DL. Esto proporciona a la red un control más ajustado de planificación del enlace ascendente, mientras que protege de retransmisiones periódicas indeseadas ya que ahora se generan únicamente cuando sean solicitadas explícitamente.

Otras mejoras en las capas 2 y 3 de NB-IoT incluyen la reducción de los tamaños máximos de los elementos de datos (SDU) y control (PDU) de la capa de convergencia de paquetes de datos (PDCP). De los anteriores 8188 octetos se pasa a un máximo de 1600 bytes. Este es un compromiso razonable ya que el tráfico tradicional de internet no excede de 1500 bytes por paquete IP.

Mientras que un dispositivo NB-IoT está en modo conectado, es posible configurar recepción discontinua (DRX) con ciclos de hasta 10.24 segundos. Respecto de la configuración de Paging, cuando se habilita DRX en modo de espera (idle), en NB-IoT la máxima duración del ciclo de DRX es de 10485.76 segundos (2.91 horas). Esto permite ahorros de energía muy notables cuando el caso de uso es compatible con la alta latencia que tendrá asociada el establecimiento de conexión.

IV. ARQUITECTURA NB-IOT

La arquitectura de NB-IoT simplifica la arquitectura existente del núcleo de red (EPC) con el propósito de cumplir con los requisitos y modelos de tráfico identificados en [1] y mostrados en la Tabla III. La transmisión de pequeños datos se basa en una arquitectura simplificada que está orientada a transportar estos pequeños bloques de información sobre mensajes de señalización con el número de red o Non Access Stratum (NAS). Se usa un nodo dedicado del núcleo de red para un perfil CIoT (terminología 3GPP para Internet de las cosas), proporcionando funciones combinadas de plano de control y de usuario, e.g. agregando funciones que tradicionalmente residían en el MME (Mobility Management Entity) y en el SGW (Serving Gateway), y en algunas instancias de PGW (Packet Data network Gateway) en una nueva entidad lógica llamada C-SGN(CIoT Serving Gateway Node). C-SGN puede implementarse para soportar únicamente la funcionalidad necesaria para casos de uso CIoT. S1-lite es una versión optimizada de S1-C (el plano de control del interfaz S1 que se basa en el protocolo S1AP) entre el eNodeB y el MME. Del S1-C solamente se soportan los mensajes S1AP necesarios (y dentro de los mensajes solamente aquellas partes requeridas) para los procedimientos CIoT. Se soportan también procedimientos de seguridad optimizados así como algunas funcionalidades para permitir el transporte de datos. El plano de usuario se transporta en estos mensajes S1AP modificados para permitir un manejo eficiente de datos pequeños. Los túneles del plano de datos del interfaz S1-U (definido en TS 23.401[9]) no son necesarios.

Esta arquitectura permite que ciertas modificaciones (por ejemplo, soporte nativo de SMS en el dominio de

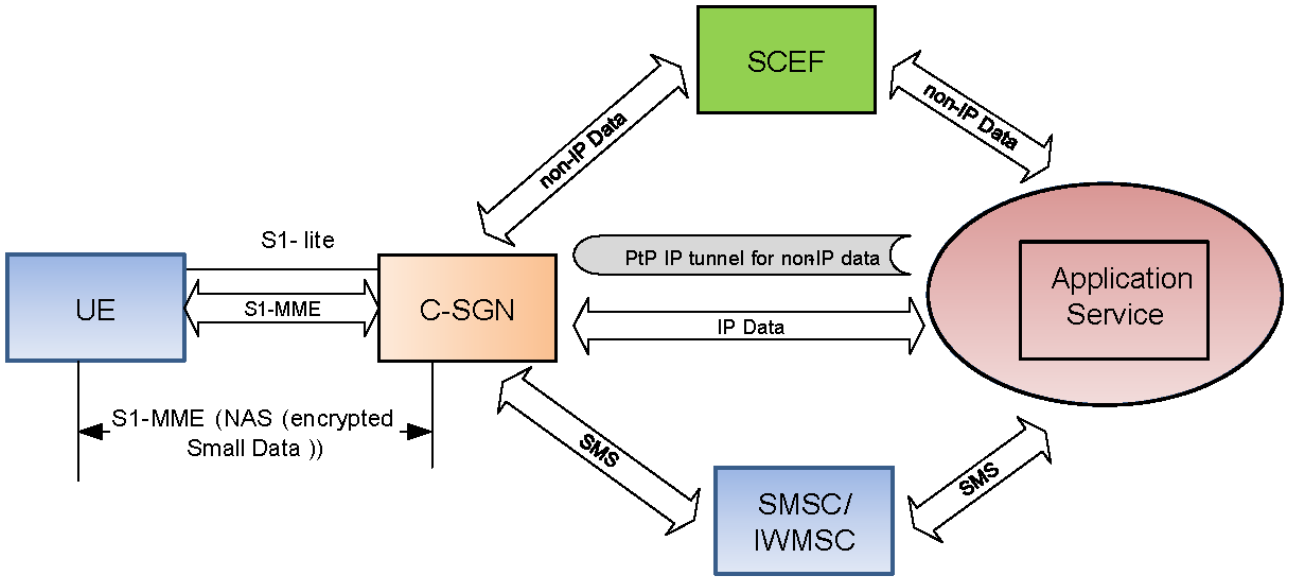


Figura 1. LTE Architecture for NB-IoT

Category	Application example	UL Data Size	DL Data Size	Frequency
Mobile Autonomous Reporting (MAR) exception reports	smoke alarm detectors, power failure notifications from smart meters, tamper notifications etc.	20 bytes	0 ACK payload size is assumed to be 0 bytes	Every few months; Every year
Mobile Autonomous Reporting (MAR) periodic reports	smart utility (gas/water/electric) metering reports, smart agriculture, smart environment etc.	20 bytes with a cut off of 200 bytes i.e. payloads higher than 200 bytes are assumed to be 200 bytes.	50 % of UL data size ACK payload size is assumed to be 0 bytes	1 day (40 %), 2 hours (40 %), 1 hour (15 %), and 30 minutes (5 %)
Network Command	Switch on/off, device trigger to send uplink report, request for meter reading	0 - 20 bytes 50 % of cases require UL response.	20 bytes	1 day (40 %), 2 hours (40 %), 1 hour (15 %), and 30 minutes (5 %)
Software update/reconfiguration model	Software patches/updates	200 bytes with a cut off of 2000 bytes i.e. payload higher than 2000 bytes are assumed to be 2000 bytes.	200 bytes with a cut off of 2000 bytes i.e. payload higher than 2000 bytes are assumed to be 2000 bytes.	180 days

Cuadro III
TRAFFIC MODELS FOR CELLULAR IOT [3GPP TR 45.820[1]]

paquetes, registro sin conexión PDN, NAS simplificado, etc.) se apliquen solamente a UEs CIoT sin requerir que los procedimientos mantengan compatibilidad hacia atrás con otros UEs.

En NB-IoT, para optimizar la señalización, se han introducido dos soluciones además del tradicional establecimiento de conexión RRC (Radio Resource Control). La primera solución, Data-over-NAS (DONAS), es obligatoria y es una optimización del plano de control basada en la solución 2 propuesta en [7]. DONAS permite la transmisión de datos sin tener que activar el plano de usuario. Implementando solamente esta opción permite implementaciones de UE con menos requisitos relativos al plano de usuarios, como el soporte para podo sin confirmación (UM) en RLC, múltiples canales lógicos dedicados, DTCH, re-establecimiento RLC ... La segunda solución es opcional y está basada en la solución 1 (especificada en [7]). La suspensión y restauración de RRC introduce mejoras para deshabilitar y recuperar el plano de usuarios de forma eficiente. En el plano de control optimizado, los datos se envían sobre Non Access Stratum (NAS), directamente desde la Mobility Management Entity (MME) en el núcleo de red hacia el UE sin interacción de la estación base.

Los procedimientos de suspensión y recuperación RRC reducen la sobrecarga de señalización y también mejoran la duración de la batería del UE. La arquitectura de protocolos de NB-IoT y LTE se separa en plano de usuario y control. El plano de control consiste en protocolos que gestionan las portadoras de acceso radio y la conexión entre el UE y la red. La capa más alta del plano de control se denomina capa de no acceso o Non-Access Stratum (NAS) y realiza el intercambio de señalización entre el UE y el núcleo de red (EPC), pasando de forma transparente a través de la red radio. Es el responsable de la autenticación, control de seguridad, gestión de movilidad y gestión de portadoras. La capa de acceso o Access Stratum (AS), es el nivel funcional por debajo de

NAS, y en el plano de control consiste en el protocolo de control de recursos radio (RRC). RRC configura los planos radio de control y de usuario de acuerdo al estado de la red. Hay dos estados principales a nivel RRC, inactivo (RRC_Idle) o conectado (RRC_Connected), y la entidad RRC controla la conmutación entre estos estados. En estado RRC_Idle, la red sabe que el UE está presente en la red y que que se puede ser alcanzado en caso de una llamada o conexión entrante. En estado RRC_Connected el UE tiene una conexión radio activa con el eNodeB, la red sabe la posición del UE a nivel de celda y el UE puede transmitir y recibir datos. Cuando no hay tráfico se libera la conexión, pasando a RRC_Idle, para ahorrar batería y recursos radio. Los procedimientos de suspensión y recuperación reducen la sobrecarga requerida para la transición del estado del UE de Idle a Connected para establecer el plano de usuario, y de vuelta a estado Idle, reduciendo los mensajes de señalización requeridos en comparación con la operativa tradicional.

De cara a soportar los procedimientos de suspensión/recuperación RRC, MME, eNodeB y UE deben incluir nuevas funcionalidades. El MME, que es responsable de gestionar la movilidad del UE, debe soportar nuevos procedimientos S1AP, en particular el procedimiento disparado por los procedimientos 'RRC Suspend' y 'RRC Resume' en el interfaz radio. El MME tiene asimismo que almacenar la información de Contexto del UE cuando la conexión RRC es suspendida y la asociación S1AP se mantiene. El eNodeB debe permitir el almacenamiento del contexto del UE así como de los parámetros relacionados con la asociación S1AP cuando el UE está en el nuevo estado CIoT RRC-Idle. Además, el eNodeB tiene que soportar los nuevos procedimientos para suspender y recuperar la conexión. Estos procedimientos deberán implementarse entre el eNB y el UE, así como entre el eNB y el MME. Finalmente el UE debe almacenar la información relevante de AS cuando el UE entra en estado CIoT RRC_Idle y soporta los procedimientos de suspensión y recuperación de la conexión RRC.

Data-overNAS se considera como la base del trabajo normativo para soportar transmisiones infrecuentes de datos (para datos IP, no-IP y SMS). El soporte de DONAS es obligatorio tanto para el UE como para la red. Esta solución permite un soporte eficiente para pequeñas transmisiones infrecuentes para IoT y soporte para datos no-IP. La solución se basa en una arquitectura ligera del núcleo de red mostrada en la Figura 1. Esta solución encapsula pequeñas cantidades de datos en el enlace ascendente en el mensaje inicial NAS, que se ha extendido, y usa un mensaje NAS descendente adicional para transportar respuestas pequeñas. De esta forma, se puede evitar el esfuerzo requerido para establecer el plano de usuario, e.g. mensajes RRC asociados y establecimiento de seguridad AS. No hay establecimiento de DRBs y portadoras S1-U. Los datos se pueden entregar también a través del SCEF (Service Capability Exposer Function), que se usa para la entrega de datos no-IP sobre el plano de control. El elemento SCEF es parte de las mejoras de arquitectura

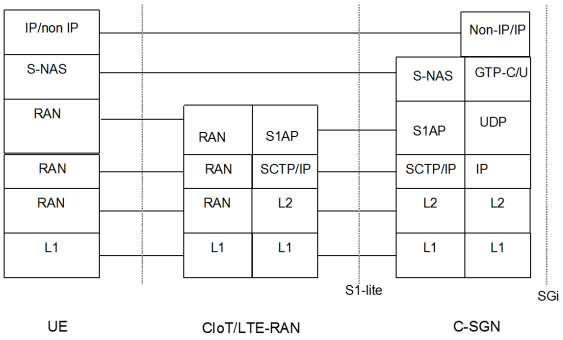


Figura 2. Protocol stack for NB-IoT

exposición de capacidades de servicios (AESE) [10] introducidas en Release 13, que muestra los servicios de red a terceras partes. La implementación de esta solución también implica modificaciones en los procedimientos disponibles en el UE, eNodeB y C-SGN/MME para soportar e.g. la gestión de sesiones 'NAS small data' y el registro sin activación de portadoras.

La pila de protocolos resultante tras la adopción de estas dos soluciones es la pila simplificada mostrada en la Figura 2, donde S-NAS indica una señalización NAS simplificada.

V. UNA PERSPECTIVA DE PRUEBA Y MEDIDA

Como se ha explicado en detalle, NB-IoT es una tecnología completamente nueva que busca adaptar las redes celulares a las demandas de conectividad y a la problemática de los dispositivos de bajo coste como sensores y contadores. Proporcionando conectividad directa, sin intermediarios, a todos los dispositivos, se abre un amplio abanico de nuevos escenarios y de nuevas oportunidades para casos de uso, servicios y dispositivos que aparecerán en un futuro y que no se hayan imaginado a día de hoy.

Esta revolución tecnológica no solamente va a traer grandes ventajas, sino que va a presentar importantes retos y desafíos a muy distintos niveles a los actores involucrados. Desde el punto de vista de un fabricantes de nuevos dispositivos o de un operador que tiene que incorporarlos a su red surge la necesidad de verificar los diseños de radiofrecuencia, comprobar el procesamiento de banda base, asegurar una correcta interoperabilidad en los distintos protocolos, verificar la estabilidad de los dispositivos y la medir las dinámicas de consumo de potencia y longevidad de las baterías. Para investigadores que buscan proponer entender la tecnología y proponer evoluciones futuras, es importante entender las dependencias entre los distintos parámetros y poder ajustar sus modelos caracterizando el impacto de la configuración de red en el enlace de comunicación NB-IoT y en los servicios que hacen uso del mismo. Para todo ello, proponemos el uso de un elemento de referencia como el UXM E7515A de Keysight Technologies [11], que ha sido validado para NB-IoT y eMTC CAT-M por el Global Certification Forum (GCF).

Usando como ejemplo un objetivo clave y ambicioso de NB-IoT, como es conseguir verificar que la duración

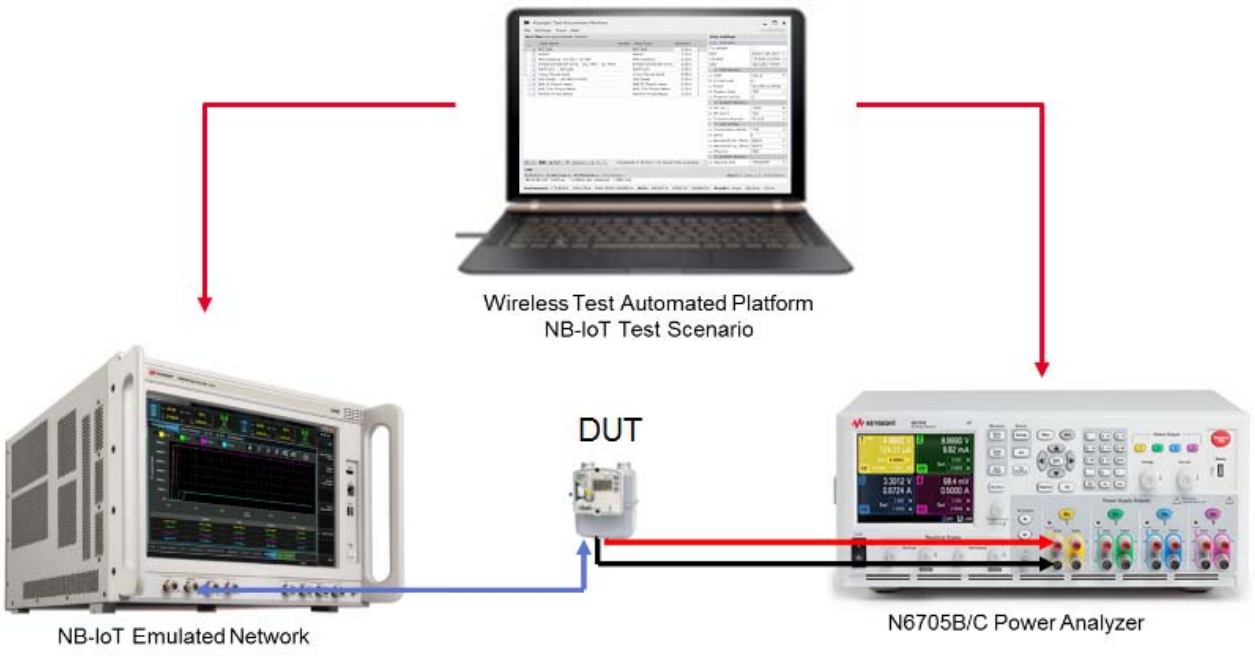


Figura 3. Entorno de pruebas y medidas para NB-IoT

Figura 4. Mensajes de control entre el UE y el eNodeB

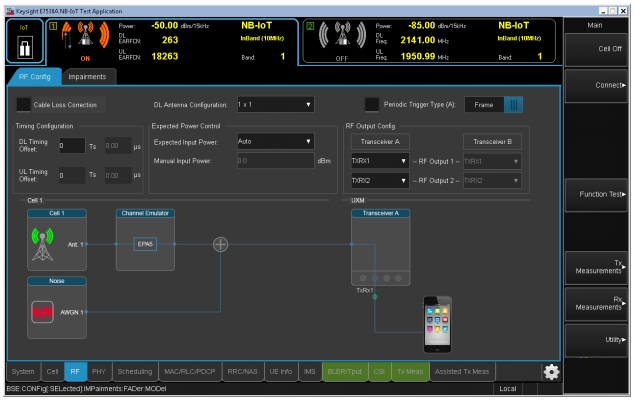


Figura 5. Emulación de propagación radio y de interferentes

de batería alcance 10 años, se requiere de herramientas de análisis de consumo de energía muy precisas y de mecanismos para realizar experimentos en condiciones realistas de funcionamiento, ya que el consumo real de un dispositivo variará en buena medida en función de las condiciones en las que se encuentre. Como se muestra en la Figura 3, esto es posible combinando las capacidades de emulación de red del UXM con un analizador de potencia de precisión como el N6705B/C de Keysight.

Conectando un dispositivo NB-IoT real al UXM, es posible reproducir una red con múltiples celdas NB-IoT, y controlar y monitorizar tanto la señalización (ver Figura 4) y los parámetros de configuración los protocolos de bajo nivel como la potencia recibida por el dispositivo y las condiciones de propagación incluyendo canales con desvanecimientos e interferentes, como se muestra en la Figura 5. Por otra parte, utilizando la aplicación de medida de la serie X integrada en el UXM, es posible realizar numerosas medidas de RF y de banda base. En la Figura 6 se muestra un ejemplo de análisis de modulación de la señal NPUSCH transmitida por un dispositivo NB-IoT. Además de la constelación IQ, es posible observar el espectro de potencia y consultar multiples indicadores de calidad, así como verificar la potencia transmitida.

En pruebas de concepto iniciales, hemos constatado el tremendo impacto que pueden llegar a tener las condiciones de propagación y la configuración de red, aumentando el tiempo de ida y vuelta de la comunicación en órdenes de magnitud en función del número de repeticiones necesarias. En la Figura 7 se muestra un escenario de conectividad IP extremo a extremo en condiciones favorables, aún así el tiempo de ida y vuelta es de unos 200 ms como se puede apreciar por la separación entre el tráfico DL y UL.

Asimismo, debido la capacidad extremadamente reducida de NB-IoT, en pruebas iniciales hemos verificado que

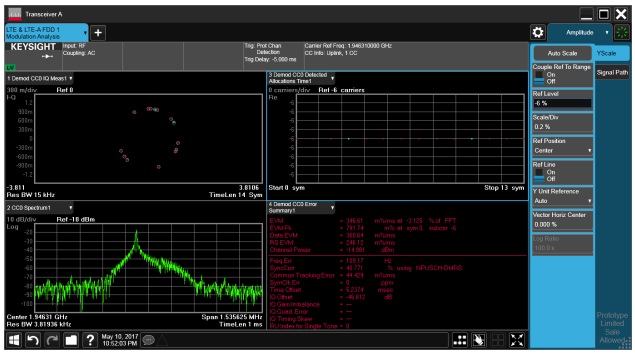


Figura 6. Medidas de espectro, potencia de señal y calidad

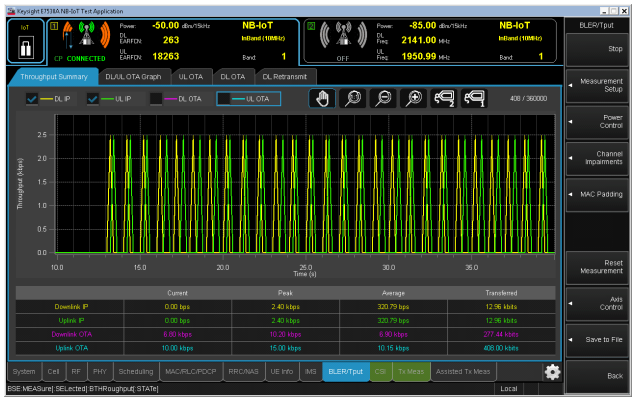


Figura 7. Tráfico IP extremo a extremo de baja tasa binaria

puede ser muy dañino que cualquier tráfico aunque sea residual proveniente de un equipo conectado a un dispositivo NB-IoT. Por tanto, en donde los dispositivos NB-IoT vayan a funcionar como modem, hemos identificado la necesidad de utilizar de forma sistemática firewalls u otros mecanismos que dejen pasar solamente el tráfico deseado. Otra interesante lección aprendida es que algunas de las herramientas tradicionales de prueba de tráfico tienen problemas para operar en escenarios de muy baja tasa de transmisión y particularmente ante picos de retardo muy elevados como los presentes en algunas configuraciones de NB-IoT, por ello es recomendable utilizar soluciones contrastadas.

En trabajos futuros analizaremos de forma detallada las dependencias entre los principales parámetros, realizando experimentos repetibles y automáticos mediante el software de control Wireless Test Automation Platform.

VI. CONCLUSIONES

Mientras que eMTC (LTE-M) es una tecnología puramente LTE, NB-IoT se basa en un nuevo interfaz radio que puede coexistir con los sistemas LTE, UMTS y GSM actuales. Respecto de los casos de uso de ambas tecnologías, la principal diferencia es la movilidad y la tasa de transmisión. NB-IoT ofrece tasas y movilidad reducidas en comparación con eMTC, pero permite una mayor cobertura y densidad, lo que la hacen más adaptada a escenarios IoT, como las redes de sensores mientras que eMTC presentaría ventajas en nichos como el de los dispositivos wearables. En Releases posteriores, se tiende a introducir mejoras adicionales para IoT en la forma de soporte para multicast, reducción de latencia para posibilitar escenarios (V2V), posicionamiento, movilidad y mejoras en la continuidad de servicio y nuevas clases de potencia entre otras.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio Español de Economía y Competitividad (TIN2015-67083-R), FEDER y el programa de investigación e innovación Horizonte 2020 de la Unión Europea (grant agreement No 688719).

REFERENCIAS

[1] 3GPP TR 45.820 Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)
[2] 3GPP TS 36.101 Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception
[3] 3GPP 36.888, Study on provision of low-cost Machine-Type Communications (MTC) UEs based on LTE
[4] J. Gozalvez, "New 3GPP Standard for IoT [Mobile Radio],ïn IEEE Vehicular Technology Magazine, vol. 11, no. 1, pp. 14-20, March 2016.
[5] S. Landström, J. Bergström, E. Westerberg, D. Hammarwall, "NB-IoT: A sustainable technology for connecting billions of devices", in Ericsson Technology Review, vol. 93, no. 3, pp. 2-11, April 2016
[6] 3GPP TR 23.770, Study on System Impacts of Extended DRX Cycle for Power Consumption Optimization
[7] 3GPP TR 23.720 Study on architecture enhancements for Cellular Internet of Things
[8] 3GPP TS 36.300 LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radico Access Network (E-UTRAN); Overall description; Stage 2
[9] 3GPP TS 23.401 General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access
[10] 3GPP TR 23.708 Architecture Enhancements for Service Capability Exposure (AESE)
[11] UXM Wireless Test Set, http://www.keysight.com/en/pd-2372474-pn-E7515A/uxm-wireless-test-set?cc=US&lc=eng

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6450

Sistema videowall de bajo coste basado en Raspberry Pi, personalizable y configurable dinámica y remotamente vía Web

Pau Salvador, Fernando Boronat, Mario Montagud, Dani Marfil
Departamento de Comunicaciones, Immersive Interactive Media R&D (IIM) Group,
Universitat Politècnica de València – Campus de Gandia
Calle Paraninf, 1, 46730, Grao de Gandia, Valencia (SPAIN)
{pasallla@epsg, fboronat@dcom, mamontor@, damarre@dcom}.upv.es

***Resumen*- En este artículo se presenta una propuesta de sistema videowall, tanto con distribución uniforme como disforme, formado por dispositivos de bajo coste, personalizable y configurable dinámica y remotamente vía Web. Se han analizado los sistemas de videowall existentes y las diferentes posibilidades en cuanto a su desarrollo con dispositivos de bajo coste. Se ha optado por la utilización de Raspberry Pi como dispositivo hardware y OMXPlayer como software reproductor multimedia. La sincronización entre los diferentes reproductores se consigue mediante el uso de una señal de reloj común y el intercambio de mensajes de control entre dispositivos, así como adoptando técnicas (agresivas y suavizadas) de ajuste de los procesos de reproducción. El sistema incluye una interface web muy amigable que facilita su gestión y control de manera remota.**

***Palabras Clave*- videowall, Raspberry Pi, OMXPlayer, NTP**

I. INTRODUCCIÓN

Un videowall consiste en un conjunto de múltiples pantallas (monitores de PC, videoproyectores, paneles LED o TVs) que reproduce contenido multimedia de manera sincronizada, simulando una pantalla de mayores dimensiones. Entre las utilidades de los videowalls se pueden citar la creación de pantallas gigantes, paneles publicitarios (aplicación comercial), paneles informativos (p. ej. información sobre salidas y llegadas en aeropuertos, estadísticas…), etc. Su uso está muy extendido, entre muchos otros ejemplos, en escaparates, centros comerciales, aeropuertos y cartelería digital.

Las pantallas del videowall pueden estar distribuidas de forma *uniforme*, es decir, pantallas de idéntico tamaño distribuidas de forma matricial, bien formando un muro plano o bien curvo (en configuración NxM, con N filas de M pantallas cada una), o bien de forma *disforme*, con pantallas de igual o diferentes tamaños y en diferentes orientaciones, para mostrar uno o varios elementos multimedia, como imágenes, vídeos, etc. (ver Fig. 1).

Aunque en el mercado ya existen numerosas soluciones de videowall comerciales, la mayoría requieren de hardware (HW) especial y software (SW) propietario, lo cuál las convierte en soluciones de coste muy elevado. Como ejemplos, se pueden citar los sistemas de *Userful*[1], *MagicInfo* de Samsung[2] y el sistema de LG[3]. Además, la mayoría de soluciones están basadas en la utilización de controladores del videowall o servidores especializados. Los precios suelen ser bastante altos, la gestión suele ser en local, y la configuración y expansion de los mismos por parte de los usuarios finales es prácticamente imposible.

[1] https://www.userful.com/ (último acceso: mayo 2017)
[2] http://www.samsung.com/es/business/solutions-services/smart-signage-solutions/smart-signage-solutions/magicinfo-videowall (último acceso: mayo 2017)
[3] http://www.lg.com/us/business/commercial-display/displays-tvs/video-walls (último acceso: mayo 2017)

Fig. 1. Videowall uniforme plano (izquierda) y disforme (derecha)

En este artículo, se presenta la arquitectura, tanto HW como SW, de una solución económica[4] para constituir un sistema de videowall formado por pantallas distribuidas, ya sea de forma uniforme o disforme, versátil, personalizable, y configurable dinámicamente y de manera remota vía IP, mediante una interfaz web. Su desarrollo se ha basado en el uso de dispositivos de bajo coste, Raspberry Pi (RPi en adelante), y se controla a través de una aplicación web, sin necesidad de instalar ninguna aplicación de control determinada, específica para cada plataforma o dispositivo.

Las pantallas del videowall se pueden combinar de varias formas: 1) todas las pantallas formando una única pantalla de dimensiones máximas, reproduciendo cada una una porción del mismo contenido (ver Fig. 1); 2) las pantallas se pueden agrupar de manera personalizable para diferentes contenidos formando secciones (Fig. 2).

En cada una de dichas pantallas, la imagen de vídeo correspondiente debe ser reproducida de forma perfectamente sincronizada con el resto. En caso contrario, una diferencia de unos pocos fotogramas entre las distintas pantallas de la misma sección del videowall puede ser notable y molesta (especialmente, cuando hay un cambio de plano o de escena), ya que las pantallas están muy cercanas o pegadas unas a otras.

Fig. 2. Videowall 2x3 con dos secciones (2x2 y 2x1)

En el videowall propuesto, la reproducción de la parte del vídeo representada en cada pantalla está controlada por un dispositivo electrónico independiente, ejecutando un SW reproductor multimedia (player) independiente. Cada dispositivo reproduce el contenido con una referencia de reloj independiente, y, por tanto, de manera distinta, ya que los relojes, aunque tengan la misma frecuencia nominal, generalmente presentan diferentes grados de precisión y desviación de la misma. Con el objetivo de conseguir el efecto de videowall, con todas las partes del vídeo siendo reproducidas de forma sincronizada, se deberá establecer alguna solución de sincronización para conseguir que la reproducción de las distintas partes del contenido por los distintos dispositivos y pantallas esté completamente sincronizada. En la sección IV, se presenta la solución adoptada en el videowall propuesto.

El artículo está estructurado de la siguiente manera. La sección II recopila trabajos anteriores relacionados con la creación de videowalls. En la sección III se describen las configuraciones HW y SW del sistema videowall propuesto. En la sección IV, se analizan las diferentes posibilidades de sincronización de las pantallas del videowall, con sus ventajas e inconvenientes. La sección V presenta la aplicación web para la gestión y control del videowall. Por último, el artículo finaliza con la sección VI, con las conclusiones y líneas de trabajo futuro.

II. TRABAJOS RELACIONADOS

En esta sección, se revisan las tecnologías para el desarrollo de videowalls, las ventajas e inconvenientes de las implementaciones de videowall existentes, así omo una comparativa cualitativa de las mismas con la propuesta presentada. Además, se presentan algunos trabajos relacionados y se comparan con el presentado.

Actualmente, existen tres opciones en cuanto a las tecnologías de acceso y distribución de contenidos multimedia para ser reproducidos en las pantallas que forman el videowall (Fig. 3): 1) distribuir el contenido multimedia sobre estándares de distribución HW de video (normalmente, basados bien en *splitters* o divisores de señal de vídeo, o bien en pantallas especiales que se conectan entre sí); 2) distribuir el contenido por streaming IP; o 3) acceder directamente a contenido almacenado en una unidad de disco compartida en red.

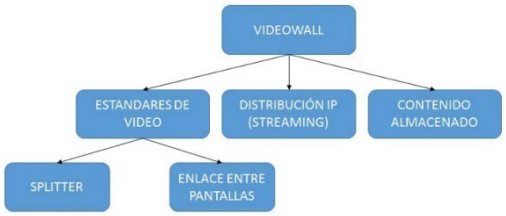


Fig. 3. Tecnologías videowall

Con respecto a la primera opción, por un lado, se pueden utilizar *splitters* de vídeo que son dispositivos activos que tienen una entrada y varias salidas de vídeo. Por la entrada se recibe el contenido del vídeo con la resolución completa a ser reproducido en el videowall, y el propio dispositivo se encarga de dividir dicho vídeo (sus imágenes) en diferentes partes (o no, si se desea que todas las pantallas reproduzcan exactamente el mismo contenido) que distribuirá a cada una de sus salidas a las que se conectarán las diferentes pantallas que forman el videowall. Esta opción está pensada para casos en los que todas las pantallas sean iguales y con configuraciones

[4] Los costes de los sistemas comerciales rondan los 26.000€ (LG) y los 15.000€ (MagicInfo y Userful) para videowall disforme de 3x3 con pantallas de 46" y el sistema propuesto con RPi tendría un coste de 9.500€.

sencillas. Tiene la ventaja de que es simple de instalar, pero tiene los inconvenientes de ser una solución poco escalable, se necesita HW específico, la fuente de contenido debe estar próxima a las pantallas y las opciones de configuración son bastante limitadas.

Por otro lado, dentro de la primera opción mencionada, también cabe la posibilidad de enlazar las pantallas para formar videowalls. Se basa en sistemas propietarios creados por determinadas empresas fabricantes de pantallas (ej. LG o NEC), que permiten que las pantallas se puedan enlazar entre ellas (por ejemplo, a través del DisplayPort en los sistemas NEC) para utilizarlas como un videowall. Esta opción tiene la ventaja de ser escalable (el sistema de NEC admite hasta 100 pantallas, en una matriz 10x10), ya que sólo hay que ir cableando las diferentes pantallas en serie. Sin embargo, tiene varios inconvenientes, como el coste elevado y el hecho de que todas las pantallas han de ser del mismo fabricante. Además, existe un problema asociado con el mantenimiento, ya que posibles futuras averías de pantallas implica sustituirlas por pantallas del mismo fabricante y, probablemente, modelo.

Con respecto a la segunda opción, basada en la distribución de los contenidos vía IP (streaming), cabe resaltar que es escalable y adaptable (p.ej., Samsung dispone de la solución *MagicInfo* que admite hasta 250 pantallas de gran formato), a la vez que muy simple en cuanto a instalación física se refiere, ya que sólo requiere la interconexión de todos los dispositivos mediante una red IP local. Sin embargo, será necesaria la disponibilidad de una red que proporcione el ancho de banda suficiente para poder distribuir los diferentes contenidos al mismo tiempo a las diferentes pantallas. También se requiere el uso de un ordenador potente que se encargue de dividr el vídeo en varios flujos (uno por cada pantalla) y realice el envío (streaming) vía IP del contenido correspondiente a ser reproducido en cada pantalla, especialmente cuando se realizan varios envíos simultáneos. La necesidad de dicho ordenador con suficientes recursos encarece el sistema. Además, para cada pantalla se requiere de HW y SW específicos (p.ej., dispositivos denominados *Zero clients* o RPi, con SW capaz de decodificar y reproducir el vídeo recibido por streaming) para decodificar el vídeo y entregarlos a los interfaces de salida o pantalla (por HDMI, por ejemplo). Con el sistema de distribución del contenido por IP podemos encontrar el sistema de Samsung nombrado anteriormente y el definido en [1].

La tercera opción indicada permite prescindir de dicho ordenador potente y, por tanto, abaratar el coste del videowall, si en cada pantalla se coloca un dispositivo de bajo coste que accede a contenido almacenado en una unidad de disco compartida. En dicho caso, cada dispositivo o bien un dispositivo de control se puede encargar de seleccionar la/s parte/s del vídeo a reproducir, sin necesidad de realizar streaming. De esta forma, se puede crear un videowall con dispositivos económicos y con menos recursos. En cuanto a inconvenientes, en este caso también se necesita una red que proporcione el suficiente ancho de banda para poder realizar los diferentes accesos simultáneos al contenido almacenado.

Por un lado, en el videowall propuesto se desea que tanto la configuración del contenido a reproducir en las diferentes secciones del videowall como la configuración del número de pantallas que componen las mismas sea dinámica y controlable remotamente. Por tanto, la primera opción de utilizar estándares de vídeo no ofrece la flexibilidad deseada y, por tanto, fue descartada desde el principio. Por otro lado, hoy en día existe electrónica de red de bajo coste que proporciona conectividad LAN con velocidad suficiente para trabajar con contenido almacenado, así como para cargar y descargar contenido a una unidad de disco compartida en red durante el funcionamiento del videowall, sin necesidad de implementar un servidor de streaming, lo cual encarecería el sistema. Es por ello que se ha optado, en esta primera versión, por la tercera opción, basada en el uso de contenido almacenado y de dispositivos RPi.

Se han encontado varias propuestas de videowalls basadas en RPi y su reproductor *OMXPlayer*: [1], [2] y [3], que se han mejorado en el presente trabajo. Dichas propuestas no satisfacen las necesidades mínimas para el videowall requerido, por tanto, se procede a realizar un desarrollo completo del sistema. La Tabla I presenta una comparativa resumida entre dichas propuestas y la presentada en este artículo.

Tabla I
COMPARATIVA TRABAJOS RELACIONADOS

	[2]	[3]	[1]	Sistema Propuesto
Programaciones	No	NS/NC	NS/NC	Si
Playlist	No	NS/NC	NS/NC	Si
Diferentes secciones	No	NS/NC	NS/NC	Si
Videowall disforme	Si	Si	No	Si
Rotar pantallas	No	No	No	Si
Slideshow de imágenes	No	No	No	Si
Sistema de sincronía	Si	Si	Si	Si
Configuración dinámica	No	No	Si	Si
Configuración vía web	No	No	Si	Si

III. DESARROLLO DEL SISTEMA VIDEOWALL

En esta sección se presenta toda la arquitectura HW y SW que se ha realizado para el desarrollo del sistema videowall, personalizable, controlable dinámica y remotamente vía web.

A. Funcionalidad

En la arquitectura propuesta se puede seleccionar cualquier combinación de pantallas, ya sea uniforme o disforme (incluso incluyendo pantallas con rotación) y formar grupos de pantallas con diferentes contenidos dentro del videowall completo (véase la Fig. 2). A estos grupos o combinaciones de pantallas se las denomina secciones del videowall, en adelante, en este documento.

Mediante una aplicación web (SW multi-plataforma) de control se puede gestionar y configurar el videowall, seleccionando el contenido a mostrar en cada una de las pantallas del mismo, sea cual sea su distribución. Además, la aplicación de gestión dispone de la posibilidad de programar las reproducciones (horarios de reproducción) de cada sección del videowall, reproducir *playlists*, poner la reproducción de los vídeos en bucle, así como controlar qué audio o audios se escucharán por los altavoces del videowall.

Por lo que respecta a los contenidos, como se ha comentado previamente, se almacenarán en un disco duro externo accesible a través de la red local por los dispositivos del videowall (RPi).

B. Hardware

Para el control y la gestión del videowall, se van a utilizar dispositivos RPi, en concreto, la Raspberry Pi 3 Modelo B[5]. Serán necesarios tantos como pantallas tenga el videowall, más uno adicional que se encargará del control de todo el sistema y de la comunicación con el dispositivo utilizado por el usuario para la gestión vía web. En la Fig. 4 se puede ver un ejemplo de videowall formado por una combinación 2x3. Para dicha configuración, el HW del sistema está formado por seis pantallas, altavoces, siete RPi, un disco duro externo USB 3.0, un switch Fast-Ethernet, y un dispositivo externo (móvil, tablet, PC, portátil…) con un navegador y conectividad IP.

Excepto la de control, cada RPi se conectará por HDMI a su pantalla correspondiente. Todas las RPi estarán conectadas en LAN a través de un switch a 100 Mbits/s. En este caso, la RPi de control y el dispositivo de usuario deberán tener conectividad IP (p. ej., de forma inalámbrica a través de un punto de acceso Wifi externo al videowall, preferiblemente, con conexión a Internet, para facilitar su configuración de forma remota).

A nivel de la red LAN, todos los dispositivos estarán configurados con direcciones IP estáticas de la red IP 192.168.0.0/16. En su conexión con esta red, a la RPi de control del videowall se le asignará siempre la IP 192.168.0.1/16. La RPi de control estará conectada a dos redes, la LAN 192.168.0.0/16 conectada a través del cable de Ethernet y también a una red (p. ej., Wifi) externa si se desea que el videowall pueda pueda ser configurado remotamente. La configuración IP de la conexión externa podrá ser estática o conseguirla a través de DHCP y permitirá la gestión remota del videowall.

Por un lado, en el caso de videowalls con distribución uniforme, para facilitar la gestión, aunque se pueda asignar cualquier dirección IP de forma dinámica y controlar las asignaciones mediante tablas, a cada RPi conectada a un monitor se ha decidido asignarle la dirección IP/máscara de red 192.168.x.y/16, siendo *x* el número de fila (siendo la fila superior la fila 1) e *y* el número de columna (siendo la columna de la izquierda la número 1), tal y como se muestra en la Fig. 4. Por otro lado, en el caso de videowalls con distribución disforme, la asignación de direcciones IP puede seguir cualquier lógica y quedará registrada en la base de datos del videowall. En este caso, se ha decidido numerar las pantallas y registrar en una tabla la asignación de direcciones a cada una de ellas.

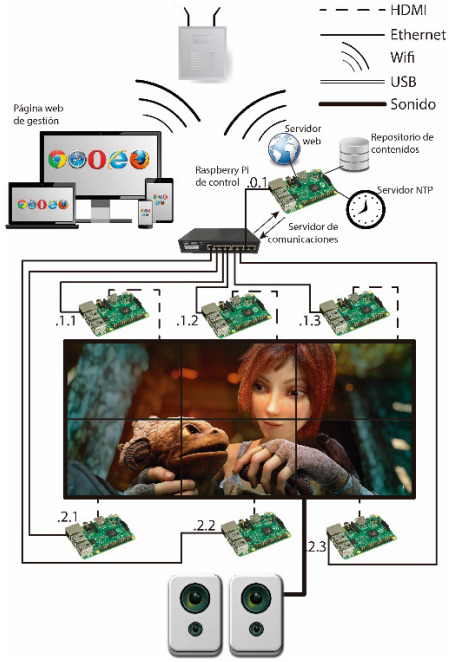


Fig. 4. Configuración del escenario

Para realizar un control de los diferentes flujos de audio a ser reproducidos por los altavoces del videowall, se ha diseñado un circuito electrónico para cada canal (Left/Right), que se muestra en la Fig. 5.

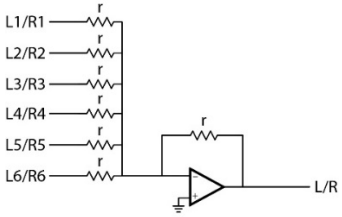


Fig. 5. Circuito electrónico del audio

Se puede apreciar que está formado por dos sumadores con amplificadores operacionales. En cada sumador están las seis salidas de cada canal de audio (L o R, respectivamente) de las RPi (L1 a L6 o R1 a R6), así como la entrada de los altavoces (L y R). Cada circuito cuenta con siete resistencias (r) en el sumador de cada canal, cuyo valor es indiferente. Todas las resistencias de cada sumador deben ser iguales para no amplificar ni

[5] La Raspberry Pi 3 Modelo B es un microcomputador de bajo coste que ofrece, entre otras, las siguientes prestaciones: Tarjeta 100 Base-T (Fast Ethernet), Wifi 802.11n, SW de código abierto y decodificación por HW de contenido H264 con una resolución de 1080p, con una tasa de 30 imágenes por segundo. Proyecto Raspberry. https://www.raspberrypi.org/ (último acceso mayo 2017)

atenuar las señales de audio. A través del SW de gestión del videowall, se pueden seleccionar los audios que se quieren reproducir y estos se sumarán en el circuito descrito.

C. Software

C.1. Sistema Operativo

Todas las RPi empleadas incorporan Raspbian[6] como sistema operativo (S.O.). Está basado en Debian, pero simplificado para estos dispositivos.

C.2. Servidor de Contenidos

En la RPi de control se ha habilitado un servidor FTP para el acceso a la unidad de disco USB compartida, en el que se han configurado los permisos adecuados para cada tipo de usuario. De esta forma, cada usuario sólo puede acceder a determinadas carpetas, y subir y/o seleccionar contenido para ser mostrado en las diferentes secciones del videowall.

C.3. Reproductor de vídeo

A la hora de escoger el reproductor de vídeo a ejecutar en las RPi, se tuvieron en cuenta dos opciones que incluyeran o permitieran implementar soluciones de sincronización entre la reproducción de diferentes instancias del mismo en diferentes dispositivos, conectados a través de redes IP. Las dos opciones fueron la plataforma *Gstreamer* [4] (incluyendo los elementos para la reproducción y sincronización) y el reproductor *OMXPlayer* [5].

Al principio, se instaló en las RPis el framework *Gstreamer* [4] y se intentó utilizar su reproductor, puesto que GStreamer también soporta de manera nativa una solución de sincronización multimedia entre dispositivos, ya utilizada por nuestro grupo en otros proyectos, en dispositivos basados en Linux. Se comprobó que el rendimiento de reproducción no era el deseado, y, por tanto, no fue posible reproducir contenido multimedia de manera correcta.

Sin embargo, el reproductor *OMXPlayer* [5] funcionó a la perfección y, por tanto, fue el escogido para el videowall propuesto. Dicho reproductor está específicamente creado para la RPi, y aprovecha la decodificación por HW del propio dispositivo, de forma que libera mucho la carga del sistema S.O. Para ejecutar este reproductor, se ha utilizado una librería desarrollada para Python[7], que permite controlar su proceso de reproducción a través de un sistema de comunicación entre procesos, denominado *D-Bus* (Desktop Bus[8]).

OMXPlayer no contempla la opción de reproducción de videos con un ángulo de rotación diferente a 90, 180 o 270 grados. Es por ello que, como el videowall propuesto, cuando se le configura una distribución disforme, permite colocar las pantallas en cualquier disposición, se ha implementado un *script* basado en la herramienta *ffmpeg*[9], que crea copias del contenido con el/los ángulo/s necesario/s. Dicho script se ejecuta tras la configuración de videowalls disformes con pantallas con inclinación, así como cada vez que se sube contenido a ser reproducido en ellas.

C.4. Servidor de Tiempo Global

Para conseguir la sincronización de la reproducción de las porciones de los vídeos correspondientes a cada pantalla de las secciones del videowall (es decir, entre los reproductores ejecutándose en cada una de las RPi), todos los procesos de reproducción deben disponer de una referencia de reloj global común. En el videowall propuesto, el reloj de todas las RPi estará sincronizado mediante un servidor NTP (Network Time Protocol, RFC 5905 [6]), instalado en la RPi de control. Se ha escogido la opción de NTP frente a PTP (*Precission Time Protocol*, estándar IEEE 1588), ya que la precisión que se puede conseguir en una LAN es más que suficiente para el videowall y no se requiere de HW específico.

C.5. Servidor de Comunicaciones

En la RPi de control se ha implementado un servidor de Websockets, basado en *Node.js*[10] y las librerías Socket.IO, para permitir el intercambio de mensajes entre todas las RPis que forman el videowall. Cada RPi, al iniciarse, se conectará a dicho servidor de Websockets y quedará a la espera de recibir mensajes de control (de la RPi de control, según la programación o configuración del contenido del videowall realizada) o de sincronismo (explicados en la siguiente sección).

En cuanto a los mensajes de control, cada Rpi puede recibir de la RPi de control, de forma unicast, tres posibles mensajes de tipo texto, iniciados con las cadenas "VIDEO", "AUDIO" o "IMAGEN", según el contenido a reproducirse en la pantalla asociada a dicha RPi, seguidas de varios parámetros separados por '%':

"VIDEO%ruta%X1%Y1%X2%Y2%posición%tiempo%master"
"ACCION%acción%tiempo"
"IMAGEN%ruta%X1%Y1%X2%Y2%posición%tiempo"

En el primer mensaje, relativo a contenido de vídeo, el parámetro *ruta* contendrá la ruta del vídeo a reproducir; los parámetros *X1* e *Y1* se corresponden con las coordenadas *x* e *y* de la esquina superior izquierda de la porción de video que ha de reproducir; *X2* e *Y2* se corresponden con las coordenadas *x* e *y* de la esquina inferior derecha de la porción de video que ha de reproducir; *posición* es la posición donde debe ir centrado el vídeo a reproducir y puede tener uno de los siguientes valores: {*L,R,U,D,C*} que equivalen a izquierda, derecha, arriba, abajo o centro, respectivamente, y sirve para

---

6 Distribución Raspbian. Página de descarga: https://www.raspberrypi.org/downloads/raspbian/ (último acceso mayo 2017)

7 OMXPlayer para Pyhton. Repositorio oficial: https://github.com/willprice/python-omxplayer-wrapper (último acceso mayo 2017)

8 https://www.freedesktop.org/wiki/Software/dbus/ (último acceso mayo 2017)

9 https://ffmpeg.org/ (último acceso mayo 2017)

10 https://nodejs.org/es/ (último acceso mayo 2017)

ajustar la posición de las porciones de los vídeos en las pantallas en casos especiales (pantallas periféricas del videowall o pantallas en distribución disforme); y *tiempo* contiene el instante en que ha de empezar a reproducirse el video. Este parámetro se ha incluido para conseguir que todas las pantallas inicien la reproducción en el mismo instante de tiempo NTP (sincronización inicial de la reproducción). El parámetro *master* será de tipo booleano e indicará a cada RPi si tiene rol maestro (master) o esclavo. Como ejemplo, en el videowall uniforme de la Fig. 4, el mensaje que recibiría la pantalla superior derecha (IP 192.168.1.3), si en todo el videowall (una única sección de 2x3 pantallas) se va a reproducir un vídeo de resolución 1920x1080, sería:

VIDEO%ruta%1440%0%1920%540%L%tiempo%false

En el segundo mensaje, relativo a la Acción a realizar, el parámetro *acción* puede contener los valores "*play*", "*pause*", "*mute*" o "*unmute*", para iniciar o pausar la reproducción, bloquear o desbloquear el audio, respectivamente; mientras que el parámetro *tiempo* contiene el instante NTP en el que ha de ejecutarse exactamente dicha acción.

En el tercer mensaje, relativo al contenido de imágenes, el parámetro *ruta* contendrá la ruta de la imagen a mostrar por pantalla; y los parametros *X1*, *Y1*, *X2*, *Y2*, *posición* y *tiempo* tienen el mismo significado anteriormente descrito para el primer mensaje para contenido de video. En este caso, al tratarse de mostrar una imagen estática en una sección del videowall (o en todo) no será necesario implementar mecanismos de sincronización de procesos reproductores. Tan sólo será necesario sincronizar el instante en el que todas las pantallas de la sección en la que vaya a aparecer una porción de la imagen empiecen a mostrarla en el mismo instante NTP.

Al recibir el primer o tercer mensaje, si la RPi ya está reproduciendo contenido diferente, detendrá dicha reproducción y empezará a reproducir el nuevo contenido en el instante indicado dejando alguna parte de la pantalla en negro si fuera necesario para respetar la relación de aspecto del vídeo o la imagen original.

C.6. Servidor Web

Se ha instalado un servidor web basado en *node.js* en la RPi de control, a través del cual se podrá gestionar todo el sistema. Los usuarios pueden acceder, a través de un navegador, a dicho servidor para realizar la gestión/configuración del videowall. La aplicación/interface web desarrollada se presenta en la sección V.

IV. SOLUCIÓN DE SINCRONIZACIÓN

A. Sincronización inicial

Como se ha visto en los mensajes, cuando cada una de las RPi del videowall recibe una orden de la RPi de control, indicando la reproducción de contenido, esta incluye el instante NTP en el que ha de iniciarse dicha reproducción (para su cálculo se tiene en cuenta el retardo de las comunicaciones entre ellas). Por tanto, si todas las RPi tienen sus relojes sincronizados con el servidor NTP de la RPi de control, todas las pantallas de cada una de las secciones del videowall iniciarán su reproducción al mismo tiempo y de forma sincronizada.

B. Sincronización fina durante la reproducción

Los relojes de cada una de las RPi y las tasas de consumo de sus procesos de reproducción no tienen por qué ser exactamente iguales, sino que pueden sufrir desviaciones durante la sesión. Una vez iniciada la reproducción, dichas desviaciones pueden llevar, incluso en poco tiempo, a asincronías acumuladas que resulten en diferencias de un cierto número de fotogramas entre las distintas pantallas de la misma sección. Esto puede ser notable y molesto para los usuarios (especialmente en cambios de plano o de escena). Es por ello que se necesitan mecanismos de ajuste frecuente de los procesos de reproducción de las diferentes pantallas que conforman cada sección del videowall que permitan corregir dichas desviaciones.

En este caso, para sincronizar los puntos de reproducción de diferentes instancias del reproductor *OMXPlayer* se ha optado por utilizar la misma referencia de reloj global (proporcionada por NTP) para sincronizar el reloj local de cada una de las instancias del reproductor, así como utilizar el módulo *pyOmxSync*[11], desarrollado en Python, realizándole algunas modificaciones. Dicho módulo permite la sincronización de los procesos de reproducción de diferentes instancias del reproductor *OMXPlayer* conectadas a través de redes IP, siguiendo un esquema de control maestro/esclavo. La RPi de una de las pantallas de cada sección del videowall se comporta como reproductor maestro (de referencia), en cuanto al proceso de sincronización, del resto de procesos de las RPis de las otras pantallas pertenecientes a esa misma sección. Aunque se puede cambiar en la configuración, por defecto, sea cual sea la distribución del videowall (unifrome o disforme), se escogerá como dispositivo maestro a la RPi con menor dirección IP de cada sección.

La RPi con rol de maestro de cada sección del videowall, envía, a través del servidor Websocket a todas las demás RPi de su misma sección, cada cierto tiempo configurable (por ejemplo, un segundo), su punto de reproducción del vídeo (posición respecto al principio del vídeo) y el instante global de tiempo NTP (timestamp) correspondiente a dicho punto de reproducción. Las demás RPis, configuradas como dispositivos esclavos de esa misma sección y que, por tanto, están reproduciendo el mismo vídeo, cuando reciben dicha información la utilizan para comparar su propio punto de reproducción con el del dispositivo maestro. Si la diferencia entre

[11] https://github.com/markkorput/pyOmxSync (último acceso mayo 2017)

ambos supera un cierto margen o umbral de asincronía permitido (también configurable), corregirá su proceso de reproducción mediante ajustes, bien agresivos (saltos y pausas) o bien adaptativos o suaves (*Adaptive Media Playout* o AMP, [7]). Tras cada ajuste, y una vez sincronizado el proceso de reproducción, se esperarán un *tiempo de guarda* (configurable) para volver a sincronizar y dejar que, mientras, se estabilice dicho proceso. Durante el mismo hará caso omiso a los mensajes recibidos del maestro, en caso que los reciba.

Se ha comprobado en el laboratorio que con este método, en media, se puede mantener las asincronías por debajo de 60 ms. En https://goo.gl/PiFn5E se encuentra disponible un video mostrando la funcionalidad del sistema así como el grado de sincronización conseguida.

V. Aplicación de Gestión del Videowall

En esta sección se presenta la aplicación web realizada para la gestión y configuración del videowall. Para este propósito se ha instalado un servidor web mediante Node.js en la RPi de control. La gestión del videowall se realiza conectándose a este servidor web mediante un navegador (vía *http://IP_Rpi_de_control*).

La aplicación de configuración tiene varias pantallas. En primer lugar, se muestra la pantalla de entrada al servidor, la página de inicio de sesión (Fig. 6). Se ha implementado un control de usuarios con diferentes roles para que cada uno de ellos tenga acceso a unas carpetas y archivos en concreto de la unidad compartida y sólo pueda actuar sobre ellos (gestionando el contenido a ser seleccionable por dicho usuario para ser reproducido en el videowall). También se ha dado de alta un usuario *Admin* con rol de administrador que podrá, además de administrar el videowall, añadir y eliminar usuarios.

Fig. 6. Página de inicio de sesión

Una vez introducidas las credenciales, se pasa a la página mostrada en la Fig. 7, en la que aparecerán diferentes opciones de la aplicación web, dependiendo del rol de usuario que haya iniciado la sesión. Incluye las opciones de configuración general del videowall, gestión de usuarios y parámetros de red (todas ellas accesibles para el usuario *Admin*), y la de selección de contenido (accesible para todos los usuarios). La selección de cada una de las opciones lleva a una determinada página web para realizar la operación correspondiente.

Fig. 7. Página con todas las opciones de configuración (*Admin*)

La Fig. 8 muestran las ventanas de configuración general del videowall (uniforme a la izquierda y disforme a la derecha) a la que sólo pueden acceder los usuarios con rol de administrador (*Admin*). Se permite seleccionar si se trata de un videowall uniforme o disforme, así como configurar los siguientes parámetros de sincronización: el intervalo de envío de información de sincronización por parte de los dispositivos maestros, el valor de la asinconía máxima permitida entre la reproducción de las diferentes pantallas del videowall, el intervalo de guarda y el tipo de ajuste de sincronización (agresiva o AMP).

Fig. 8. Página de configuración general del videowall

La parte de la izquierda de la Fig.8 muestra la página de configuración del videowall en caso de tener una distribución de pantallas uniforme (matriz NxM). Permite seleccionar el número de pantallas en horizontal y el número de pantallas en vertical. La parte de la derecha de la Fig. 8 muestra la página de configuración del videowall en caso de tener una distribución de pantallas disforme. Se permite añadir pantallas al videowall, seleccionando su tamaño y número. Se crean en un panel y se permite arrastrarlas con el ratón, así como proporcionarles un ángulo de rotación para colocarlas, tal y como estarán las pantallas físicamente en el videowall real. Las pantallas se pueden seleccionar de forma independiente y cambiar sus propiedades (tamaño, posición y rotación).

La Fig. 9 muestra la página de gestión de usuarios, en la que se pueden crear, editar y eliminar usuarios, además de configurar los permisos de acceso a las carpetas correspondientes. A esta página sólo tendrán acceso los usuarios con rol de administrador (*Admin*). Al pulsar el botón de 'Añadir Nuevo Usuario', aparecerá la ventana de la parte inferior de la figura.

La Fig. 10 muestra la Gestión de la información de red de los dispositivos del videowall (sólo usuario *Admin*). Al pulsar el botón de 'Añadir Pantalla', aparecerá la ventana de la parte inferior de la figura.

La Fig. 11 muestra la página con la distribución del videowall (uniforme, en este caso) en la que se pueden seleccionar las pantallas que conforman cada sección del videowall, así como el contenido a mostrar en cada una de ellas. En caso de tratarse de un videowall uniforme, se deben seleccionar la pantalla de la esquina superior izquierda y la pantalla de la esquina inferior derecha de cada sección, e inmediatamente aparece una ventana para

seleccionar el contenido que se quiere visualizar en dicha sección. Si el contenido que se pretende enviar contiene audio, se preguntará si se quiere reproducir con audio o sin él (opción *muted*). En el caso de un videowall disforme se deben seleccionar, una a una, las pantallas en las que se quiere mostrar el contenido y pulsar sobre el botón 'Seleccionar Vídeos/Imágenes' y aparecerá la ventana ya comentada para seleccionar el contenido a visualizar. En esta página también aparecen botones para reproducir/reanudar y pausar la reproducción, así como tres opciones a destacar: '*Programación*', '*Seleccionar Audio*', '*Cargar Configuración*' y '*Reset Videowall*'.

Fig. 9. Página de gestión de usuarios

Fig. 10. Gestión de parámetros de red.

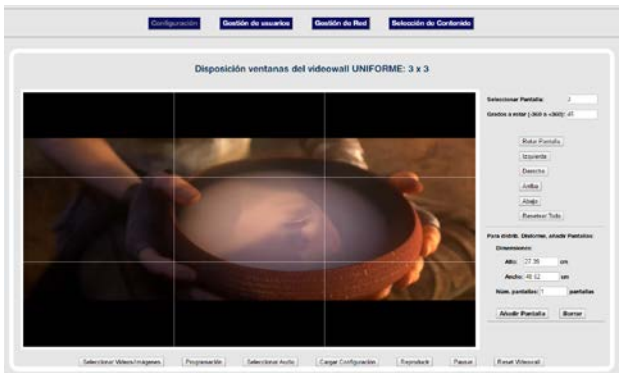


Fig. 10. Ventana de selección del contenido y programación

'*Programación*' permite guardar secuencias de reproducción programadas para cada sección del videowall. '*Seleccionar Audio*' permite activar la reproducción de los canales asociados a cada contenido siendo reproducido en alguna sección del videowall. Aparece la lista con los vídeos siendo reproducidos y, mediante controles de tipo *checkbox*, se puede seleccionar (o deseleccionar) cualquiera para que su audio empieze a sonar (o no) por los altavoces del videowall. '*Cargar Configuración*' envía la configuración a la RPi de control que será la encargada de enviar los mensajes pertinentes a las RPis de cada pantalla. '*Reset Videowall*' detiene la reproducción de las pantallas del videowall y borra las asignaciones de contenido a cada una de las secciones del mismo.

VI. Conclusiones y Trabajo Futuro

En este artículo se ha presentado una propuesta HW y SW para diseño y configuración de sistemas videowalls, tanto uniformes como disformes, muy completo y flexible, controlable dinámicamente vía web y, lo más importante, utilizando dispositivos de muy bajo coste (*cost-effective*). Tras analizar las diferentes opciones existentes para el desarrollo de videowalls, se ha definido la arquitectura y componentes apropiados, se ha implementado un prototipo siguiendo dicha propuesta.

Se han analizado distintos trabajos sobre videowalls basados en RPi y se han mejorado, añadiendo más funcionalidades, utilizando reproductores independientes en cada pantalla, y dispositivos de bajo coste. Además, se ha incluido una gestión integral del videowall mediante una aplicación web y, por tanto, multiplataforma.

Como trabajo futuro se pretende realizar evaluaciones objetivas y subjetivas para analizar su idoneidad, aumentar las funcionalidades (p.ej., la visualización de contenido en directo a través de streaming IP basado en RTP [8] o MPEG-DASH). También se pretende añadir multiplexores analógicos al circuito de audio, para tener diferentes equipos de sonido y enviar a cada uno el audio o audios que se deseen, controlables mediante salidas GPIO (*General Purpose Input/Output*) de las RPi.

Agradecimientos

Los autores quieren agradecer a Mª José Canet y José Vicente Llario, profesores del Dept. de ingenieria electrónica de la UPV, su ayuda en la realización del circuito electrónico para controlar las salidas de audio.

Referencias

[1] A. Jiménez, "Videowall disforme sobre redes IP", Trabajo Final de Grado, Universitat Politècnica de Catalunya, julio 2016.
[2] *PiWall:* http://www.piwall.co.uk (ult. acceso mayo 2017).
[3] *Video Wall using the Raspberry Pi* (ult. acceso mayo 2017). https://www.yodeck.com/news/video-wall-using-raspberry-pi/
[4] GStreamer: open source multimedia framework. https://GStreamer.freedesktop.org/ (último acceso mayo 2017)
[5] Proyecto OMXPlayer, http://omxplayer.sconde.net/ (último acceso mayo 2017)
[6] D. Mills, J. Martin, J. Burbank, and W. Kasch, "NetworkTime Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.
[7] M. Montagud, F. Boronat, B. Roig, A. Sapena, "How to perform AMP? Cubic adjustments for improving the QoE", Computer Communications, Volume 103, 1 May 2017, Pages 61-73, ISSN 0140-3664.
[8] H. Schulzrinne, et al, "RTP: A Transport Protocol for Real-Time Applications", IETF Standard, RFC 3550, July 2003.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6485

Experiencia de Implantación de Estrategias de Autoevaluación y Coevaluación en el Grado de Ingeniería Telemática

Jaume Ramis Bibiloni, M. Magdalena Payeras Capellà, Loren Carrasco Martorell
Departament de Ciències Matemàtiques i Informàtica
Universitat de les Illes Balears
Carretera de Valldemossa, Km. 7,5 07122 Palma
jaume.ramis, mpayeras, loren.carrasco@uib.es

***Resumen*—En este artículo se presenta una propuesta metodológica aplicada a la asignatura *Planificación de redes* del tercer curso del Grado de Ingeniería Telemática de la Universidad de las Islas Baleares. Se fundamenta en la utilización de técnicas de autoevaluación y coevaluación (evaluación entre iguales) con el objetivo de mantener a los alumnos continuamente informados sobre su proceso de aprendizaje, estimulando la cooperación entre ellos asi como su interacción con el profesor. Esta metodología consigue además mejorar su capacidad para formular críticas constructivas respecto del trabajo propio y del de los compañeros. Los resultados obtenidos ponen de manifiesto la creciente calidad de la tarea de evaluación realizada por los alumnos y evidencian la buena aceptación de esta propuesta metodológica entre ellos.**

***Palabras Clave*—Evaluación continuada, Autoevaluación, Coevaluación, Cooperación, Rúbrica**

I. INTRODUCCIÓN

La evaluación continuada es uno de los pilares en los que se sustenta una docencia de calidad. Se debe diferenciar entre *información* continuada y *calificación* continuada [1]. La primera es imprescindible para una docencia de calidad. En cambio, la segunda no siempre es recomendable, ya que calificar al alumno antes de la última fase del curso, cuando todavía no ha asimilado los contenidos de la asignatura, puede sesgar su calificación final. Como herramientas para lograr informar de forma continua y con prontitud a los alumnos podemos utilizar la autoevaluación y la evaluación entre iguales o coevaluación. Para lograr que esta colaboración de los alumnos en la evaluación continuada goce de buena salud es imprescindible instruirlos en el procedimiento a seguir para dicha tarea. La utilización de rúbricas constituye una buena estrategia para facilitar el proceso de evaluación. La rúbrica es un documento que delimita lo que es evaluable y lo concreta en niveles de eficiencia.

Este artículo recoge las experiencias y los resultados obtenidos en la aplicación de estrategias de autoevaluación y coevaluación en la asignatura *'Planificación de redes'* del tercer curso del Grado en Ingeniería Telemática. La implantación ha sido gradual, incorporando en primer lugar la autoevaluación y, posteriormente, la coevaluación a título ilustrativo, para finalmente incorporarla como un elemento más en el proceso de evaluación. Ello responde a la utilización de diversos métodos de evaluación que pretenden no únicamente calificar al alumno sino proporcionarle herramientas para la mejora en el desarrollo de competencias.

Este proceso ha permitido, por una parte, detectar un cambio en la actitud y en el comportamiento del alumnado frente al trabajo evaluado y, por otra, observar como las notas de la coevaluación se han ido aproximando, paulatinamente, a las notas del profesor, evidenciando que los alumnos han ido aprendiendo a utilizar las rúbricas y por tanto ha mejorado su capacidad de detección de los errores propios y ajenos. De esta manera se ha fomentado el hábito de reflexión sobre el trabajo realizado, elemento fundamental para el aprendizaje autónomo. Así mismo, la evaluación entre iguales fomenta la capacidad de emitir críticas constructivas y juzgar el trabajo de los compañeros, habilidad fundamental para el ejercicio profesional.

A continuación se describirán los métodos de evaluación utilizados antes de la adopción de la propuesta detallada en el presente trabajo. Seguidamente se explicará en qué consiste nuestra propuesta metodológica; en concreto, se presentará la rúbrica de evaluación adoptada y el procedimiento que se ha seguido para la familiarización de los alumnos con su utilización. Se mostrarán después los resultados obtenidos, que serán analizados con el objetivo de valorar la efectividad de la metodología planteada. Para finalizar se concluirá el artículo con una reflexión sobre su aplicación práctica y las posibilidades de mejora.

II. Metodología de evaluación previa

Los contenidos de la asignatura *Planificación de Redes* del tercer curso del Grado en Ingeniería Telemática consisten básicamente en el modelado formal de protocolos mediante la utilización de cadenas de Markov. La metodología didáctica utilizada se basa fundamentalmente en la realización de explicaciones teóricas por parte del profesor y en la resolución de problemas y realización de prácticas por parte de los alumnos con el objetivo de consolidar los conceptos y conocimientos introducidos por el profesor. Los alumnos trabajan en grupos reducidos, generalmente de dos miembros, y deben entregar sus actividades unos días después de finalizar cada tema, según marca el calendario indicado por el profesor, a través de *Campus Extens* (plataforma Moodle). La resolución de ejercicios y prácticas se lleva a cabo mediante la realización de pequeños programas utilizando el Matlab [2]. Además se realizan dos controles a lo largo del cuatrimestre, siendo necesario superar ambas pruebas (nota superior a cinco) para aprobar la asignatura.

Uno de *los siete principios de la docencia de calidad* [3] consiste en *proporcionar retroalimentación a tiempo*. Es por ello que resulta imprescindible disponer de las herramientas necesarias para mantener al alumno informado sobre su progreso a lo largo de todo el curso, sin que tenga que esperar a los exámenes para ello. Así pues, el alumno debe disponer de esta retroalimentación de forma rápida para poder emprender las acciones que considere oportunas.

Con este objetivo, en el primer año de impartición de la asignatura *Planificación de Redes*, ésta se diseñó de manera que tras cada entrega, consistente en el conjunto de actividades correspondientes a un tema, el profesor llevaba a cabo su corrección y posteriormente informaba a los alumnos de la calificación obtenida por su trabajo. Dada la naturaleza de los trabajos, este esquema resultó del todo ineficiente e improducivo: por una parte implicó una elevada carga de trabajo para el profesor, dada la gran cantidad de ejercicios a resolver a pesar del no muy elevado número de alumnos y, por otro lado, tan sólo consitía en una *calificación* continuada y estaba lejos de ser una *información* continuada.

A partir de la segunda edición del curso, se decidió mantener la realización de las mismas entregas de actividades por parte del alumnado, pero se sustituyó la corrección de todos los ejercicios por la corrección de una muestra de ejercicios seleccionados por el profesor. Se trataba de una muestra suficientemente respresentativa de la tipología de actividades de las entregas realizadas por los alumnos. Entonces el profesor llevaba a cabo la corrección de esta selección de ejercicios y asignaba la calificación correspondiente. Con esta estrategia se redujo considerablemente la carga de trabajo del profesor sin que ello respresentara una reducción significativa de la retroalimentación que recibían los alumnos. Pero los alumnos seguían sin disponer de la información continuada sobre su progreso.

Este esquema se mantuvo hasta el curso pasado, cuando se decidió modificar la metodología didáctica con el objetivo de proporcionar una retroalimentación de calidad a los alumnos, apostando por una estrategia que estimulase además la cooperación entre ellos y que favoreciese el contacto entre alumnos y profesor, elementos que también se encuentran entre *los siete principios de la docencia de calidad* [3]. En la siguiente sección se detalla en qué consite esta propuesta.

III. Metodología propuesta

La metodología didáctica propuesta incorpora la autoevaluación y la coevaluación (o evaluación entre compañeros) como herramientas que facilitan la retroalimentación a tiempo y de forma continuada a los alumnos sobre el progreso de su trabajo a lo largo de todo el curso. Para ello es fundamental la utilización de una rúbrica para la evaluación de los programas realizados por los alumnos utilizando la herramienta Matlab [2]. Resulta imprescindible que el evaluador, ya sea el propio autor del trabajo, un compañero de clase o bien el profesor, *haga suya* la rúbrica. Sólo así se logrará el éxito de la presente propuesta.

A. Rúbrica para la evaluación

La rúbrica utilizada se muestra en la Tabla I. Se halla estructurada en tres columnas, la primera de las cuales especifica cada criterio de evaluación considerado, mientras que la segunda corresponde a una descripción detallada de los aspectos que se deben considerar para el criterio en cuestión y, finalmente, la tercera especifica cómo asignar la puntuación.

De acuerdo con los criterios considerados, se valora la estructura y claridad del código así como su corrección. En cuanto a los resultados se valora tanto la claridad en la presentación de los resultados como su corrección, así como la interpretación de los mismos.

La rúbrica no tan sólo es un herramienta para la evaluación sino también para el aprendizaje, puesto que detalla las características que debe satisfacer un trabajo para ser evaluado de forma positiva. Su validez depende de la objetividad en su aplicación, permitiendo a distintos evaluadores asignar puntuaciones muy similares a un mismo trabajo. La existencia de diferencias significativas en las evaluaciones puede denotar la falta de objetividad en la rúbrica y/o el poco dominio del evaluador en su utilización.

B. Procedimiento

Tanto la planificación de la asignatura como su evaluación se diseñaron dando máxima prioridad a la familiarización del alumnado con la utilización de la rúbrica de evaluación. Así pues, la asignatura se divide en dos grandes bloques: el primero de *conceptos básicos* y el segundo de *modelado formal de protocolos*. Como ya se ha comentado en la sección II, a lo largo de todo el curso los alumnos trabajan en grupos de dos alumnos y van realizando las entregas de las actividades de cada tema. Además, tras finalizar cada bloque, los alumnos

Tabla I
RÚBRICA ORIGINAL UTILIZADA EN LA ASIGNATURA "PLANIFICACIÓN DE REDES"

Criterio	**Descripción**	**Puntuación**
Estructura y claridad del código (2 puntos)	El código esta organizado y estructurado correctamente El código está identado correctamente. Las instrucciones incluyen descripciones explicativas y aclaradoras.	2 puntos si se cumplen todos los criterios 1 punto si falta un criterio por cumplir. 0 puntos si faltan dos o más criterios por cumplir.
Corrección del código (3 puntos)	No hay errores en el código	3 puntos si el número de errores es 0 2 puntos si el número de errores es 1 1 puntos si el número de errores es 2 0 puntos si el número de errores es 3 o más
Resultado de la ejecución del código (2 puntos)	No hay duda de cual es el resultado obtenido con la ejecución. Todos los gráficos incluyen la descripción de los ejes y la leyenda En caso de que se requieran diversos gráficos, utilizar subplots	2 puntos si se cumplen todos los criterios 1 punto si falta un criterio por cumplir 0 puntos si faltan dos o más criterios por cumplir
Corrección de los resultados (2 puntos)	los resultados (numéricos y gráficos) son correctos.	2 puntos si se cumple en el 100% de los resultados. 1 punto si se cumple en más del 75%. 0 puntos si se cumple en menos del 75%.
Comentarios de los resultados (1 punto)	Los resultados están completamente y correctamente comentados	2 puntos si se cumple en el 100% de los resultados. 1 punto si se cumple en más del 75%. 0 puntos si se cumple en menos del 75%.

realizan, de forma individual, sendas pruebas de control. A continuación se detalla la organización del procedimiento.

1) Aprendizaje del uso de la rúbrica y autoevaluación de las entregas de actividades: Con el objetivo de familiarizar a los alumnos con la rúbrica, a lo largo de la primera parte del curso, correspondiente al bloque de *conceptos básicos*, tras cada entrega de actividades se destina la siguiente hora de clase a su corrección. El profesor selecciona un alumno para cada ejercicio, quien explica su resolución al grupo con la ayuda del proyector (cada alumno realizará esta intervención en público como mínimo dos veces a lo largo del curso). Seguidamente el profesor realiza los comentarios oportunos y los compañeros pueden consultarle cualquier aspecto que sea de su interés. A continuación se procede a la calificación del ejercicio utilizando la rúbrica de evaluación: alumnos y profesor consensuan las puntuaciones de cada apartado, aclarando cualquier duda sobre el procedimiento. Las calificaciones así obtenidas constituirán la *nota de seguimiento* del curso, con un peso del 10% sobre la nota final. Para acabar el proceso, cada alumno se autoevalúa su trabajo. Se establece como requisito indispensable para la validez de las evaluaciones que se justifiquen convenientemente las calificaciones asignadas en cada apartado de la rúbrica.

Tras finalizar la sesión de revisión de la entrega de actividades en cuestión, los alumnos, de forma individual, suben a 'Campus Extens' el resultado de la autoevaluación de su entrega. Con el objetivo de alentar a los alumnos a que realizen esta tarea de autoevaluación, un 10% de la nota final de la asignatura se corresponde a la *nota por las entregas realizadas*. Ésta se obtiene en función del número de entregas de actividades con sus respectivas autoevaluaciones que hayan llevado a cabo los alumnos y no de las calificaciones obtenidas en ellas. Con ello se pretende potenciar la aplicación objetiva de los criterios de calificación especificados en la rúbrica. Además, los alumnos son informados de que estas autoevaluaciones son revisadas por el profesor, quien, si lo cree necesario, puede proponer a aquellos alumnos que considere recomendable, la realización de problemas complementarios y tutorías para mejorar sus resultados.

El número total de entregas a realizar a lo largo del curso es 10. La distribución de calificaciones es la siguiente:

- El 100% de entregas + autoevaluación: 10%
- Entre 100% y el 90% de entregas + autoevaluación: 7.5%
- Entre 90% y el 80% de entregas + autoevaluación: 5%
- Menos del 80% de entregas + autoevaluación: 0%

2) Prueba ilustrativa de autoevaluación y coevaluación del primer control: Tras este proceso, que comprende unas siete semanas, se realiza la primera prueba de control, cuyo peso sobre la nota final es del 40%. Los alumnos suben sus resoluciones a 'Campus Extens', donde se halla la tarea *Control1*, que consiste en un *taller* de la plataforma Moodle [4]. Esta herramienta permite configurar actividades para la corrección entre iguales, obteniéndose una *calificación de la actividad* realizada por el alumno y una *calificación de la tarea de evaluación* de sus compañeros. Además permite incorporar cualquier tipo de rúbrica para el proceso de evaluación. La calificación del taller está configurada para que la *nota del control* se obtenga de la siguiente manera:

- la *calificación de la actividad* representa el 80%, correspondiente a:
  - calificación asignada por el profesor: 75%
  - calificación media asignada por los compañeros: 25%
- la *calificación media de la tarea de evaluación* de sus compañeros: 20%

Con el objetivo de primar la evaluación realizada por el profesor, el peso de su calificación de la actividad es tres veces mayor que el peso correspondiente a la calificación de la actividad por parte de los compañeros. Así, la calificación asignada por el profesor corresponde al 60% de la nota del control, la calificación recibida por parte de sus compañeros representa el 20% y su tarea de evaluación de los compañeros corresponde al 20% restante. Como se observa, estos dos últimos porcentajes son iguales, lo cual significa que evaluar correctamente a sus compañeros representa la misma recompensa que sobrepuntuar a sus compañeros. Con ello se incita a los alumnos a corregir de la forma más objetiva posible.

La clase posterior a la realización del control se destina a su evaluación. Ésta comienza con la resolución de

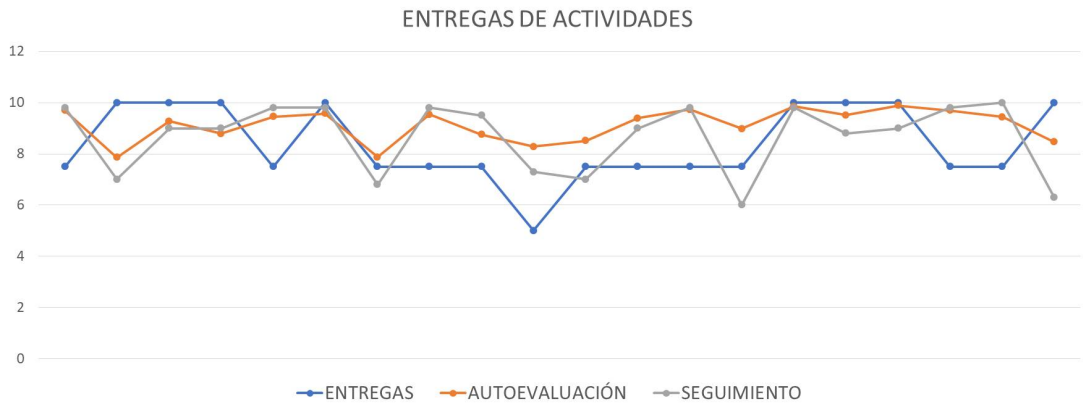


Fig. 1. Resultados del seguimiento de las entregas de actividades de los alumnos

los ejercicios del control por parte del profesor ante el grupo de clase. Seguidamente cada alumno realiza su autoevaluación y la evaluación de tres compañeros. Para ello utilizan la rúbrica incorporada en el taller de Moodle, siendo imprescindible que los alumnos justifiquen sus calificaciones mediante los comentarios oportunos.

Como primer paso en la incoroporación de la autoevaluación y la coevaluación, se decidió llevar a cabo este proceso a nivel tan sólo ilustrativo, haciendo que la calificación de esta primera prueba de control se correspondiese exclusivamente con la calificación del profesor.

3) Autoevaluación de las entregas de actividades: A lo largo de la segunda parte del curso, correspondiente al bloque de *modelado formal de protocolos*, los alumnos siguen realizando sus entregas de actividades y siguen llevando a cabo su resolución en clase. En este punto del curso los alumnos están altamente familiarizados con el uso de la rúbrica, de manera que durante la sesión de resolución de los ejercicios en clase, las explicaciones de los alumnos seleccionados por el profesor se centran básicamente en un comentario general del programa y de los resultados obtenidos. El profesor realiza los comentarios oportunos y, a continuación, cada alumno se autoevalúa su trabajo - no existe ya la necesidad de aplicar la rúbrica de forma conjunta y consensuada.

4) Autoevaluación y coevaluación del segundo control: La última semana se destina a la realización de la segunda prueba de control y a su evaluación, siguiendo el mismo proceso descrito para la primera prueba. Ahora ya sí que la calificación obtenida corresponde al resultado del procedimiento antes detallado.

IV. Resultados

El número de alumnos matriculados el curso 2016/2017 ha sido 26, de los cuales 20 alumnos han asistido a más del 80% de clases. Los 6 restantes son alumnos a tiempo parcial y no se han adherido al itinerario presencial, por lo que no han seguido la propuesta metodológica objeto del presente trabajo.

En esta sección analizaremos los resultados obtenidos, que se han organizado de la siguiente manera:

- familiarización con la rúbrica e incorporación de la autoevaluación,
- autoevaluación y coevaluación de las pruebas de control,
- resultados finales.

A. Autoevaluación de las entregas de actividades

Como ya se ha indicado en la sección III-B, la calificación obtenida por los alumnos dependerà del número de entregas de actividades y autoevaluaciones, y no de las calificaciones obtenidas en ellas. La asignación de calificaciones es muy estricta, propiciando que los alumnos lleven a cabo la mayoría de sus *entregas de actividades*, lo que se refleja en los resultados obtenidos, que se muestran en la Figura 1. El eje horizontal de todas las figuras se corresponde con cada uno de los 20 alumnos que han seguido la metodología propuesta en este trabajo. Exceptuando un alumno, que realizó el 85% de las entregas, el resto supera el 90% y casi la mitad llega al 100%. Se representan también las calificaciones medias obtenidas en las *autoevaluaciones* de todas las entregas de actividades a lo largo del curso, así como la nota de *seguimiento* asignada por el profesor según del procedimiento detallado en la subsección III-B1. En general, se aprecian las mismas oscilaciones en estas dos últimas curvas, evidenciando la correlación existente entre ambas. La discrepancia promedio entre ellas es inferior a 0.5 puntos.

B. Autoevaluación y coevaluación de las pruebas de control

En primer lugar, cabe destacar que la implementación del proceso mediante la herramienta *'taller'* de la plataforma Moodle, ha facilitado la utilización de la rúbrica en la autoevaluación y coevaluación de las pruebas de control.

Debe tenerse en cuenta que la calificación de la tarea de evaluación no corresponde a la simple comparación de la calificación otorgada por el alumno coevaluador en relación a la asignada por parte del profesor. De hecho la comparación se realiza con respecto a la media ponderada de todas las calificaciones tal como se detalló en la subsección III-B2, donde se explicó que se asignan

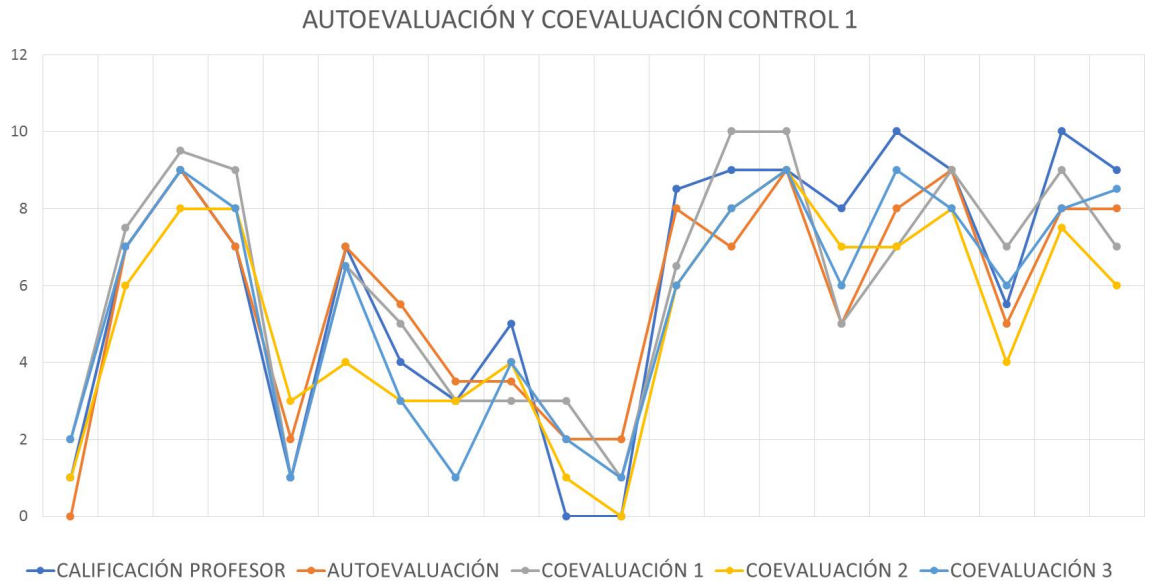


Fig. 2. Resultados de la evaluación de la primera prueba de control

pesos con el objetivo de primar la evaluación realizada por el profesor. Entonces, la tarea de evaluación se califica teniendo en cuenta las puntuaciones asignadas a cada criterio de evaluación presente en la rúbrica. Ello puede dar como resultado que la puntuación global del ejercicio según la evaluación del profesor coincida con la del alumno coevaluador, pero que sin embargo la calificación de la tarea de evaluación del alumno reciba una puntuación muy baja. Se trata por tanto de puntuar cada ítem de la rúbrica de forma correcta y no simplemente de asignar la nota de forma general al ejercicio.

En nuestro caso cada ejercicio ha sido evaluado por su autor, por el profesor y por tres compañeros. El análisis de los resultados representados en las figuras 2 y 3 muestran como, en general, convergen las calificaciones otorgadas por los cinco revisores. Puede observarse claramente que para el segundo control existe una mayor similitud entre todas ellas. Se aprecia que tan sólo existen muy pocos casos con discrepancias significativas (detectamos diferencias de hasta 3 puntos (sobre 10) para el primer control y de hasta 2 puntos en el segundo). Estos resultados ponen de manifiesto el progresivo dominio de los alumnos en la correcta utilización de la rúbrica. Cabe destacar además que el número de coevaluaciones realizadas en

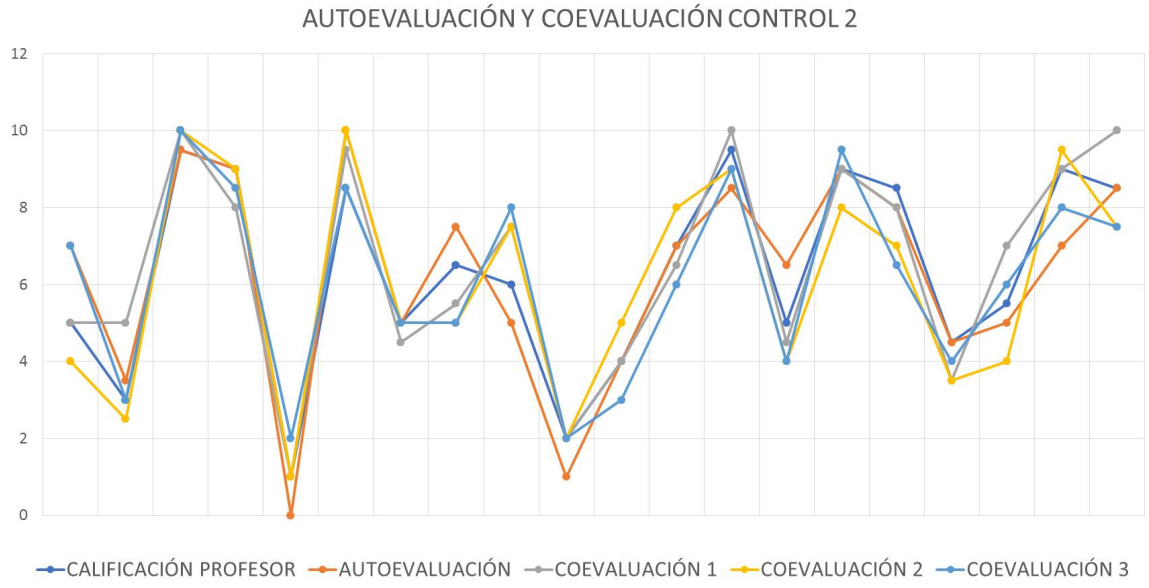


Fig. 3. Resultados de la evaluación de la segunda prueba de control

las pruebas de control debe ser suficientemente grande para garantizar la fiabilidad del proceso. Para los casos en que se producen discrepancias significativas entre las calificaciones asignadas existe la posibilidad de solicitar una segunda revisión más exhaustiva de los ejercicios al revisor en cuestión. Incluso podría descartarse dicha revisión divergente. No olvidemos que una tarea de evaluación incorrecta se traduce en una peor nota para el revisor.

Las figuras 4 y 5 representan las notas obtenidas en las dos pruebas de control, detallando la calificación de la *actividad* y la calificación de la *tarea de evaluación*. Además se incluyen las calificaciones asignadas por el profesor. En ambas figuras se aprecia un elevado grado de simililitud entre la calificación del profesor y la asignada a la actividad, básicamente coincidentes en la segunda prueba de control. En cuanto a la calificación de la tarea de evaluación, se aprecia una clara mejora en el segundo control en comparación con el primero. En el primer control el 75% de los alumnos realizaron esta tarea de forma satisfactoria (nota de la tarea de evaluación superior a 5), alcanzándose el 100% en el segundo control. Ello influye de forma positiva en la nota obtenida por los alumnos: si bien en el primer control la tarea de evaluación influyó negativamente en el 65% de los alumnos con una reducción promedio de -0.85 puntos en su nota del control, en el segundo control este porcentaje se reduce al 50% con una reducción promedio de tan solo -0.21 puntos. Los resultados ponen de manifiesto la fiabilidad del proceso, sin necesidad de descartar ninguna revisión. Ello es fruto del elevado número de revisiones de cada ejercicio.

C. Comentario de las calificaciones del curso

La figura 6 resume las calificaciones obtenidas por los alumnos. Recordemos que la nota final se obtiene a partir de la siguiente distribución:

- 10% nota por las entregas realizadas, de acuerdo a lo descrito en la subsección III-B1
- 10% nota de seguimiento, según se especifica en la subsección III-B1
- 40% prueba de control 1, tal como se detalla en la subsección III-B2
- 40% prueba de control 2

Todos los alumnos han recibido una valoración positiva del número de entregas realizadas y del seguimiento. En la segunda prueba de control se aprecia una mejora en las calificaciones en el 70% de los alumnos con respecto al primer control. El número de alumnos que superan positivamente la asignatura es del 80% en la convocatoria ordinaria, resultado altamente satisfactorio.

V. Análisis de la experiencia, propuestas de mejora y conclusiones

La experiencia nos muestra que un elemento clave para el éxito de la propuesta es el diseño de una rúbrica fácilmente utilizable por parte de alumnos y profesores, que defina los criterios de evaluación de forma clara y objetiva. Además debe evitarse cualquier tipo de solape entre los criterios a evaluar. Es también fundamental distribuir los pesos proporcionalmente a la importancia que se le otorga a cada criterio. El diseño propuesto inicialmente debe ser mejorado a partir de la experiencia de su utilización. Para ello puede contarse con la colaboración de los alumnos. En el caso que nos atañe se decidió modificar la rúbrica para su utilización en la siguiente edición de la asignatura. La rúbrica mejorada se muestra en la tabla II. Se diferencia de la versión original en dos aspectos:

- se iguala el peso de todos los criterios excepto el de *'corrección del código'* al que se le otorga mayor predominancia;
- se describe con mayor detalle el criterio *'corección del código'* con el fin de distinguir entre errores *leves* y errores *graves*, especificando a título ilustrativo algunos ejemplos de cada tipología.

Además se hizo evidente que resulta imprescindible elaborar los enunciados de los ejercicios teniendo en todo momento presente que deberán ser evaluados utilizando la rúbrica. Así pues, el diseño de los ejercicios seguirá una estructura que facilitará la identificación de cada uno de los criterios detallados en la rúbrica así como su evaluación siguiendo las pautas en ella detalladas. Esto resulta especialment relevante en los ejercicios que conforman las pruebas de control.

La aplicación de la metodología propuesta en el presente trabajo ha resultado una herramienta altamente eficaz para alcanzar el objetivo que nos habíamos propuesto como primer propósito: *proporcionar retroalimentación de calidad y a tiempo*. La autoevaluación de cada una de las entregas de actividades a lo largo del curso ha representado un cambio muy significativo en la dinámica con respecto a las ediciones anteriores. Los alumnos perciben la autoevaluación de forma positiva dado que la rúbrica representa una guía para el aprendizaje, puesto que detalla las características que debe satisfacer su trabajo para ser evaluado de forma positiva y, al mismo tiempo, les informa continuamente sobre la calidad y la evolución de su desempeño.

Los alumnos han *interiorizado* la rúbrica con gran facilidad, lo que ha sido un elemento clave para el éxito de esta propuesta. Gracias a ello la calidad de las autoevaluaciones y coevaluaciones ha sido en general elevada. Los principales beneficios obtenidos son [1], [5]:

- Los alumnos hacen suyos los criterios de evaluación y mejoran cada vez más sus respuestas, fomentando el hábito de reflexión sobre el trabajo realizado, elemento fundamental para el aprendizaje autónomo.
- En cuanto a las coevaluaciones, permiten a los alumnos ver soluciones alternativas a los ejercicios, contribuyendo también a la reflexión no sólo sobre el trabajo de los demás, sino incluso sobre el propio. El hecho de ser evaluados por sus compañeros les motiva para realizar un buen trabajo. De la misma manera, evaluar a sus compañeros fomenta su capacidad de ser crítico con el trabajo de los demás.

Otro beneficio de esta propuesta metodológica ha sido que ha representado un verdadero estímulo para la coop-

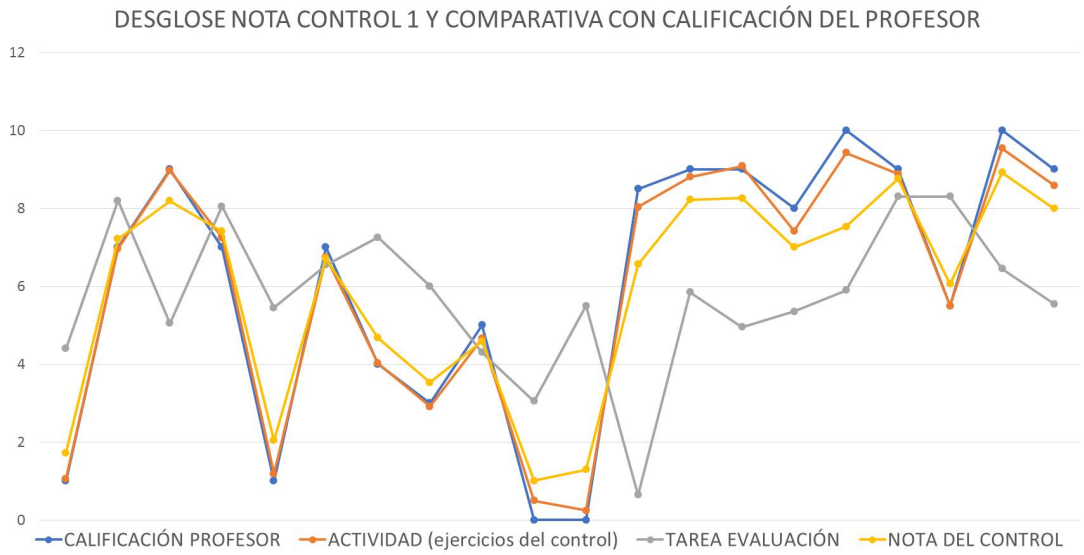


Fig. 4. Resultados de la primera prueba de control: desglose y comparativa con la calificación del profesor

Tabla II
RÚBRICA ADAPTADA UTILIZADA EN LA ASIGNATURA "PLANIFICACIÓN DE REDES"

Criterio	**Descripción**	**Puntuación**
Estructura y claridad del código (1,5 puntos)	El código esta organizado y estructurado correctamente El código está identado correctamente. Las instrucciones incluyen descripciones explicativas y aclaradoras.	1,5 puntos si se cumplen todos los criterios 0,75 puntos si falta un criterio por cumplir. 0 puntos si faltan dos o más criterios por cumplir.
Corrección del código (4 puntos)	Se distinguen dos tipos de errores. Errores Graves: de concepto. Errores Leves: bucles innecesarios, repeticiones, pequeños errores de comando,...	4 puntos si el número de errores es 0 Cada error grave resta dos puntos. Cada error leve resta un punto.
Resultado de la ejecución del código (1,5 puntos)	No hay duda de cual es el resultado obtenido con la ejecución. Todos los gráficos incluyen la descripción de los ejes y la leyenda En caso de que se requieran diversos gráficos, utilizar subplots	1,5 puntos si se cumplen todos los criterios 0,75 puntos si falta un criterio por cumplir 0 puntos si faltan dos o más criterios por cumplir
Corrección de los resultados (1,5 puntos)	los resultados (numéricos y gráficos) son correctos.	1,5 puntos si se cumple en el 100% de los resultados. 0,75 puntos si se cumple en más del 75%. 0 puntos si se cumple en menos del 75%.
Comentarios de los resultados (1,5 puntos)	Los resultados (numéricos y gráficos) estan completamente y correctamente comentados	1,5 puntos si se cumple en el 100% de los resultados. 0,75 punto si se cumple en más del 75%. 0 puntos si se cumple en menos del 75%.

eración entre alumnos. El trabajo en equipo ha resultado un elemento clave en el proceso de aprendizaje de la utilización de la rúbrica de evaluación. Además ha favorecido el contacto entre alumnos y profesor, puesto que éste ha sido quien continuamente ha aclarado todo tipo de dudas en cuanto a la aplicación de la rúbrica. Las consultas al

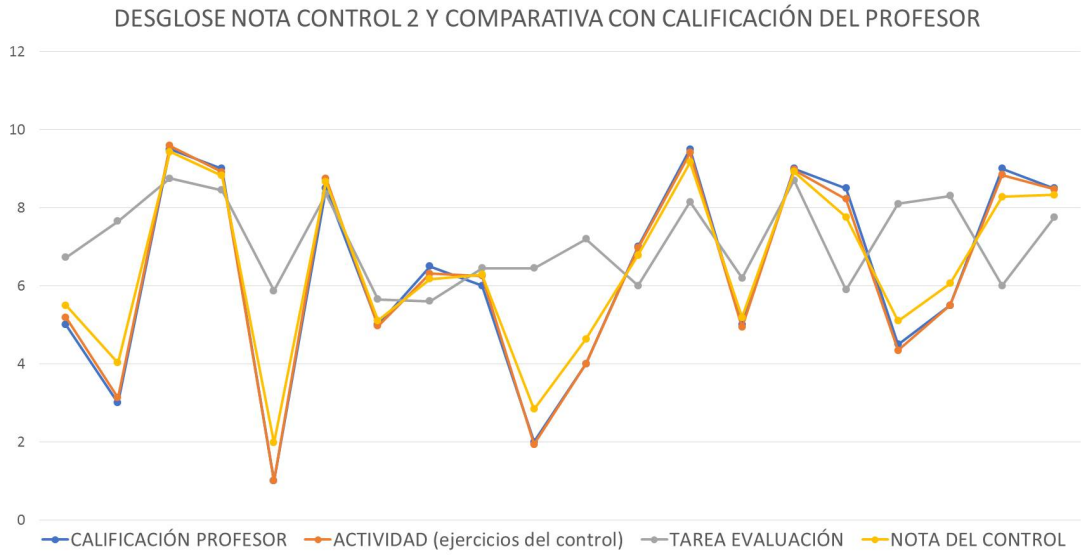


Fig. 5. Resultados de la segunda prueba de control: desglose y comparativa con la calificación del profesor

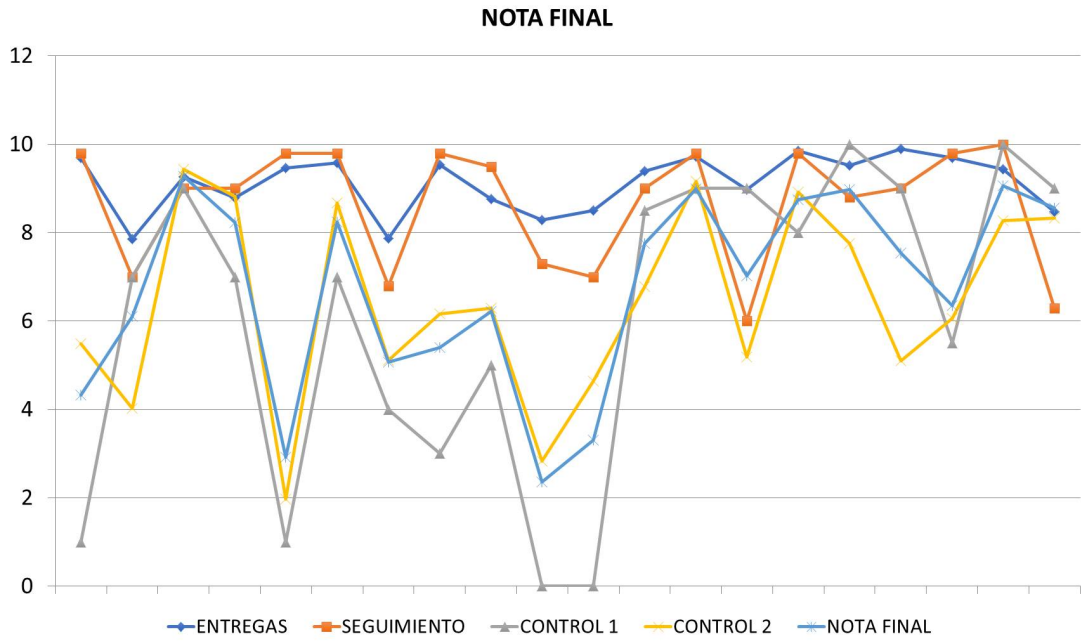


Fig. 6. Resultados de la calificación final del curso

profesor han proliferado de forma especialmente significativa durante las sesiones de autoevaluación y coevaluación de las pruebas de control, dado el interés de los alumnos en aplicar correctamente la rúbrica para obtener así una buena calificación de su tarea de evaluación. En definitiva, la participación del alumnado en el desarrollo de todo el proceso de aprendizaje, tanto autónomo como en clase, ha sido mucho más activa.

Como evidencia objetiva del grado de aceptación de esta propuesta metodológica, la tabla III muestra la evolución del resultado correspondiente al ítem *'nivel de satisfacción'* del *informe de opinión del alumnado de grado sobre la tarea docente del profesorado*, aplicado por el *Servei d'Estadística i Qualitat Universitària* [6]. Puede observarse la significativa mejora en el nivel de satisfacción de los alumnos correspondiente a la última edición del curso con respecto a la anterior. Ello evidencia la buena acogida de esta propuesta metodológica entre el alumnado.

Tabla III
NIVEL DE SATISFACCIÓN DEL ALUMNADO DE LA ASIGNATURA 'PLANIFICACIÓN DE REDES'

2012/2013	2013/2014	2014/2015	2015/2016	2016/2017
7,25	5,91	7,17	6,7	8,32

Es obvio el ahorro en el trabajo de revisión de las entregas de actividades por parte del profesor y la mejora en la calidad de la retroalimentación, a tiempo y de forma continuada, que reciben los alumnos sobre el progreso de su trabajo a lo largo de todo el curso. De todas formas, ha sido necesario invertir esfuerzos por ambas partes, alumnos y profesor, para la aplicación en el aula de esta metodología didáctica. Han sido necesarias 8 horas para llevar a cabo las sesiones de revisión de las actividades en clase, lo que ha significado un incremento de las horas de trabajo autónomo de los alumnos, puesto que los ejercicios no se han reducido con respecto a las ediciones anteriores de la asignatura. Por otra parte, el profesor ha visto incrementada significativamente la carga horaria asociada a la impartición de la asignatura debido al diseño de la metodología en sí, la configuración de la rúbrica, la implementación de los talleres de Moodle para las pruebas de control y el seguimiento y control de todo el procedimiento. A pesar de ello, la buena aceptación por parte del alumnado, los satisfactorios resultados académicos obtenidos y las expectativas de seguir aplicando esta metodología a lo largo de los próximos cursos, compensan con creces el esfuerzo realizado.

Para finalizar, cabe indicar que se pueden utilizar procedimientos similares a la metodología aquí propuesta a otra tipología de trabajos, como pueden ser colecciones de problemas, informes o trabajos escritos o también presentaciones orales. En este sentido los autores están perfilando el diseño de estrategias de coevaluación para presentaciones orales de trabajos en equipo.

REFERENCIAS

[1] Miguel Valero-García, Luis M. Díaz de Cerio, Autoevaluación y coevaluación: estrategias para facilitar la evaluación continuada, Actas del Simposio Nacional de Docencia en la Informática, SINDI2005 (AENUI), pp.25-32 ISBN: 84-9732-443-9, 2005
[2] https://www.mathworks.com/products/matlab.html
[3] A. W. Chickering y Z. F. Gamson, Seven Principles for Good Practice in Undergraduate Education, March 1987, American Association of Higher Education and Accreditation (AAHEA) Bulletin, Vol. 39(7), pages 3-7
[4] https://moodle.org
[5] A.W. Bangert, Peer Assessment: A Win-Win Instructional Strategy for Both Students and Teachers, J. Cooperation & Collaboration in College Teaching, Vol. 10, No. 2, pp. 77-84, January 2001.
[6] http://sequa.uib.cat/Avaluacio-docent/Avaluacio-docent-graus-i-masters/

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6496

Estudio longitudinal de las calificaciones de evaluación continua en la asignatura de Arquitectura de Redes II del Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Jorge E. López de Vergara*, Ricardo Olmos†
*Departamento de Tecnología Electrónica y de las Comunicaciones,
†Departamento de Psicología Social y Metodología
Universidad Autónoma de Madrid
Campus de Cantoblanco, s/n, 28049 Madrid
{jorge.lopez_vergara, ricardo.olmos}@uam.es

***Resumen*—En este artículo se realiza un estudio longitudinal de las calificaciones de evaluación continua de la asignatura Arquitectura de Redes II, que se imparte en el Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación ofertado en la Universidad Autónoma de Madrid. Para ello, se analizan las calificaciones obtenidas en los veinticuatro ejercicios de evaluación continua realizados durante seis cursos académicos. Los resultados muestran que, en general, existe una alta correlación en las calificaciones que obtienen los estudiantes entre los distintos ejercicios de evaluación de cada curso, lo cual permite corroborar que estos han sido correctamente planteados. Igualmente, se comparan los resultados obtenidos entre cursos distintos, observándose que los resultados tienen similitudes en las distintas promociones, si bien se observa cierta variación en las calificaciones a lo largo de los años.**

***Palabras Clave*—Correlación de calificaciones, correlación corregida, α de Cronbach, consistencia interna.**

I. INTRODUCCIÓN

Con la introducción de los grados del Espacio Europeo de Educación Superior, las asignaturas han pasado a utilizar metodologías en las que suele realizar una evaluación continua de las mismas [1]. Esto permite un seguimiento continuado del trabajo de los estudiantes, lo cual debe redundar en una mayor calidad de la enseñanza. No obstante, conviene revisar si dicha evaluación continua ha sido adecuada, o por contra, está teniendo algún tipo de desviación que no sea fácilmente identificable. Para ello, se pueden aprovechar las técnicas propuestas en Psicometría [2] para estudiar la fiabilidad de las calificaciones obtenidas por los estudiantes.

Por ello, en este artículo se plantea la necesidad de analizar en profundidad las calificaciones de la asignatura Arquitectura de Redes II, del grado en Ingeniería de Tecnologías y Servicios de Telecomunicación que se oferta en la Universidad Autónoma de Madrid desde el curso 2011-12. La asignatura, de 6 ECTS, se imparte en el segundo semestre de segundo curso, y se cubren cuatro temas: Teoría de Colas, Nivel de Enlace, Redes Inalámbricas y Móviles, y Seguridad y Gestión de Redes.

Desde su implantación, la asignatura ha seguido exactamente la misma metodología docente y de evaluación. La metodología de evaluación continua para la parte teórica de la asignatura consiste fundamentalmente en la realización de cuatro ejercicios parciales, uno por tema, a lo largo del cuatrimestre, combinado con la realización de problemas en clase por parte de los estudiantes. La superación de todos los ejercicios de evaluación permite aprobar la asignatura sin necesidad de realizar un examen final, calificándose la nota final como la ponderación de los ejercicios de evaluación y los problemas de clase. Si un estudiante suspende alguno de los ejercicios de evaluación debe entonces realizar el examen final de la asignatura, si bien puede seguir la evaluación continua, pues con ello puede mejorar la nota final. En este último caso, la calificación obtenida en evaluación continua se pondera como un 40% de la nota final.

Los ejercicios parciales se realizan mediante preguntas de opción múltiple (tipo test), con 4 opciones posibles, una sola válida, y penalizaciones de 1/3 de la puntuación para respuestas incorrectas, siguiendo las indicaciones habituales para este tipo de evaluación [3]. El primer ejercicio de evaluación se corresponde con la resolución de un problema de Teoría de Colas junto con algunas preguntas sueltas de teoría, con una duración de 50 minutos. Este primer ejercicio, al estar sujeto al problema a realizar, varía entre 12 y 15 preguntas. Los otros tres ejercicios de evaluación consisten en 20 preguntas relativas a ejercicios

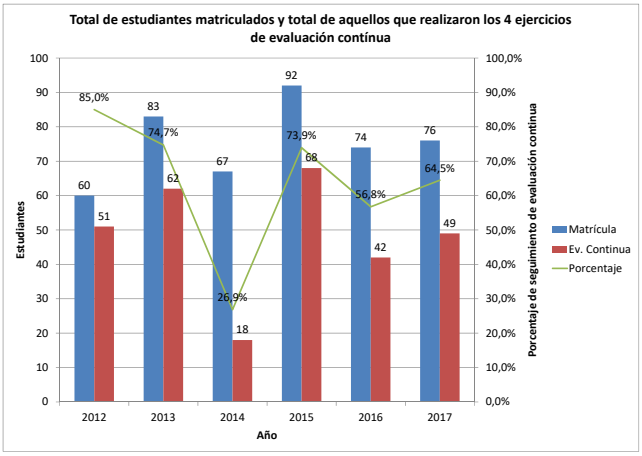


Fig. 1. Población de estudiantes del estudio y porcentaje de seguimiento de la evaluación continua para los distintos cursos estudiados.

cortos y preguntas teóricas de los tres temas restantes, con una duración de 30 minutos.

En este artículo se analizan las siguientes cuestiones:

1) La correlación entre ejercicios parciales con respecto a la nota final, y la nota final corregida sin tener en cuenta cada ejercicio (correlación corregida), para aquellos estudiantes que realizan todos los ejercicios de evaluación continua. Cabe mencionar en este caso que, respecto del total de matriculados, suele haber muchos abandonos del itinerario de evaluación continua, con lo que no se puede hacer el estudio sobre todos los estudiantes, lo cual seguramente genere algún sesgo en el análisis. La figura 1 presenta la comparación entre estudiantes totales matriculados y estudiantes que finalizaron el itinerario de evaluación continua. Según se observa, dependiendo del año, los abandonos son variables, lo que también va a influir sobre la calificación media final de cada curso, como se verá más adelante.
2) Adicionalmente, se estudia si existe coherencia en el conjunto de los ejercicios de evaluación en cada curso académico. Para ello se utiliza el coeficiente α de Cronbach [4].
3) La evolución de la asignatura a lo largo de los años, de forma que se pueda ver si las correlaciones obtenidas previamente son más o menos estables entre distintos cursos académicos o cohortes.
4) De forma añadida a lo anterior, se analiza si la dificultad de la asignatura ha variado sustancialmente, estudiando cómo ha variado el rendimiento en los cuatro ejercicios de evaluación planteados.

El estudio de la fiabilidad de las calificaciones muestra parcialmente si una evaluación es justa, porque informa de si el fenómeno en estudio se mide con precisión y con poco error. Para que una evaluación sea justa es condición necesaria tener una alta fiabilidad (no suficiente, puesto que la fiabilidad no garantiza que los conceptos evaluados representen bien el universo de contenidos que deben evaluarse en una asignatura). La evaluación de una asignatura con poca fiabilidad presentará variabilidad en las calificaciones por cuestiones ajenas al nivel de conocimientos de nuestros estudiantes, lo cual provoca una situación de injusticia [5].

Por ello, las respuestas a las cuestiones aquí planteadas permitirán disponer de una metodología formal para estudiar la fiabilidad del proceso de evaluación continua en esta u otras asignaturas de características similares. Tras realizar una búsqueda bibliográfica, no hemos encontrado estudios que se asemejen a este, o al menos que sean tan específicos como este, en el ámbito de la docencia en el área de Ingeniería Telemática.

Como trabajo relacionado al aquí expuesto, es relevante citar al de McKenzie y Schweitzer [6], donde se buscaban factores que predijeran el rendimiento académico en universitarios mediante un estudio longitudinal, y en el que se observó que el rendimiento académico está muy correlacionado. Esto es, el comportamiento habitual es que cada estudiante obtenga calificaciones parecidas a lo largo de los cursos.

Para llevar a cabo el trabajo planteado, primeramente se presentan en la sección II los indicadores psicométricos que se utilizan en este estudio. A continuación se realiza un análisis anual de correlación entre las calificaciones de los cuatro ejercicios parciales en la sección III, y posteriormente se estudia su evolución a lo largo de los distintos cursos en la sección IV. Finalmente se discuten los resultados obtenidos, proporcionando las conclusiones de todo el análisis y se proponen líneas futuras de continuación.

II. INDICADORES PSICOMÉTRICOS

Antes de realizar el análisis propuesto, y para comprender mejor los fundamentos en que se basa, es necesario presentar los indicadores psicométricos que se van a utilizar.

El primero de estos indicadores es el coeficiente de correlación de Pearson, que se define como:

$$r_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N \cdot S_X \cdot S_Y} \qquad (1)$$

donde X e Y son las variables aleatorias a comparar, $\bar{X}$ e $\bar{Y}$, sus valores esperados, N el número de elementos, y S_X y S_Y las desviaciones típicas de X e Y.

El coeficiente de correlación de Pearson toma valores en el intervalo $[-1, 1]$. Cuanto más próximo sea su valor a 1, mayor será la relación directa entre las variables aleatorias, y cuando más próximo a 0, menor. Si la correlación es negativa, la relación entre ambas variables será inversa.

En el presente trabajo se utiliza el coeficiente de correlación para comparar las calificaciones de cada ejercicio de evaluación continua con la nota final obtenida como media de los cuatro ejercicios de cada curso. Adicionalmente, se utiliza la correlación corregida, donde cada ejercicio se compara con una nota media de la que se excluye del cómputo global dicho ejercicio. Desde un punto de vista psicométrico, se recomienda que los valores de correlación sean superiores a 0,2 [2].

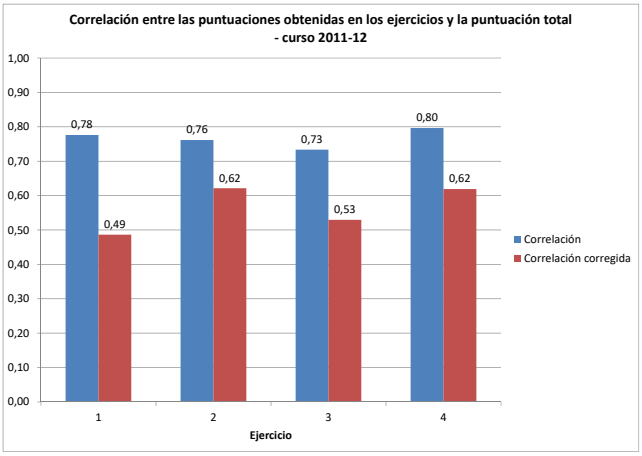


Fig. 2. Correlación de puntuaciones en el curso 2011-12.

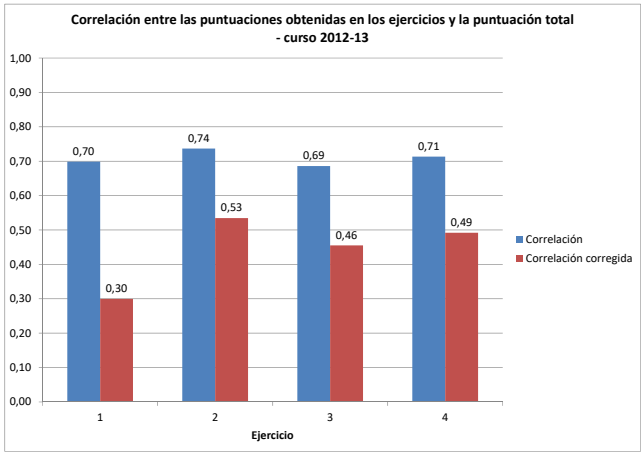


Fig. 3. Correlación de puntuaciones en el curso 2012-13.

El segundo indicador es el parámetro α de Cronbach [4], definido como:

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} S_j^2}{S_X^2}\right) \qquad (2)$$

donde J se corresponderá con el número de ejercicios de evaluación (cuatro en el caso de estudio) y S_j^2 la varianza del ejercicio de evaluación j y S_X^2 la varianza de la calificación global de evaluación continua. Desde un punto de vista psicométrico, se debe buscar que la consistencia de los ejercicios de evaluación tenga valores de $\alpha > 0,7$ [7].

No todos los expertos están de acuerdo en que 0,7 sea un buen estándar como referencia mínima de la fiabilidad de un test [8]. Obviamente, este umbral dependerá del contexto, de las consecuencias y de las implicaciones que tiene el instrumento de medida para una persona (por ejemplo, no es lo mismo una evaluación de la que depende quién accede y quién no a un empleo que otra que forme parte de un estudio piloto para evaluar la calidad de un producto). Un examen universitario se considera una evaluación de altas consecuencias por lo que se debe ser riguroso a la hora de fijar el valor mínimo de α.

Finalmente, el tercer indicador es el índice de dificultad, que se obtiene como la media de las calificaciones en cada ejercicio. Si la media es alta, el ejercicio fue fácil, y si por contra es baja, difícil.

III. CORRELACIÓN ENTRE LAS CALIFICACIONES DE LOS DISTINTOS EJERCICIOS

Para responder a la primera de las cuestiones planteadas en la introducción, las figuras 2 a 7 muestran el resultado de haber obtenido la correlación y la correlación corregida (esto es, excluyendo cada ejercicio de la media para calcular su correlación con el resto) obtenida para las calificaciones obtenidas en los cuatro ejercicios de evaluación a lo largo de los distintos cursos, aplicando la Ec. 1.

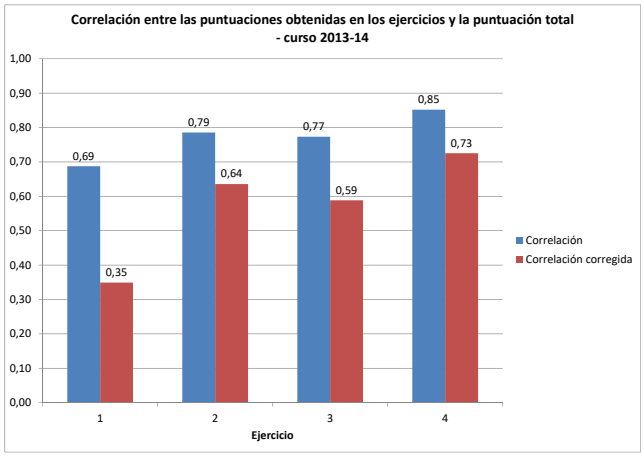


Fig. 4. Correlación de puntuaciones en el curso 2013-14.

Según se puede advertir, en general, la nota de los distintos ejercicios parciales tiene una correlación alta con la nota promediada. Al estar dicha correlación en el entorno de 0,7, es un resultado bueno desde un punto de vista psicométrico en relación a cómo se ha realizado la evaluación. Si se utiliza la correlación corregida, sigue siendo igualmente alta, en el entorno de 0,5 aunque con una variación mayor según cada curso.

Únicamente se observa un caso del total de veinticuatro ejercicios parciales analizados, el primero del curso 2014-15, cuyas correlaciones son realmente bajas. Este hecho ha supuesto la necesidad de estudiar este ejercicio en más profundidad. La figura 8 presenta un diagrama de dispersión donde se puede ver este fenómeno con mayor detalle, representando las calificaciones de dicho ejercicio respecto de la nota media y la nota media corregida excluyendo la calificación de dicho ejercicio.

Como puede inferirse a partir de los puntos rojos con forma de aspa (X), donde se representa la calificación obtenida en el ejercicio frente a la nota media del resto de parciales, no existe una tendencia clara. De hecho, si se realiza una regresión, se obtiene una pendiente prácticamente plana y un coeficiente de determinación R^2 cercano a cero, lo que indica que la nota obtenida en dicho

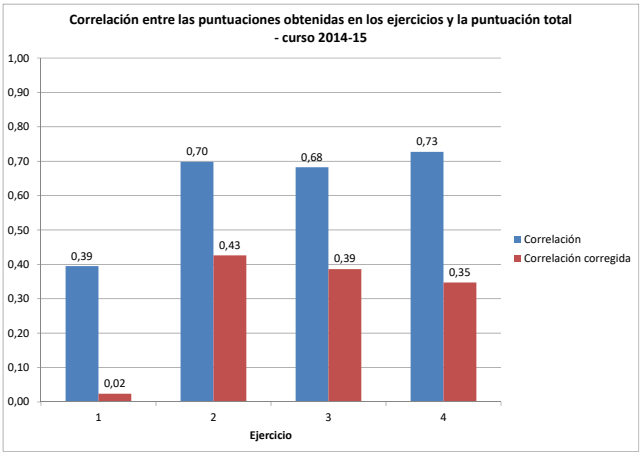


Fig. 5. Correlación de puntuaciones en el curso 2014-15.

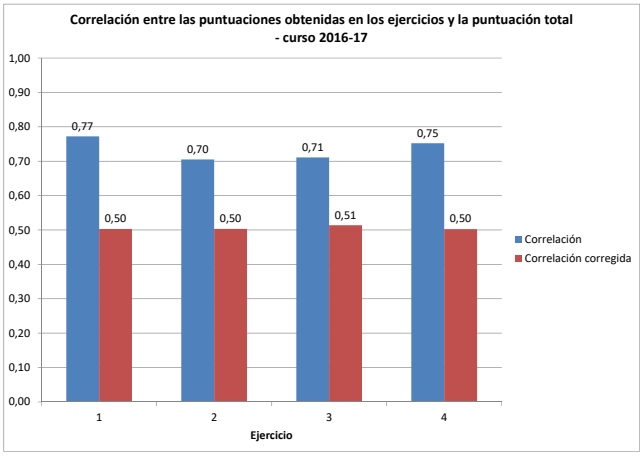


Fig. 7. Correlación de puntuaciones en el curso 2016-17.

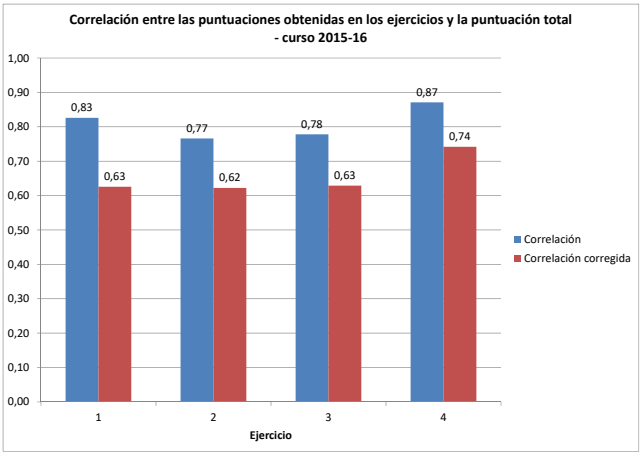


Fig. 6. Correlación de puntuaciones en el curso 2015-16.

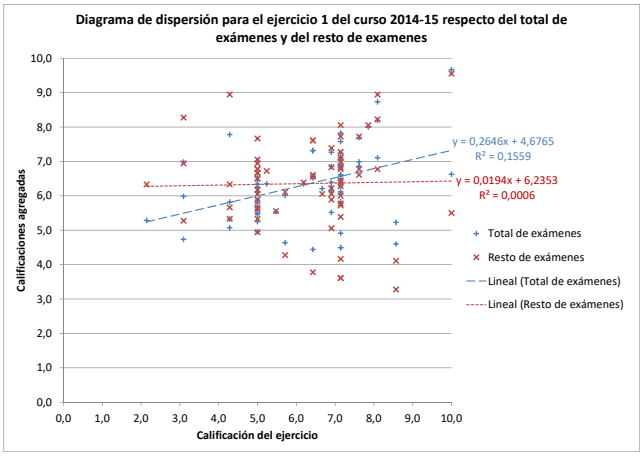


Fig. 8. Diagrama de dispersión para el ejercicio 1 del curso 2014-15 respecto del total de ejercicios de evaluación y del resto de ejercicios.

ejercicio no permite explicar la nota del resto de ejercicios de evaluación continua de ese curso académico.

Las posibles causas de este hecho podrían ser las siguientes:

- Al ser un ejercicio de evaluación continua, se realizaba en las horas de clase, y coincidió con otro examen de otra asignatura en las horas previas, lo que posiblemente provocó que muchos alumnos no pudieran dedicarle el tiempo de estudio suficiente, a diferencia del resto de ejercicios de evaluación en dicho curso.
- De forma complementaria, al ser el primer ejercicio de evaluación del curso, los estudiantes podrían estar todavía tanteando la dificultad de la asignatura, y no supieron valorarla adecuadamente. Este hecho puede verse reflejado igualmente en el resto de cursos.

Por otro lado, según se estudiará más adelante, se puede determinar que el ejercicio 4, de final de curso, es el que suele obtener en la mayoría de los casos mejores valores de correlación con la nota final y con el resto de notas. Esto puede deberse, siguiendo el razonamiento del caso anterior, a que en este caso los estudiantes ya tienen más clara la dificultad de la asignatura, tras haber ido realizando los exámenes anteriores.

IV. EVOLUCIÓN A LO LARGO DE LOS CURSOS

Para conocer cómo ha sido la evolución de la dificultad de la asignatura a lo largo de los cursos, se han realizado distintas medidas. La primera de ellas es la de las notas medias de los ejercicios de evaluación continua para los estudiantes que realizaron todos los ejercicios del curso, y la nota media final de los mismos, que se representa en la figura 9, con los valores también detallados en la tabla I.

Para el caso de la nota media de cada curso, en la figura se representa adicionalmente dónde se encontrarían los valores de +/- una desviación típica para el conjunto de los individuos estudiados en dicho curso. Para entender mejor estos valores medios, es necesario también comparar con el número de estudiantes que realizaron los cuatro ejercicios de evaluación cada año, según se mostraba en la figura 1.

Según se observa, las puntuaciones fueron en promedio evolucionando de forma positiva hasta el curso 2013-14, año en el que siguieron la evaluación continua un menor porcentaje de estudiantes, siendo por tanto el curso con mayor sesgo. Posteriormente, las calificaciones han caído en los años siguientes. En el caso de los cursos 2015-16 y 2016-17 las calificaciones de los estudiantes muestran

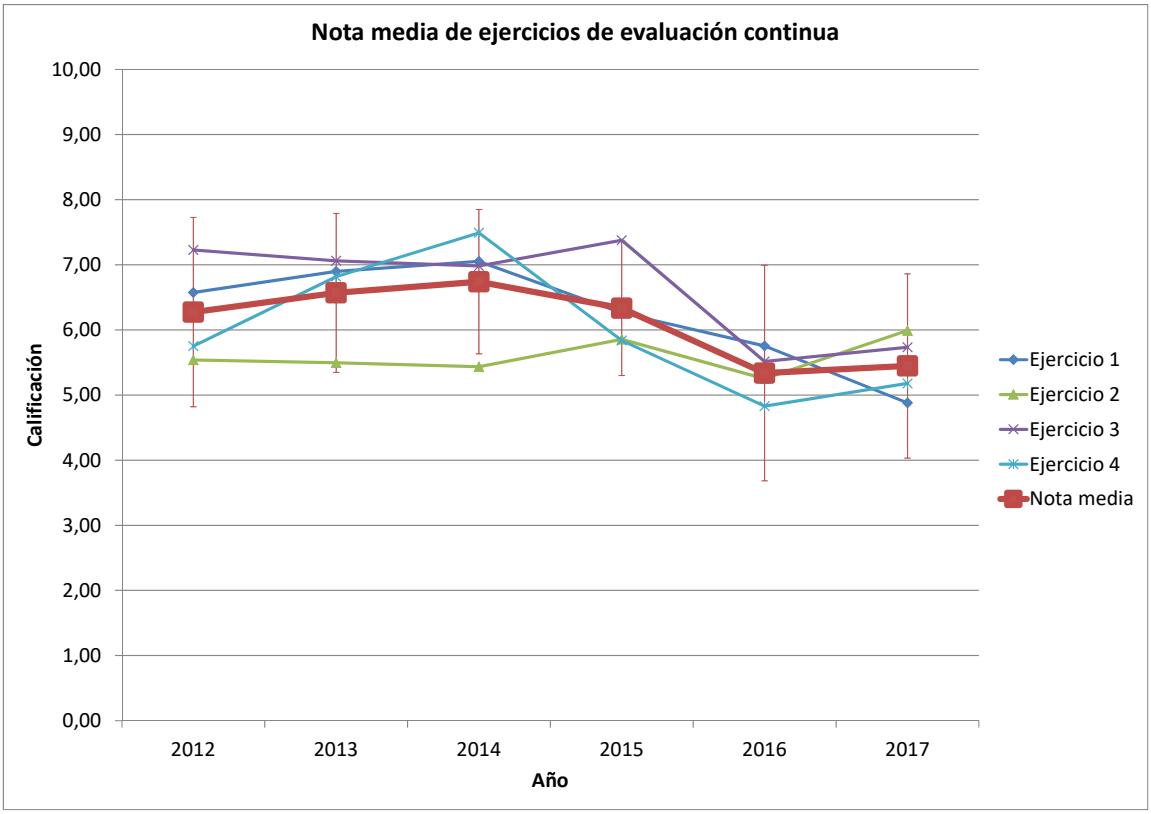


Fig. 9. Evolución de la nota media de los ejercicios de evaluación continua.

Tabla I
NOTA MEDIA DE LOS EJERCICIOS DE EVALUACIÓN CONTINUA. DETALLE DE VALORES PARA LOS DISTINTOS CURSOS ACADÉMICOS.

Curso	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17
Nota media global	6,27	6,57	6,74	6,33	5,34	5,45
Desviación típica	1,45	1,22	1,11	1,03	1,66	1,41
Ejercicio 1	6,58	6,90	7,05	6,26	5,75	4,88
Ejercicio 2	5,54	5,50	5,44	5,86	5,25	5,99
Ejercicio 3	7,23	7,06	6,98	7,38	5,52	5,73
Ejercicio 4	5,75	6,82	7,49	5,84	4,83	5,18

que hubo muchos que permanecieron realizando todos los ejercicios pese a haber suspendido uno o varios de ellos. Este hecho explicaría la caída de la nota media global en los últimos cursos, así como las notas medias por debajo del 5,0 en algunos ejercicios, al reducirse también el sesgo que se planteaba inicialmente.

Además, con respecto a las notas medias y su desviación típica, se observa que estos indicadores no son capaces de identificar que en el curso 2014-15 existiese la singularidad descubierta usando las correlaciones entre el primer ejercicio y el resto.

En relación a la evolución temporal de las correlaciones, las figuras 10 a 13 muestran cómo ha ido variando la correlación y correlación corregida para cada uno de los ejercicios de evaluación a lo largo del tiempo. Puede decirse que la evolución para los cuatro parciales ha sido bastante semejante, existiendo la singularidad explicada en la sección anterior del curso 2014-15, en el que se produce la incorrelación del ejercicio 1 con respecto al resto de ejercicios. Adicionalmente, el hecho de que el primer ejercicio sea el de inicio del curso y que su ejecución sea algo distinta provoca que las correlaciones de este ejercicio en general son más bajas que las del resto.

No obstante, en general existe una correlación alta en las calificaciones. Para corroborar este extremo, se ha calculado igualmente el valor del parámetro α de Cronbach definido en la Ec. 2 para los distintos cursos académicos, como se muestra en la figura 14, para tener una medida de la consistencia interna entre los cuatro exámenes parciales.

Según queda reflejado en la gráfica, la consistencia es bastante alta, por encima de 0,7 para cuatro de los cursos estudiados. En el curso 2014-15 se produce el valor más bajo para el parámetro α, siendo este hecho coherente con los valores de las correlaciones explicadas anteriormente.

V. CONCLUSIONES

Como conclusiones al estudio realizado respecto de la evaluación continua en la asignatura Arquitectura de Redes II, y en respuesta a las cuestiones planteadas en la introducción, se puede extraer que:

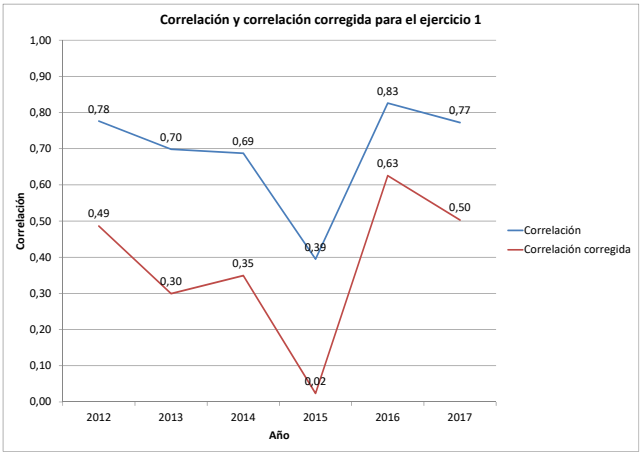


Fig. 10. Evolución de la correlación y correlación corregida para el ejercicio 1.

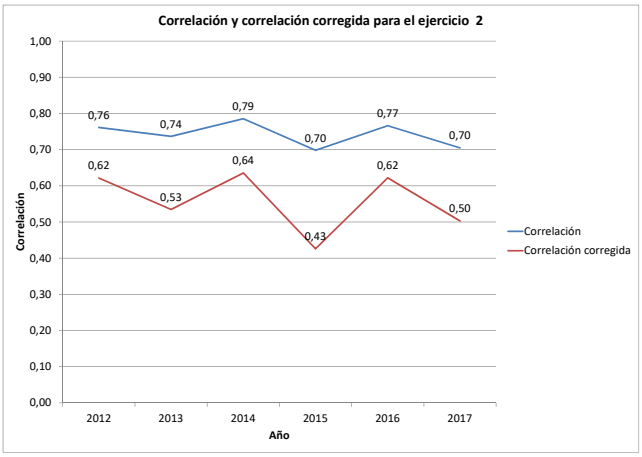


Fig. 11. Evolución de la correlación y correlación corregida para el ejercicio 2.

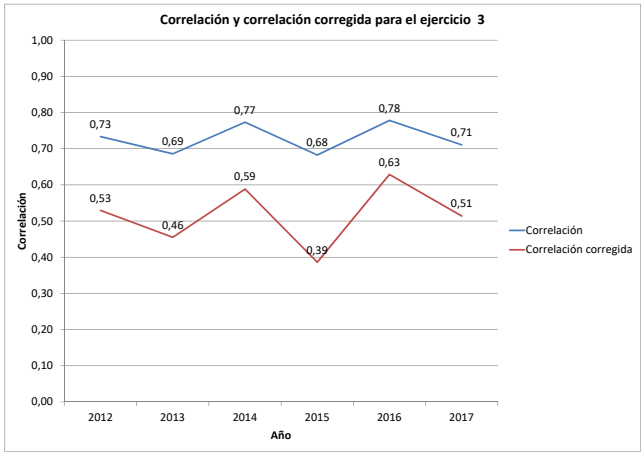


Fig. 12. Evolución de la correlación y correlación corregida para el ejercicio 3.

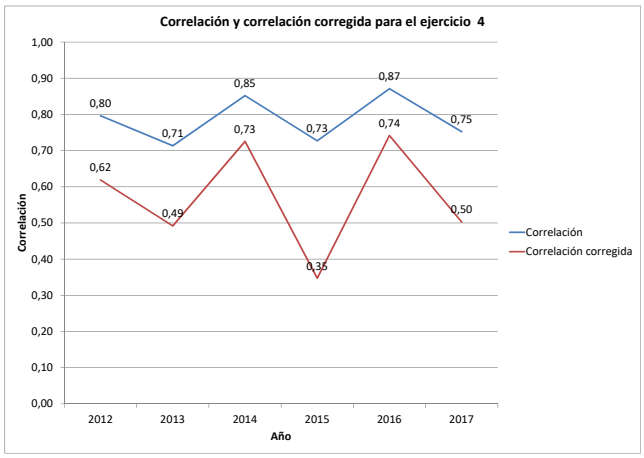


Fig. 13. Evolución de la correlación y correlación corregida para el ejercicio 4.

1) En general hay una correlación alta entre ejercicios parciales con respecto a la calificación final de evaluación continua, así como con la nota corregida que se obtiene excluyendo cada ejercicio para calcular la correlación. Se ha observado únicamente un caso de los veinticuatro ejercicios parciales estudiados en el que dicha correlación podría considerarse no adecuada. En general, el último ejercicio suele tener una mayor correlación con la nota final y la nota del conjunto de los otros tres parciales. Por otro lado, el primer ejercicio suele ser el que menos correla con el resto de calificaciones, si bien en casi todos los cursos está dentro de los márgenes aceptables.
2) A partir del valor del parámetro α de Cronbach se ha visto igualmente que en general hay coherencia en el conjunto de los ejercicios de evaluación de cada curso académico, a excepción del curso 2014-15, en el que se produce el caso singular identificado previamente, que reduce el valor de α en dicho curso.
3) En general, exceptuando el caso singular, las correlaciones que se obtienen para los distintos parciales a lo largo de los años son bastante semejantes. Hay que tener en cuenta que las correlaciones (y medidas que dependen de ellas como el α de Cronbach) cuentan con un error muestral relativo mayor que el que tienen otros estadísticos como las medias o proporciones [9]. Por lo tanto, las fluctuaciones en este indicador también pueden ser debidas en parte a efectos del muestreo si la muestra es pequeña ($N < 100$), como de hecho ocurre en todos los cursos estudiados.
4) Adicionalmente, la dificultad de los ejercicios de evaluación parece haber ido en aumento en los últimos cursos, si bien esto puede deberse también a que muchos estudiantes han seguido presentándose a los parciales pese a haber suspendido alguno de los mismos en los últimos cursos estudiados. Esta cuestión no ocurrió así en el curso 2013-14, donde todos los que siguieron la evaluación continua aprobaron todos los ejercicios, explicando este hecho que sea el curso con mejor media final, debido al sesgo introducido.

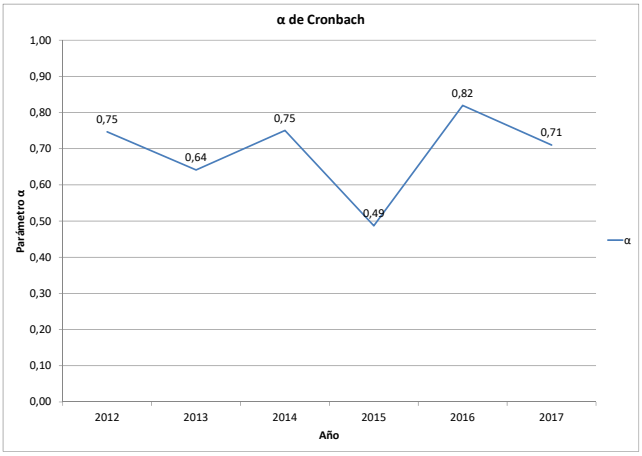


Fig. 14. Evolución del parámetro α de Cronbach para los distintos cursos académicos.

La versión final de este artículo ha incluido también en el estudio al curso 2016-17, cuyos datos no estaban disponibles en la primera versión enviada a revisión. En relación con esta ampliación, se ha comprobado que la metodología planteada permite realizar un estudio riguroso que analice en detalle los resultados de cada curso. A la vez, es relativamente sencillo de implementar con una hoja de cálculo, sin tener que acudir a herramientas estadísticas sofisticadas. Además, los resultados proporcionan un soporte adicional al personal docente frente a una posible reclamación de las calificaciones obtenidas por los estudiantes. No obstante, esta metodología plantea también el inconveniente de que, hasta que no estén todas las calificaciones de todos los ejercicios disponibles, no es posible tener una visión de la marcha del curso, por lo que sería oportuno estudiar la aplicación de otras metodologías que sí lo permitan.

Por tanto, como trabajos futuros se plantea, por un lado, una continuación de este estudio para los cursos subsiguientes, de forma que puedan detectarse problemas en años venideros; y por otro lado, estudiar cómo llevar a cabo adecuadamente un estudio individual de fiabilidad de los distintos ejercicios de evaluación.

Dicho estudio de fiabilidad de cada una de las preguntas de cada ejercicio de evaluación permitirá hacer un seguimiento continuado del proceso de evaluación continua. No obstante, no ha sido posible para el trabajo presente, dado que las evaluaciones se realizan usando 4 modelos de test con las preguntas y respuestas barajadas, lo que dificulta realizar un procesado con el programa TAP [10] que se utiliza habitualmente para este tipo de análisis. Una posible opción para alcanzar este objetivo puede venir dada por el uso de herramientas de generación de exámenes de opción múltiple, tales como *Auto Multiple Choice* [11].

Finalmente, también puede ser de interés llevar a cabo un estudio longitudinal que analice formalmente los resultados obtenidos por los estudiantes que abandonaron el itinerario de evaluación continua y optaron por acudir al examen de evaluación final, comparándolos con los aquí presentados.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional, a través del proyecto Tráfica (MINECO/FEDER TEC2015-69417-C2-1-R).

Los autores agradecen igualmente a la Universidad Autónoma de Madrid por la realización dentro de su Programa de Formación del Profesorado del curso "Evaluación del Aprendizaje", que ha posibilitado el desarrollo de este estudio.

REFERENCIAS

[1] J. M. Giménez Guzmán, E. de la Hoz, M. T. López y M. Moreno, "La evaluación continua en la docencia de Ingeniería Telemática en el EEES", en *Actas de las II Jornadas de Innovación Educativa en Ingeniería Telemática*, Santander, 29 de septiembre de 2011.

[2] F. J. Abad, J. A. Olea, V. Ponsoda y C. García, "Medición". Pirámide. Madrid, 2011.

[3] R. F. Burton, "Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers". *Assessment and Evaluation in Higher Education*, Vol. 26, pp. 41–50, 2001.

[4] L. J. Cronbach, "Coefficient alpha and the internal structure of tests". *Psychometrika*, Vol. 16(3), pp. 297-334, 1951.

[5] C.S. Wells y J.A. Wollack, "An instructor's guide to understanding test reliability". *Testing & evaluation Services*. University of Wisconsin, 2003

[6] K. McKenzie y R. Schweitzer, "Who succeeds at university? Factors predicting academic performance in first year Australian university students". *Higher education research & development*, Vol. 20(1), pp. 21-33, 2001.

[7] N. Schmitt, "Uses and abuses of coefficient alpha". *Psychological assessment*, Vol. 8(4), p. 350, 1996.

[8] C. E. Lance, M. M. Butts y L. C. Michels, "The sources of four commonly reported cutoff criteria: What did they really say?" *Organizational research methods*, Vol. 9(2), pp. 202–220, 2006.

[9] L. Feldt, D. J. Woodruff y F. A. Salih, "Statistical Inference for Coefficient Alfa". *Applied Psychological Measurement*, Vol. 11(1), pp. 93–103, 1987.

[10] G. P. Brooks y G. A. Johanson, "Test Analysis Program", *Applied Psychological Measurement*, Vol. 27, pp. 305-306, 2003.

[11] A. Bienvenüe, "AMC: Auto Multiple Choice", http://home.gna.org/auto-qcm/

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6508

Servicio centralizado de proyección de material docente

Jorge Navarro-Ortiz, Sandra Sendra,
Pablo Ameigeiras, Angel de la Torre, Luz Garcia, Angel M. Gomez,
Juan M. Lopez-Soler, Sonia Mota, Pablo Padilla, Jonathan Prados-Garzon,
Javier Ramirez, Juan J. Ramos-Munoz, Antonio Ruiz-Moya, José C. Segura
Departamento de Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada
C/ Periodista Daniel Saucedo Aranda s/n, 18071 Granada
jorgenavarro@ugr.es

***Resumen*—En los últimos años las tecnologías TIC se han ido incorporando en los diferentes ámbitos de la enseñanza, desde las pizarras electrónicas para las clases magistrales hasta el uso de tabletas para la visualización de libros docentes en formato electrónico. De hecho, resulta cada vez más frecuente que los docentes empleen sus portátiles para presentar su material en formato de transparencias. No obstante, esto implica que los profesores deben llevar sus portátiles al aula y conectarlos a través de un cable, sea VGA o HDMI, al proyector. Esto resta movilidad al profesor, anclado a través del cable al proyector, además de requerir que disponga de un portátil que ha de llevar al aula. Dado que, en la actualidad, casi la totalidad de la población dispone de móviles inteligentes, este artículo presenta la solución propuesta en un proyecto de innovación docente (PID 14-61) desarrollado en la Universidad de Granada. En éste, se propone una solución en la que el profesor sólo deberá llevar su móvil (o alternativamente una tableta o un portátil) al aula. El material docente será subido a un servidor central desde su despacho, y la visualización en el proyector será controlada a través del móvil usando una interfaz muy amigable y sencillo.**

***Palabras Clave*—Proyección, inalámbrico, sistema centralizado, proyecto innovacion docente.**

I. INTRODUCCIÓN

En la actualidad muchas de las aulas de docencia de la Universidad de Granada cuentan con equipamiento para realizar la proyección de material docente, típicamente en formato de transparencias, que ayude a la impartición de las clases en la modalidad de presentación magistral. Esta proyección requiere conectar un ordenador portátil a través de un cable VGA, o bien transferir el material docente desde una memoria USB a un ordenador fijo (en el caso de las aulas donde este ordenador esté disponible). Sin embargo, esta forma de presentación conlleva tres limitaciones importantes: 1) el tipo de dispositivos que se puede utilizar, 2) la movilidad del profesor en el aula, y 3) la necesidad de llevar al aula un soporte informático (sea un portátil o una memoria USB) con el material docente. Así, por ejemplo, con el sistema actual un profesor no podría utilizar una tableta o un smartphone para presentar su material. Por otro lado, el profesor no puede moverse a través del aula si tiene que cambiar de transparencia, debido al cable VGA. Finalmente, no existe la posibilidad de que el profesor elija en su despacho el material que debe estar disponible en el aula sin tener que llevarlo en algún tipo de soporte.

Para solventar estas limitaciones, se propone el diseño e implementación de un sistema informático que permita 1) almacenar en un servidor el material docente a visualizar en clase, sin necesidad de llevarlo en ningún tipo de soporte informático; 2) elegir dicho material a través de una gran variedad de equipos ampliamente utilizados, como un ordenador portátil, un móvil o tableta Android, un iPAD o un iPhone; 3) poder avanzar y retroceder en las transparencias mostradas a través de los equipos comentados; y 4) todas estas funcionalidades (selección de material, navegación por transparencias) deben ser accesibles inalámbricamente, de forma que no sea necesario la conexión del equipo a través de un cable VGA (o de otro tipo), máxime cuando muchos de estos dispositivos no incluyen este tipo de interfaces.

Este trabajo es el resultado de un proyecto de innovación docente (PID 14-61) desarrollado en la Universidad de Granada. Los autores de este artículo han trabajado activamente en la fase de desarrollo y/o en la fase de pruebas del mismo.

El resto del paper se estructura como sigue. La sección II presenta algunas herramientas existentes que facilitan la reproduccion de contenido multimedia durante una clase magistral. La sección III describe el sistema desarrollado. A lo largo de la sección IV se exponen los objetivos que

persigue este dispositivo. La sección V muestra el funcionamiento de la plataforma desarrollada. Los resultados obtenidos y servicios que proporciona nuestro sistema es mostrado en la sección VI. Finalmente, la sección VII expone las conclusiones acerca del trabajo realizado.

II. ESTADO DEL ARTE

En esta sección se comentan algunos dispositivos que podrían tener una funcionalidad similar a la del sistema propuesto pero, por diferentes motivos, no son adecuados.

A. Chromecast de Google

Chromecast [4] es un dispositivo de reproducción de contenido multimedia que se conecta al puerto HDMI de una pantalla (sea TV, monitor o proyector). A través de un dispositivo móvil se le puede enviar diferentes tipos de contenidos multimedia, como vídeos, música, etc.

Además de ser compatible con la mayoría de dispositivos móviles del mercado (iPhone/iPad, móviles y tabletas Android, portátiles Mac/Windows y Chromebooks), tiene un precio reducido (39 euros [4]) y cuenta con miles de applicaciones disponibles [5]. Entre éstas se incluyen aplicaciones para la reproducción de vídeo/audio, pero también para la visualización de documentos en los formatos habituales (e.g. Google Slides, Polaris Office, OfficeSuite, etc.).

Sin embargo, su principal inconveniente es que está orientado a hogares, no a empresas. Así, no soporta la autenticación mediante IEEE 802.1X para la conexión Wi-Fi (típicamente conocida como *WPA-Enterprise*), por lo que no es capaz de conectarse a redes empresariales o universitarias (e.g. eduroam). Esta carencia provoca que no se haya contemplado Chromecast para el presente trabajo.

B. Apple TV

Apple TV [7] es un receptor digital multimedia diseñado por Apple. Este dispositivo permite reproducir todo el contenido de los dispositivos iOS y Mac a través del protocolo AirPlay. Además proporciona acceso a diferentes contenidos, disponibles en iTunes. Su funcionamiento es similar al de Chromecast, y se puede utilizar junto con un iPad para sustituir en las aulas a las pizarras electrónicas.

Sin embargo, presenta dos grandes inconvenientes para los objetivos del PID. Por un lado, tampoco soporta *WPA-Enterprise*, por lo que no se podría utilizar en la mayoría de redes universitarias. Por otro lado, sólo funcionaría con dispositivos de Apple lo que limitaría su uso[1].

C. Proyectores inalámbricos

Existen proyectores de los principales fabricantes con tecnologías Wi-Fi o Bluetooth para la conexión con el ordenador que proporciona las imágenes. Sin embargo, esta opción está disponible sólo para los proyectores de gama alta (e.g. PowerLite 975W WXGA 3LCD Projector, con un coste de $1,599 [6]) o a través de un módulo inalámbrico vendido aparte (e.g. Wireless LAN module de EPSON, con un coste de $99 [7]) y habitualmente más caro que el sistema propuesto.

El principal inconviniente de este tipo de sistemas es que su uso implicaría la sustitución de todos los proyectores disponibles, lo que implicaría un coste muy elevado. Además, en general utilizan soluciones propietarias lo que reduciría su flexibilidad y la obligación de utilizar un fabricante concreto. Por estos motivos, tampoco se contempló esta opción para la presente propuesta.

[1]Existen implementaciones de pago para Android/Windows, pero en general tienen limitaciones y son susceptibles de no funcionar en el futuro ya que no están soportadas por Apple.

III. DESCRIPCIÓN DEL SISTEMA

Tal como se ha introducido, este trabajo surge con la idea de que los profesores 1) no necesiten llevar el material docente a clase y 2) no tengan que estar conectados al proyector a través de un cable. Para ello, se ha diseñado e implementado un sistema con tres entidades principales:

1) Un **servidor central**, con varias funcionalidades: a) **almacenamiento** del material docente, b) **servidor web** que aloja todas las páginas web que conforman la interfaz con el usuario, c) **servidor VPN** para que todas las comunicaciones sean seguras, d) **controlador** de todos los dispositivos conectados a los proyectores, ordenando a los mismos ejecutar las acciones indicadas por el usuario.
2) Un **dispositivo conectado al proyector** en el aula: este dispositivo muestra una interfaz sencilla con la información necesaria para que el profesor se pueda conectar al mismo. Recibe órdenes por parte del servidor central, de forma que éste controla el material a visualizar, la navegación por las transparencias y cualquier otra acción que sea necesaria.
3) **Móvil del profesor** (o una tableta o un portátil, con cualquiera de los principales sistemas operativos e.g. Android, iOS, Windows, Linux). El profesor seleccionará, a través de una interfaz web, el aula/dispositivo a utilizar y el material docente a visualizar. Igualmente, a través de una interfaz muy amigable podrá navegar a través de las diferentes transparencias. Esta solución cumple todos los objetivos que se detallarán en la siguiente sección.

IV. OBJETIVOS

Los objetivos de este trabajo son los siguientes:

1) **Diseño e implementación de un servidor para el almacenamiento de material docente**. Este almacenamiento será realizado por el profesor desde su despacho (o desde otras ubicaciones como su propia casa), típicamente desde un ordenador, véase la Fig. 1. La interfaz de usuario ha sido realizado para ser amigable, usándose para ello diferentes páginas web con las opciones oportunas. Además, utiliza el sistema de autenticación de la Universidad de Granada, por lo que es seguro y se evita tener que guardar credenciales (e.g. usuario y clave) de los profesores. Este sistema también permite la eliminación del material por parte del usuario que lo almacenó, así como la modificación del nombre de los ficheros.
2) **Diseño e implementación de un servicio que permita seleccionar el material docente**. Esta

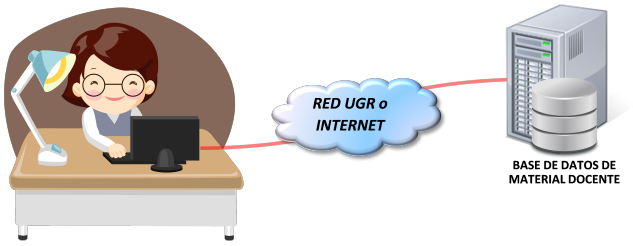


Fig. 1. Escenario típico para el objetivo 1 (almacenamiento de material docente en un servidor por parte del profesor).

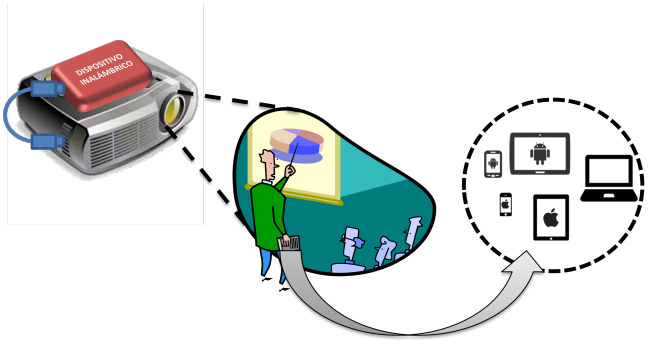


Fig. 2. Escenario típico para la selección de material, la visualización y la navegación a través de las transparencias.

selección se puede realizar por cualquiera de los dispositivos contemplados (portátil, tableta, móvil) desde el aula donde se impartirá la docencia. El único requisito es que el dispositivo soporte un navegador web con *javascript*, por lo que se soportan todas las plataformas habituales (Android, iOS, Windows, Linux). La interfaz es amigable y se implementa a través de diferentes páginas web. Se continúa utilizando el sistema de autenticación de la Universidad de Granada, lo que permite que las comunicaciones sean seguras. Además, al seleccionar el material, el usuario tiene que elegir el aula donde quiere visualizarlo. En dicha aula habrá un dispositivo TV Stick (se ha utilizado un MK809iii con Android 4.4.2) conectado al proyector y que será el encargado de realizar la proyección, siempre controlada desde el móvil/tableta/portátil del profesor. Además, el profesor debe introducir un código que se proyecta desde el dispositivo, de forma que así se evitan posibles confusiones con dispositivos en otras aulas.

3) **Diseño e implementación de un servicio para la visualización de las transparencias y navegación a través de las mismas**. Desde el aula y a través de cualquiera de los dispositivos soportados, el profesor controlará la proyección avanzando o retrocediendo a través de sus transparencias (véase la Fig. 2). También está disponible la opción de ir a una transparencia concreta, introduciendo su número. Las acciones realizadas por parte del profesor en su móvil/tableta/portátil son enviadas, a través del servidor central, al dispositivo TV Stick conectado al proyector de su aula. Este control se realiza a través de una interfaz amigable usando diferentes páginas web. Estos dispositivos TV Stick están conectados a través de HDMI al proyector. En el caso de la ETSI Informática y de Telecomunicación, muchas aulas contaban con proyectores con entrada HDMI, pero otras sólo disponían de entrada VGA (e.g. proyectores EPSON EB-X7), por lo que se han utilizando además conversores HDMI a VGA, de manera que la solución es válida para cualquier tipo de proyector.

V. ACTIVIDADES REALIZADAS

Las acciones realizadas en este trabajo han estado dirigidas a diseñar e implementar las funcionalidades necesarias para cumplir los objetivos ya comentados. Así, se ha utilizado un **servidor central** que tiene varias funciones:

- Por un lado, sirve para **almacenar todo el material docente**. Este almacenamiento se realiza utilizando un protocolo seguro (FTP sobre SSH), usando el **sistema de autenticación de la Universidad de Granada** basado en SAML (*Security Assertion Markup Language*, véase [1]) para asegurarse de que el usuario es un profesor o un alumno de la Universidad de Granada.
- La interfaz de usuario de las diferentes funcionalidades está implementada a través de diferentes páginas web [2]. El servidor central implementa un **servidor web** (Apache) para servir dichas páginas. Para la ejecución de acciones en la parte del servidor (e.g. cuando actúa como intermediario entre el móvil del profesor y el dispositivo en el proyector) se utilizan scripts PHP. Para la ejecución de acciones en el cliente web (móvil del profesor) se utiliza javascript/jQuery/AJAX. Además, se utilizan bases de datos MySQL en el servidor para las diferentes opciones de configuración que hay que almacenar.
- Todas las comunicaciones son seguras. Para ello, el servidor central implementa un **servidor VPN** al que se conectarán todos los dispositivos conectados a los proyectores, que ejecutarán clientes VPN. La conexión VPN es de tipo IPSec Xauth PSK. El uso de una VPN aporta dos grandes ventajas: 1) la conexión es completamente segura, independientemente del protocolo utilizado y 2) es posible utilizar servidores en la red inalámbrica de la universidad. Nótese que la red eduroam en la UGR emplea direcciones IP privadas y además los cortafuegos no permiten el uso de servidores, algo necesario para este trabajo ya que los dispositivos conectados a los proyectores ejecutan servidores para atender las acciones ordenadas por el profesor, e.g. avanzar una transparencia.

Por otro lado, el dispositivo conectado al proyector es un TV Stick, en concreto un MK809iii. Este modelo presenta varias ventajas. Por un lado, tiene una relación calidad/precio excelente. Se trata de un dispositivo Android (la versión usada es la 4.4.2), con un procesador de 4 núcleos, 2 GB de RAM y 8 GB de memoria interna para

almacenamiento (tanto del firmware como de los datos de usuario y de aplicaciones). Y se puede encontrar por menos de 50 euros.

El uso de Android presenta también numerosas ventajas. Por un lado, existen numerosas librerías y herramientas de desarrollo para este sistema. Entre ellas cabe destacar el uso de la herramienta ADB (Android Debug Bridge), que permite la ejecución de comandos desde un PC remoto (e.g. nuestro servidor central). Por otro lado, cuenta con conexiones inalámbricas Wi-Fi y Bluetooth (la primera utilizada en este trabajo y la segunda se utilizará para algunas mejoras futuras). Además, los principales programas para abrir transparencias están disponibles de forma gratuita. En concreto, en este trabajo se ha utilizado la aplicación PowerPoint de Microsoft (gratuita para Android) y Adobe Reader (también gratuito para Android).

En cuanto al **TV Stick MK809iii**, se ha creado una versión de **firmware** con toda la funcionalidad necesaria para este proyecto. A través de la modificación de los ficheros de configuración necesarios, este firmware implementa las siguientes características:

- Puede ejecutar comandos con permisos de administrador (***root***).
- Se ha habilitado la ejecución de scripts durante el arranque (funcionalidad **init.d**).
- Se ha configurado **ADB** (Android Debug Bridge) para usar conexiones TCP/IP.
- Utiliza un **gestor de pantalla principal** (*home screen*) **muy ligero y personalizable**, lo que ha permitido que el usuario vea una pantalla de presentación y quede oculto que se trata de un dispositivo Android.
- El firmware instala las aplicaciones necesarias al arrancar por primera vez. En concreto, se instalan -entre otras- las aplicaciones Microsoft PowerPoint, Adobe Reader, y un cliente VPN que soporta IPSec Xauth PSK.
- De forma similar, durante el primer arranque también se realizan todas las configuraciones necesarias (e.g. uso de eduroam para la conexión Wi-Fi y configuración del cliente VPN).
- El arranque está personalizado para mostrar el **logo de la Universidad de Granada**, así como la pantalla principal (que muestra además información sobre el proyecto de innovación docente).
- La pantalla principal incluye un *widget* que muestra un número aleatorio cada vez que se arranca el dispositivo, que servirá como **código de verificación** para que el profesor no interfiera con otras aulas por error.
- Incluye un directorio con todos los scripts necesarios para que el servidor central le solicite acciones (e.g. abrir un fichero, cambiar de transparencia, pedirle el número aleatorio para su verificación, etcétera).
- Se incluye la **eliminación automática de *popups*** de diferentes programas y del propio sistema Android, dado que no hay interacción directa con el dispositivo.

Tras realizar una primera versión funcional del proyecto, se probó por parte de los componentes del equipo en sus respectivas asignaturas, identificándose algunos puntos a mejorar. Los más importantes estuvieron relacionados con la usabilidad:

- La versión de Android de Microsoft **PowerPoint no soporta los ficheros con extensión PPT**, sólo soporta PPTX. Para evitar que un profesor no se dé cuenta y suba un fichero PPT (que después no podrá mostrar en el aula), se muestra un aviso indicándole que no puede subir el fichero y que sólo se soportan extensiones PPTX.
- Elementos que **indiquen al usuario que se está llevando a cabo una determinada acción**, especialmente cuando esta acción tarda un cierto tiempo (e.g. la que más tardaba era subir un fichero al dispositivo en el proyector). Para ello se ha incluido una barra de avance (similar a la de los navegadores) para la mayoría de las páginas web desarrolladas, y una animación que indica qué porcentaje del fichero se ha subido al dispositivo.
- **Moverse a una transparencia concreta**. Si bien esto estaba ya implementado para los ficheros PPTX, no lo estaba para los ficheros PDF. Aunque resulta complicado automatizar ciertas acciones en aplicaciones Android desarrolladas por terceros, finalmente se consiguió implementar esta funcionalidad.
- **Recuperación de la sesión de forma rápida ante fallos**. Si el navegador utilizado en el móvil/tableta/PC del profesor se cierra (por descuido del profesor o fallo de la aplicación), resulta tedioso volver a seleccionar el dispositivo, cargar el fichero, abrir la aplicación para su visualización y finalmente volver a la transparencia por la que el docente se había quedado. Esto además se agravaba por dos motivos: 1) cuando no se realiza ninguna transmisión a través de la red VPN, a veces ésta se desconectaba; 2) algunos móviles cierran el navegador (o no mantienen la sesión) cuando el móvil se bloquea (algo que pasa tras algunos segundos si no se utiliza). El primer problema (falta de transmisión a través de la VPN) se resolvió usando dos mecanismos. Por un lado, un script en el dispositivo detecta si la conexión VPN se ha caído (e.g. porque falle la conexión Wi-Fi) y la recupera de forma automática. Por otro lado, una vez que el dispositivo está conectado al servidor VPN, un script realiza un *ping* cada 30 segundos para evitar que pase demasiado tiempo sin ninguna transmisión. Para el segundo problema (cierre accidental o por error del navegador), cuando un usuario vuelve a autenticarse en la plataforma, se comprueba cuál fue el último dispositivo utilizado, el último fichero visualizado y la última transparencia mostrada, dándosele la opción -a través de un *popup*- de volver a utilizarlos.

Éstas y otras actualizaciones han mejorado la usabilidad de esta solución, que se está planteando para ser implantada en la ETSI Informática y de Telecomunicación de la Universidad de Granada.

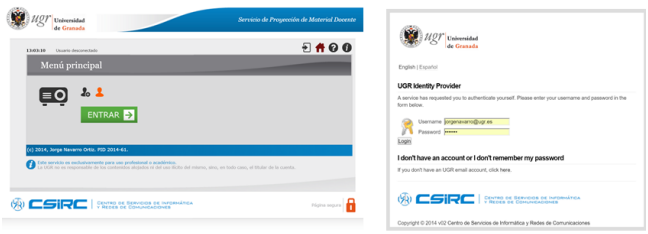


Fig. 3. Autenticación en la web inicial.

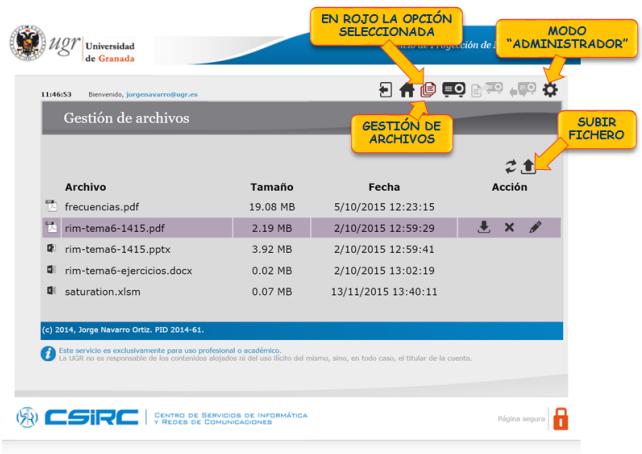


Fig. 4. Web para la gestión de los archivos.

VI. RESULTADOS OBTENIDOS

Como resultado del presente trabajo, se ha obtenido una solución completa al problema que se planteó y que cumple con los objetivos iniciales del proyecto de innovación docente. A continuación se resumen los diferentes resultados.

A. Servicio para el almacenamiento de material docente

Este servidor está terminado y se puede encontrar en [2]. Como se ha comentado, se ha utilizado un servidor seguro usando el protocolo FTP sobre SSH. La Fig. 3 muestra cómo se realiza la autenticación con el sistema de autenticación de la Universidad de Granada. La Fig. 4 muestra cómo se realiza la gestión de los ficheros a través de la web correspondiente. Cabe destacar que los móviles/tabletas iOS/Android pueden subir ficheros desde los principales servicios en la nube (e.g. Google Drive y Dropbox, véase la Fig. 5).

B. Servicio para la selección del material docente

En este -y otros- servicios, se ha implementado una interfaz web accesible desde cualquier dispositivo. Además,

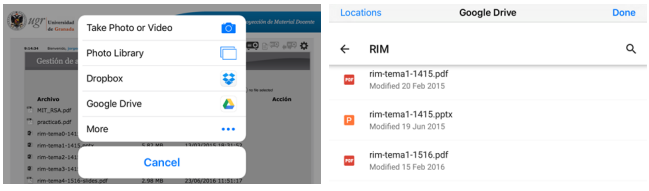


Fig. 5. Ejemplo de interacción con servicios en la nube (Google Drive).

Fig. 6. Selección de material docente.

se ha empleado el diseño web adaptable (responsive web design) para que las páginas web se adapten a las capacidades del dispositivo (principalmente su resolución). La Fig. 6 muestra cómo el docente elige el material que quiere presentar.

C. Servicio para la visualización y navegación a través del material docente

Nuevamente la interfaz web permite que este servicio sea accesible desde cualquier dispositivo. La Fig. 7 muestra cómo se manda el archivo seleccionado anteriormente al dispositivo (aula) seleccionado. En el ejemplo se muestra el nombre del dispositivo "SPMD05", pero en un despliegue real se pondría el nombre del aula en el que está ubicado (e.g. "ETSIIT_aula_1_1"). La Fig. 8 muestra la interfaz principal desde el que el profesor podrá navegar a través de las diferentes transparencias. Como se observa, hay 3 zonas definidas: 1) funciones generales (refrescar presentación, cerrar presentación, cambiar manualmente el modo de presentación, e ir a una transparencia concreta), 2) flechas para navegar por la presentación y 3) zona con la hora actual y algunos botones para uso en caso excepcional (pulsación en la pantalla del proyector, pulsación de las teclas ESCAPE y ENTER) que sólo estarán disponibles durante la fase de pruebas del sistema. Esta última zona permite capturar un pantallazo de lo que se visualiza en el proyector, y simular la pulsación de un dedo sobre dicha pantalla desde el móvil del profesor. Esto permite, entre otras cosas, actualizar el *script* de eliminación de *popups* de manera automática.

D. Servicio de gestión y administración

Todas las acciones realizadas por parte de los usuarios son almacenadas en un registro con una doble funcionalidad. Por un lado, para saber exactamente qué estaba haciendo el usuario en caso de fallo, y así ayudar a encontrar la solución de una manera más sencilla. Y, por otro lado, para poder detectar y tener toda la información de comportamientos fraudulentos, e.g. intentos de conexión por parte de usuarios sin la autorización pertienente.

La página web utilizada para esta administración, además del registro comentado, muestra información de los dispositivos conectados indicando su dirección IP, el código de verificación, la versión de firmware empleada (útil para saber qué dispositivos están completamente actualizados y cuáles no) y una serie de acciones sobre el mismo (apagar, reiniciar, eliminar los archivos de material

Fig. 7. Envío del archivo al dispositivo seleccionado.

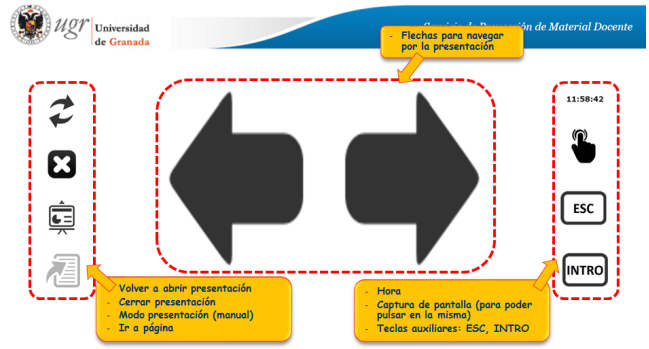


Fig. 8. Interfaz principal de navegación por las transparencias.

docente, cerrar el programa actual, simular la pulsación de las teclas ESCAPE y ENTER, y pulsar las flechas de avance/retroceso de transparencia).

También se incluyen algunas acciones que repercuten sobre toda la plataforma, i.e. sobre todos los dispositivos conectados, como son el reinicio del servicio ADB y el reinicio del servidor de VPN.

E. Solución completa

Las Fig. 9 y 10 muestran un ejemplo de uso de la solución desarrollada. La pantalla de arranque del dispositivo muestra el logo de la Universidad. La Fig. 9 muestra la pantalla principal que se visualiza en el proyector, mientras que la Fig. 10 muestra un ejemplo de navegación a través de las transparencias.

Las instrucciones para utilizar el Servicio de Proyección de Material Docente desarrollado en este PID se pueden consultar en [3].

VII. CONCLUSIONES

En este trabajo se ha realizado el diseño y la implementación de una solución que permite 1) disponer del material docente en cualquier ubicación y 2) poder presentar transparencias de forma inalámbrica, necesitando únicamente que el profesor disponga de un móvil en el aula (también se soportan tabletas y portátiles).

Fig. 9. Pantalla principal visualizada en el dispositivo conectado al proyector.

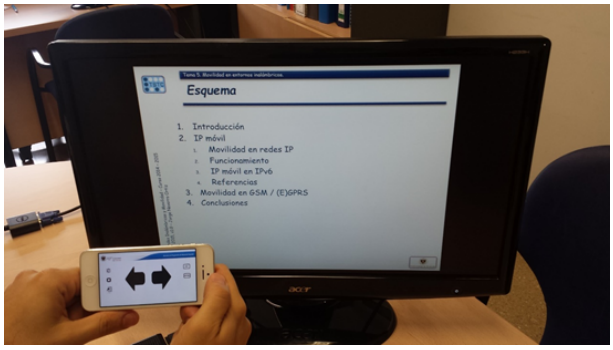


Fig. 10. Ejemplo de control de la navegación por las transparencias.

La valoración global es muy positiva, ya que se han cumplido todos los objetivos y se ha comprobado que facilita y hace más cómoda la tarea de presentar transparencias por parte del profesorado. Además, el Centro de Servicios de Informática y Redes de Comunicaciones (CSIRC) de la UGR está estudiando una posible integración en los sistemas de docencia multimedia desplegados por las aulas de los 7 campus universitarios.

AGRADECIMIENTOS

El presente trabajo ha sido financiado a través del Programa de Innovación y Buenas Prácticas Docentes del Secretariado de Innovación Docente de la Universidad de Granada, Proyecto de Innovación Docente 14-61 "Servicio de Proyección de Material Docente", dentro de la acción 1 (innovación en la gestión *on-line* de los procesos de enseñanza-aprendizaje). Parte del presente trabajo ha sido desarrollado por los alumnos D. Juan Ramón Gutiérrez Martínez, D. Daniel Álvarez González y D. David Gallardo Jiménez, siendo estos dos últimos becarios del citado PID.

REFERENCIAS

[1] Sistema de autenticación de la Universidad de Granada, disponible en http://csirc.ugr.es/informatica/ServiciosWeb/AutenticacionUsuariosUGR.html (accedida el 15 de abril de 2017).
[2] Servicio de Proyección de Material Docente, disponible en https://palas.ugr.es/SPMD/ (accedida el 15 de abril de 2017).
[3] Instrucciones del Servicio de Proyección de Material Docente, disponible en https://palas.ugr.es/SPMD/instrucciones.php (accedida el 15 de abril de 2017).
[4] Google Chromecast, disponible en https://www.google.com/intl/es_es/chromecast/tv/chromecast/ (accedida el 15 de abril de 2017).
[5] Aplicaciones de Google Chromecast, disponibles en https://www.google.es/intl/es_es/chromecast/apps/ (accedida el 15 de abril de 2017).

[6] PowerLite 975W WXGA 3LCD Projector, disponible en https://epson.com/For-Work/Projectors/Classroom/PowerLite-975W-WXGA-3LCD-Projector/p/V11H835020 (accedida el 15 de abril de 2017).
[7] Wireless LAN Module (ELPAP07), disponible en https://epson.com/Accessories/Projector-Accessories/Wireless-LAN-Module-(ELPAP07)/p/V12H418P12 (accedida el 15 de abril de 2017).

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6566

Vídeos cortos realizados por los alumnos como recurso docente. Diferentes enfoques.

Guillermo Azuara Guillén[1], Diego Fernández Iglesias[2], Ana María López Torres[1], Ana María Salinas Baldellou[1], Mª Carmen Aguilar Martín[3], José Luis Salazar Riaño[1], Julián Fernández-Navajas[1], Fidel Cacheda Seijo[2], Francisco Javier Nóvoa de Manuel[2], Víctor Manuel Carneiro Díaz[2].
[1]Departamento de Ingeniería Electrónica y Comunicaciones, [2]Departamento de Tecnologías de la Información y las Comunicaciones, [3]Departamento de Derecho de la Empresa.

[1,3]Universidad de Zaragoza. [2]Universidad de A Coruña.

C/ Atarazanas, 2. 44.003 - Teruel

gazuara@unizar.es, diego.fernandez@udc.es, {lopeztor, salinas, caguilar, jsalazar, navajas}@unizar.es, {fidel,fjnovoa, viccar}@udc.es.

***Resumen*- Este trabajo presenta el desarrollo y análisis de los resultados de un proyecto de innovación docente basado en la creación por parte de los alumnos de vídeos cortos, de 3 o 4 minutos de duración, donde deben exponer algún concepto o tema propuesto por el profesor. En esta actividad los alumnos trabajan las competencias digitales, búsqueda y síntesis de la información, comunicación, trabajo en grupo y mediante el uso de la coevaluación la capacidad de crítica. La experiencia se ha desarrollado de forma conjunta con asignaturas y docentes de diversas universidades, titulaciones y cursos. A partir de unas pautas comunes, se ha adaptado la experiencia al contexto específico de cada asignatura, lo que permite diferentes enfoques.**

***Palabras Clave*- metodologías activas, vídeos generados por los alumnos, competencias digitales, síntesis, coevaluación, Web 2.0.**

I. Introducción

En una sociedad cada vez más globalizada, la cantidad ingente de información que se genera y distribuye hace indispensable desarrollar la capacidad de procesarla y resumirla de la mejor manera posible.

Otra de las nuevas competencias cada vez más demandada en la sociedad contemporánea es la competencia digital, que engloba tanto el manejo en general de las TIC, Tecnologías de la Información y las Comunicaciones, como la propia creación de contenidos digitales.

Con el objetivo principal de trabajar estas competencias y capacidades, junto con otras como la búsqueda de información, la comunicación, la capacidad crítica o el trabajo en grupo, se definió un proyecto de innovación docente. Este proyecto se basa en que los alumnos realicen un vídeo corto, de unos tres minutos de duración (cuatro en algunos casos), donde trabajen un concepto o tema propuestos por los docentes de cada asignatura. Posteriormente los estudiantes deben evaluar los vídeos de sus compañeros. Se les da libertad en la elección del software de edición de vídeo.

Desde el curso 2011/2012, en la asignatura de Redes, de segundo curso del grado de informática de la Universidad de A Coruña, los docentes organizan un concurso de vídeos cortos de conceptos relacionados con la asignatura. De entre los presentados se premian los tres mejores, que reciben tres recompensas diferentes [1]. A partir de esta idea de proponer la realización de vídeos cortos a los alumnos, un grupo de profesores de diversas universidades, titulaciones y cursos, solicitaron en el año 2015 un proyecto de innovación docente donde desarrollar esta idea y ponerla en práctica con diferentes matices en las asignaturas que impartían.

En el siguiente apartado se presentará una breve descripción del estado del arte en la utilización de vídeos realizados por los alumnos como recurso docente, a continuación se explicará cómo se organizó

y se planificó la experiencia. Luego se presentará un breve análisis de la experiencia en diferentes asignaturas y en el último apartado se abordarán las conclusiones globales y el trabajo futuro a desarrollar.

II. VÍDEOS DE LOS ALUMNOS COMO RECURSO DOCENTE.

En la última reforma universitaria en España (Ley Orgánica 6/2001, de 21 de diciembre, de Universidades), se definió a los estudiantes como "protagonistas activos de la actividad universitaria", lo que se traduce en tener la visión del alumno como centro del sistema educativo. En este contexto se hizo fundamental un cambio de modelo educativo, y las metodologías activas pasaron a ser parte esencial del proceso de formación de los estudiantes [2].

En paralelo a este proceso, aunque hace tiempo que las TIC son utilizadas para acompañar al alumno en su aprendizaje (campus virtuales, comunicación por medios telemáticos, video-tutoriales, etc.) en los últimos años han aparecido nuevos fenómenos sociales como la "web 2.0" o "web participativa", en la que el usuario ha dejado de ser un mero consumidor de información y ha pasado a tener la capacidad de crear sus propios contenidos y difundirlos [3].

En esta sociedad donde tienen un peso muy importante las redes sociales, soportadas por la "web 2.0", tenemos varios actores claramente diferenciados.

Por un lado tenemos las personas que han crecido en este escenario, habituados a crear y compartir contenidos, conocidos como "nativos digitales" descritos en [3-5], caracterizados entre otras cosas por estar muy habituados al aprendizaje informal, donde el conocimiento que adquieren está sin estructurar, no es explícito y es difícil de expresar y transmitir [4]. Para este tipo de alumnos la experiencia propuesta les servirá para reforzar su capacidad de síntesis, organización de la información y transmisión ordenada de conceptos.

Otro tipo de alumno que podemos encontrarnos es el denominado "inmigrante digital" [6], caracterizado por no tener incluida en su rutina diaria el uso de las TIC. Esto puede deberse a que por su edad se ha iniciado más tardíamente en el uso de la tecnología, a que por sus condiciones socio-culturales no ha tenido fácil acceso a dispositivos electrónicos, o sencillamente a que no le gusta usar este tipo de tecnologías. Para estos estudiantes la experiencia les permitirá además de potenciar las mismas competencias que a los nativos digitales, mejorar también sus competencias en TIC.

Teniendo en cuenta este contexto, y a partir de la experiencia antes mencionada que sirvió como base [1], se decidió plantear un proyecto de innovación docente basado en que fueran los propios alumnos los que realizaran vídeos cortos y que permitiera trabajar todas las competencias descritas.

En la primera fase de este proyecto se realizó una revisión bibliográfica para conocer el estado del arte, localizando diversas publicaciones donde se presentaban y analizaban diferentes experiencias, con variados enfoques y sobre distintos niveles educativos. A modo de ejemplo, en [7] se describe una experiencia de vídeos generados por los alumnos, más jóvenes que los universitarios (las mayoría de los vídeos realizados por alumnos de primaria y algunos por alumnos de secundaria) y se analiza desde el punto de vista de alumnos, padres, profesores y equipos directivos; en [8], la experiencia fue realizada con alumnos de primaria (10-12 años), y los autores se centraban sobre todo en el papel del profesor en el desarrollo de los trabajos; [9], también realizada con estudiantes de primaria, se centra principalmente en analizar el comportamiento del grupo durante el desarrollo de la actividad; en [10], con estudiantes universitarios, se analiza la posibilidad de utilizar los vídeos producidos por los estudiantes como material docente, además de buscar aumentar la motivación de los alumnos y mejorar sus competencias tecnológicas; en [11] se propone la actividad como entorno de trabajo para que los estudiantes aprendan a contar historias, desarrollando todas las posibilidades narrativas de los vídeos.

En esta primera revisión, donde se encontraron bastantes experiencias de este tipo, la duración de los vídeos no era un factor crítico, y aparecía condicionado principalmente por el número de alumnos o grupos que realizaban la actividad. Por ello, una de las principales aportaciones de este proyecto es la inclusión de la limitación temporal como uno de sus pilares fundamentales. Esto obliga a los alumnos a los alumnos a desarrollar un profundo trabajo de los conceptos y contenidos para extraer la información más relevante y exponerla de manera breve, concisa y atractiva.

En una segunda revisión bibliográfica, llevada a cabo tras la conclusión del primer curso académico en el que pusimos en marcha este proyecto, sí que se detectó que la bibliografía sobre este tipo de experiencias había crecido de forma notable. De esta segunda revisión queremos señalar el estudio y conclusiones de una experiencia con ciertas similitudes a nuestro proyecto, como la descrita en [12].

III. ORGANIZACIÓN Y DESARROLLO DE LA EXPERIENCIA.

En la experiencia se perseguía analizar las potencialidades de la creación de vídeos cortos por parte de los alumnos.

Los objetivos perseguidos eran los siguientes:

- Analizar la influencia de la metodología en las competencias de comunicación, búsqueda de información, trabajo en grupo, síntesis y motivación. Esto implica ver cómo influye la realización de la actividad en estas competencias, o al menos estudiar las condiciones deseables para poder realizar de manera adecuada este examen. La información se obtuvo tanto de los estudiantes como de los profesores.

- Analizar la carga de trabajo que supone para los alumnos y cuantificar la dedicación media para la realización de los vídeos (según información suministrada por los alumnos).

- Detectar las ventajas e inconvenientes de las diferentes formas de orientar la actividad. Como se detallará en el siguiente apartado, el proyecto se planteó para que en cada asignatura se pudiera aplicar de forma adaptada a sus condiciones particulares. Por ello, y derivado de los diferentes planteamientos, se buscaron los puntos fuertes y débiles de cada una de las experiencias previas personales y de las descritas en la bibliografía, para identificar aquellas estrategias que dieran mejores resultados en contextos determinados o por el contrario identificar los aspectos negativos para su eliminación o al menos mitigación, con la consiguiente mejora de la actividad.

- Generar unas pautas para una eficiente utilización de este recurso docente. Desde el inicio se tenía claro que como producto final se debía generar una recopilación de "buenas prácticas" para facilitar su uso en diferentes ámbitos.

En las primeras reuniones de planificación del desarrollo de la actividad, se plantearon diversas formas de abordar la realización de estos vídeos, presentando diversas opciones: modalidad competitiva (concurso) / no competitiva, suministrar o no información específica relacionada con la edición de vídeo (lo que denominamos guiada y no guiada); tamaño del grupo; formación del grupo; implantación en cursos iniciales o finales de la titulación; en asignaturas obligatorias u optativas; publicación de los vídeos en webs, plataformas públicas o en entornos docentes de acceso restringido.

Se decidió que cada profesor tomara las decisiones que le parecieran más adecuadas para cada asignatura, pero se elaboraron una serie de materiales comunes (pautas generales, rúbrica de evaluación y encuestas de valoración de la actividad) para garantizar un grado de homogeneidad que facilitara la extracción de conclusiones. Sin embargo se mantuvo también la diversidad, ya que los grupos de alumnos en los que se realizó la actividad se corresponden con contextos muy diferentes.

IV. Desarrollo.

En este apartado se van a detallar las diferentes características específicas de desarrollo de la actividad en cada una de las asignaturas que han seguido (o están siguiendo) este curso el protocolo completo, mostrando para todas la misma estructura de presentación de la información. En primer lugar se describen las características de la asignatura, luego los detalles de desarrollo de la actividad, la forma de evaluación y finalmente dificultades encontradas, puntos fuertes, aspectos a mejorar y buenas prácticas detectadas.

A. Redes. Grado de Ingeniería Informática. Universidad de A Coruña.

Curso: 2. **Cuatrimestre**: 2. **Créditos**: 6 ECTS.

Tipo de asignatura: Obligatoria.

Número de alumnos matriculados: 2015/16: 223, 2016/17: 257 (este curso todavía está el plazo abierto).

Número de alumnos que han hecho vídeos: 2015/16: 28 (este curso todavía no se ha cerrado la actividad).

Modalidad: Actividad voluntaria y Concurso.

Número de miembros por grupo: Variable. Máximo 5 alumnos por grupo.

Tema del vídeo: el curso pasado se definió el mismo tema para todos. En concreto, se propuso que explicasen el protocolo DNS, que también se ve en la clase de teoría. Este curso, sin embargo, los alumnos disponen de más posibilidades a la hora de escoger el tema (FTP, SMTP, DHCP o DNS).

Duración máxima del vídeo: 3 minutos.

Realización de la actividad: no presencial y no guiada. No se les dio más pautas que el enunciado y la rúbrica. De todos modos, suele haber grupos que piden sugerencias para la realización, sobre todo acerca del contenido a abordar.

Presentación de los vídeos: mixta (Moodle / YouTube). Cuelgan su vídeo en YouTube, y a través de Moodle indican la URL usando la actividad Tarea.

Obligatorio rellenar encuesta de valoración: sí.

Número de horas dedicadas por los alumnos a la realización de la actividad: 0 horas presenciales, 14'5 horas dedicadas de media por vídeo (según los alumnos).

Evaluación: realizada por el profesor y coevaluación. Los profesores (6) valoraron todos los vídeos, mientras que cada alumno que hubiera realizado algún vídeo debía valorar únicamente 2 vídeos de sus compañeros.

Si hay varias evaluaciones de alumnos cómo se calcula la nota: se realizó una media de todas las evaluaciones (tanto procedentes de profesores como de alumnos).

Peso de evaluación profesor/estudiantes: todas valen lo mismo.

Peso en la nota final: 1 punto al mejor grupo, 0'5 al segundo, y 0'25 para el tercero.

B. Física II. Grado de Ingeniería Electrónica y Automática. Universidad de Zaragoza. 2015-2016.

Curso: 1. **Cuatrimestre**: 2. **Créditos**: 6 ECTS.
Tipo de asignatura: Formación Básica.
Número de alumnos matriculados: 25
Número de alumnos que han hecho vídeos: 25 (todos los que cursan la asignatura).
Modalidad: Práctica obligatoria.

Número de miembros por grupo: Variable (4-5).
Tema del vídeo: a escoger entre propuestos (todos ellos sobre funciones del osciloscopio).
Duración máxima del vídeo: 3 minutos.
Realización de la actividad: Guiada.
Presentación de los vídeos: Publicación en YouTube.
Obligatorio rellenar encuesta de valoración: Sí.
Número de horas dedicadas por los alumnos dedicadas a la realización de la actividad: 10 horas.

Evaluación: profesor + coevaluación. Cada vídeo fue valorado por 1 profesor y 4 alumnos.
Peso de evaluación profesor/estudiantes: No se tuvo en cuenta la evaluación de los alumnos.
Peso en la nota final: 5 % de la nota final (es uno de los cuatro trabajos tutelados que se desarrollan en la asignatura).

C. Física II. Grado de Ingeniería Electrónica y Automática. Universidad de Zaragoza. 2016-2017.

Curso: 1. **Cuatrimestre**: 2. **Créditos**: 6 ECTS.
Tipo de asignatura: Formación Básica.
Número de alumnos matriculados: 27
Número de alumnos que han hecho vídeos: 24.
Modalidad: Práctica obligatoria.

Número de miembros por grupo: Variable (2-3).
Tema del vídeo: mismo tema para todos (funcionamiento de una impresora láser).
Duración máxima del vídeo: 4 minutos.
Realización de la actividad: Guiada.
Presentación de los vídeos: Publicación en YouTube.
Obligatorio rellenar encuesta de valoración: Sí.
Número de horas dedicadas por los alumnos dedicadas a la realización de la actividad: 9.2 horas de promedio.

Evaluación: realizada por el profesor y coevaluación. Cada vídeo fue valorado por 1 profesor y 1 ó 2 alumnos.
Peso de evaluación profesor/estudiantes: No se tuvo en cuenta la evaluación de los alumnos.
Peso en la nota final: 5 % de la nota final (es uno de los cuatro trabajos tutelados que se desarrollan en la asignatura).

D. Redes de Computadores. Grado de Ingeniería Informática. Universidad de Zaragoza.

Curso: 2. **Cuatrimestre**: 1. **Créditos**: 6 ECTS.
Tipo de asignatura: Obligatoria.
Número de alumnos matriculados: 2015/16: 27, 2016/17: 31.
Número de alumnos que han hecho vídeos: 2015/16: 17 (63 %), 2016/17: 26 (84 %).
Modalidad: Actividad voluntaria.

Número de miembros por grupo: Variable. Máximo 5 alumnos por grupo.
Tema del vídeo: A escoger entre propuestos (escoger un estándar IEEE 802.3, 802.11, 802.15 y 802.16 o realizar una comparación entre los 4).
Duración máxima del vídeo: 2015/16: 4 minutos (flexibles); 2016/2017: 3 minutos (inflexibles).
Realización de la actividad: no presencial y no guiada.
Presentación de los vídeos: 2015/16: Moodle. 2016/17: mixta (Moodle / YouTube). Cuelgan su vídeo en YouTube, y a través de Moodle indican la URL usando la actividad Tarea. Dos de los grupos pidieron no colgarlo en YouTube y se colgó en Moodle, aunque hubo problemas con el tamaño del vídeo.
Obligatorio rellenar encuesta de valoración: sí.
Número de horas dedicadas por los alumnos a la realización de la actividad: 0 horas presenciales, 14 horas dedicadas de media por vídeo (según los alumnos).
Peso en la nota final: 2015/16: 55% de la nota de un trabajo que valía el 10 % de la parte de teoría (el otro 45 % era el trabajo escrito sobre los mismos temas). 2016/17: 10 % de la parte de teoría.

Evaluación: 2015/16: solo profesor; 2016/17: realizada por el profesor y coevaluación. El profesor (1) valoró todos los vídeos, mientras que cada alumno que hubiera realizado algún vídeo debe valorar todos los demás vídeos de sus compañeros.
Si hay varias evaluaciones de alumnos cómo se calcula la nota: se realizó una media de todas las evaluaciones de los alumnos.
Peso de evaluación profesor/estudiantes: profesor 70 %, alumno 30 %.

V. Análisis

En este apartado se presenta un análisis de los aspectos más destacados de la experiencia, centrándonos en las dificultades detectadas durante el desarrollo de la actividad, los puntos fuertes detectados, los aspectos a mejorar y los aspectos a destacar como buenas prácticas.

Además de las asignaturas presentadas en el punto anterior, también han participado de alguna manera en la experiencia, en el curso anterior, en el presente o en ambos, las asignaturas de la Universidad de Zaragoza: Visión por computador (Ingeniería Electrónica y

Automática), Régimen laboral en la empresa (Administración y Dirección de Empresas) y Comercio electrónico (Ingeniería de Tecnologías y Servicios de Telecomunicación, Ingeniería Informática).

La información presentada en los siguientes subapartados ha sido obtenida en base a los datos recogidos en el desarrollo de las actividades y a opiniones y valoraciones emitidas por los docentes en las reuniones de coordinación y valoración de la actividad. Los datos cuantitativos han sido obtenidos de las encuestas realizadas por los alumnos el curso pasado, ya que en el presente curso no se ha cerrado la actividad en todas las asignaturas, aunque para los aspectos tratados se consideran relevantes. Se presentan los valores medios del conjunto de las tres asignaturas descritas. Para los datos cualitativos se han enriquecido también con las experiencias de este año y las opiniones y valoraciones aportadas por profesores y estudiantes.

A. Dificultades.

Conseguir la motivación del alumnado es importante en cualquier asignatura, y mantenerla no resulta sencillo, especialmente cuando se trata de asignaturas con un gran número de matriculados.

Pocos alumnos tenían experiencia previa realizando vídeos de esta índole (sólo pequeñas piezas audiovisuales) y sólo un 20% había grabado y editado un vídeo con anterioridad.

Aunque la mayoría de los alumnos en la pregunta de qué les había constado más si hacer el vídeo o preparar los contenidos, contestaron que hacer el vídeo, luego a la hora de cuantificar el porcentaje del tiempo total dedicado al trabajo, la mayoría (59 %) señalaron que menos del 40 % del tiempo fue dedicado a la realización del vídeo (de hecho el 17 % dedicó una quinta parte del tiempo o menos y el 39 % entre un 20 y un 40 % del tiempo). Algunos alumnos manifestaron que para ellos la parte más complicada de la actividad había sido la edición del vídeo. No obstante creemos que esta dificultad no fue generalizada, y además fue superada por todos los alumnos.

Algunos de los alumnos también destacaron como dificultad el mostrar los conceptos de forma accesible a todos los públicos (sea cual sea su formación), lo que implicaba: lenguaje sencillo, analogías comprensibles...

En algunas de las experiencias, se habían explicado en clase los contenidos asociados al vídeo y se habían proporcionado referencias para obtener la información. Aun así, algunos de los grupos buscaron en Internet trabajo ya hecho (vídeos, explicaciones…) que luego presentaron de manera inconexa, demostrando que no habían entendido el tema en absoluto. Este problema fue minoritario, y sólo se detectó en dos de los grupos.

B. Puntos fuertes detectados.

Los alumnos han demostrado gran autonomía a la hora de escoger herramientas para edición de vídeo (*Sony Vegas, Powtoon, Kdenlive, Camtasia Studio, Animaker, Windows Movie Maker, Videopad*, etc.).

Han demostrado ser capaces de encontrar y seleccionar la información relevante en los casos en los que no se les proporcionaba la información sobre los conceptos a tratar en el vídeo, lo cual ha implicado organización, coordinación y planificación.

En general, la realización del vídeo en sí no supuso un hándicap, ya que un porcentaje elevado (51%) del alumnado empleó menos del 20% del tiempo de la actividad en aprender cómo se hace y edita un vídeo (ver figura 1). A la vista de los resultados creemos que puede ser uno de los puntos fuertes de la actividad, ya que el uso de esta tecnología parece que no supone un gran sobreesfuerzo a los alumnos. Destacar que estos datos han sido obtenidos en base a los alumnos de carreras de corte técnico, y cuando se tengan los datos, será interesante analizar si sucede lo mismo en otras titulaciones que no pertenezcan a la rama de ingeniería.

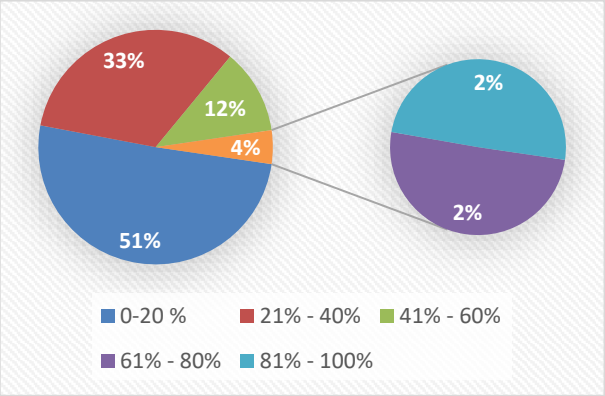


Fig. 1. Porcentaje de tiempo de la actividad dedicado a aprender el manejo de las herramientas de edición de vídeo.

Otro punto fuerte del proyecto es que en general a los alumnos no les resultó especialmente difícil plasmar los contenidos teóricos en el vídeo. En la encuesta, en la cuestión donde se les preguntaba al respecto, en una escala de 1 a 4, donde el 1 era que les resultaba difícil y el 4 fácil, un 79 % respondió con valores 3 ó 4.

Otro punto a destacar es que muy mayoritariamente opinaron que les gustó la actividad. Un 96% de los alumnos manifestaron una satisfacción con la actividad entre 3 y 4 puntos (1 la peor valoración y 4 la mejor), tal y como se indica en la figura 2.

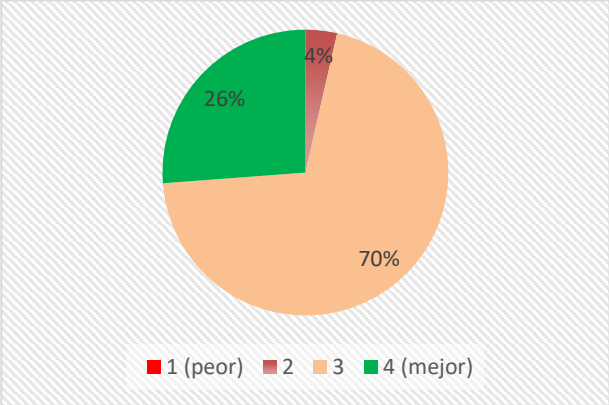


Fig. 2. Valoración de la actividad.

Los alumnos también consideran que se asimilan los conceptos mejor al "meterse más en el tema". Un 79% de los alumnos manifestaron una mejor asimilación de los conceptos que con otras actividades (ver figura 3). En este sentido se espera comprobar esta afirmación en el examen final, con preguntas sobre la temática trabajada y comparándola con resultados de años anteriores.

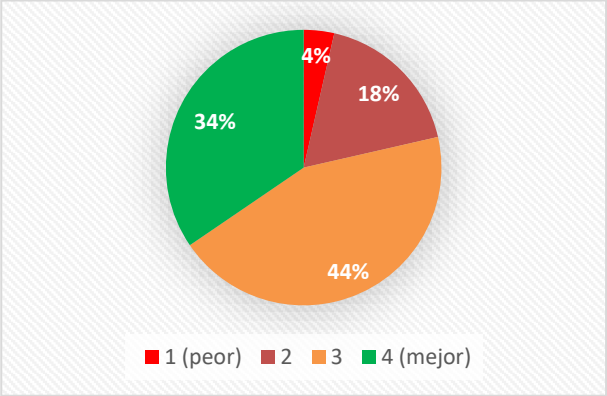


Fig. 3. Valoración sobre la mejor asimilación de conceptos con esta metodología.

Con el desarrollo de la experiencia se ha detectado mayor implicación y motivación por parte del alumnado participante, al tratarse de una actividad que ellos consideran entretenida y distinta. Los docentes perciben que los estudiantes disfrutan con este tipo de actividad, como también se recoge en [13].

C. Aspectos a mejorar en el próximo curso.

A partir del análisis de las experiencias realizadas, se detectaron una serie de aspectos a modificar de cara a la mejora general de la experiencia. En este subapartado se reseñan los más importantes.

- Proponer más de un tema para la realización del vídeo.
- Intentar minimizar el grado de subjetividad a la hora de realizar la evaluación del vídeo. Ésta queda patente cuando comparamos las distintas calificaciones otorgadas sobre un mismo vídeo basándonos en una misma rúbrica. Afinar un poco más la rúbrica y ser algo más exigentes y concretos en el planteamiento de algunos de los criterios.
- Intentar animar más a los alumnos a su participación en la actividad (en el caso de que la actividad se plantee como voluntaria).
- Utilizar los vídeos presentados como material en el aula para próximos cursos. Para ello se debería permitir al alumno modificar el video presentado, para aplicar las mejoras propuestas por docentes y compañeros, lo cual implicaría una dedicación de más tiempo al trabajo. No obstante, de esta forma se podrían reforzar competencias transversales.
- Se considera que la presencia de coevaluación en la actividad mejora la experiencia. Por ello las asignaturas que todavía no lo han hecho, creen interesante considerar la nota de la coevaluación en la nota del trabajo.
- Algunos docentes no tienen claro si los alumnos revisan su trabajo después de la evaluación. Por ello se propone probar con la auto-evaluación para que los alumnos reflexionen sobre el resultado de su tarea, o en su defecto hacer la coevaluación antes de la entrega para poder mejorar los vídeos con las indicaciones de sus compañeros. Con la inclusión de esta metodología se espera que mejore la calidad final del trabajo de los estudiantes.
- Se detectó que cuando el tema no está bien definido, los alumnos tienen problemas para centrar el trabajo, por lo que se considera muy importante definir perfectamente los conceptos a tratar.
- Proponer la actividad hacia el final de la asignatura parece que puede ser un problema, por la acumulación de carga de trabajo. Se propone realizar la actividad hacia la mitad del curso, o al inicio si es posible.

D. Aspectos a destacar como buenas prácticas.

Una programación didáctica que incluya actividades distintas a lo habitual suele tener buena acogida por los alumnos, y suele incrementar la motivación. De hecho, en una de las experiencias, sólo una persona ha suspendido la asignatura tras las convocatorias de junio y julio habiendo presentado el vídeo, lo cual supone menos del 4 % de los participantes. La asignatura, por su parte, en el curso 2015/2016 tuvo un 27'8 % de suspensos (sobre alumnos matriculados).

Mediante la elaboración del vídeo se trabajan competencias transversales, se incentiva la

socialización y la integración, se estimula la creatividad, etc.

Además, la coevaluación ha demostrado que las calificaciones propuestas por el alumnado no distan de las proporcionadas por los profesores (tienden a ser algo inferiores), cuando están basadas en una misma rúbrica. Esta actividad se ajusta bien a este tipo de evaluación. Al realizar la coevaluación los estudiantes deben ver los vídeos del resto de sus compañeros (o al menos varios vídeos), con lo que consiguen aprender de los errores y si la temática es variada adquirir nuevos conocimientos. La utilización de la actividad Taller de Moodle facilita el proceso de coevaluación. El empleo de la rúbrica ha ayudado a que tanto alumnos como profesores tengan más claro qué se va a evaluar si lo comparamos con las experiencias en cursos anteriores.

En ciertas experiencias se ha detectado que los alumnos están acostumbrados a realizar pequeños vídeos en etapas anteriores de educación, por lo que realizar la actividad no les supone un esfuerzo mucho mayor que el de realizar un trabajo al uso y rompen la monotonía.

La limitación de tiempo les hace aprender a elegir las ideas fundamentales que quieren presentar y les obliga a realizar una exposición oral, sin dedicar parte de las horas lectivas del aula.

También se cree importante realizar esta actividad en un periodo de "tranquilidad" en el desarrollo del curso, evitando fases de sobrecarga de trabajo.

El tiempo total consumido para completar la actividad ha sido de menos de 8 horas en el 67 % de los casos, lo que hace que la actividad no sobrecargue especialmente a los alumnos y tenga un fácil encaje en la planificación docente de la asignatura.

En una de las experiencias se comentó a los alumnos que podían extenderse un poco más de tiempo si era indispensable. Esto propició que se llegaran a presentar vídeos de hasta 14 minutos (sobre un tiempo máximo de 4 minutos), lo que desvirtuaba totalmente la actividad. Por ello se considera muy importante ser inflexible con la duración máxima de los vídeos.

Por último, se ha detectado que mostrar un vídeo de ejemplo antes de empezar la actividad (eran los vídeos ganadores del año anterior del concurso de la asignatura de Redes de la Universidad de A Coruña, y el mejor vídeo de la asignatura de Redes de la Universidad de Zaragoza), puede servir de inspiración y modelo, y hacer mejorar mucho la calidad de realización y originalidad de los vídeos.

E. Valoración por asignaturas.

En este apartado, se muestra la valoración de la actividad en cada una de las asignaturas que han seguido el protocolo completo.

Redes de computadores (U. de A Coruña). 2015/16 y 2016/17: A pesar de que la participación en la actividad no superó el 13% de los alumnos matriculados, los resultados han sido muy positivos, teniendo en cuenta la calidad de los vídeos entregados, las opiniones proporcionadas por los alumnos y, en último término, las calificaciones conseguidas por los alumnos participantes. Al tratarse de una disciplina técnica, los alumnos han necesitado invertir poco tiempo en su familiarización con las herramientas de edición de vídeo, lo cual es muy deseable en esta actividad, donde el manejo de las herramientas no es un objetivo principal. Asimismo, los vídeos han permitido detectar ciertos errores de concepto que se han podido corregir a tiempo (antes del examen teórico).

Física I. 2015/16: La actividad fue del agrado de los alumnos y se consiguieron los objetivos establecidos, tanto de forma (tiempo) como de contenido.

Física II. 2016/17: Dentro de la asignatura sirve a los estudiantes para profundizar en un aspecto que no se trabaja en clase. Además, es beneficioso realizar actividades diferentes (no todo controles o trabajos escritos) para incrementar la motivación.

Redes de Computadores (U. de Zaragoza). 2015/16 y 2016/17: Tanto el profesor como los alumnos valoran la actividad de forma muy positiva. Los alumnos reconocen que les cuesta un esfuerzo, tanto la búsqueda y tratamiento de la información como la planificación, realización y postproducción del vídeo, pero todos consideran que se ve compensado. También destacar el importante salto de calidad de un curso al siguiente, probablemente asociado entre otras cosas a una mejor planificación y exposición de la actividad a los alumnos y a mostrarles vídeos de ejemplo.

VI. CONCLUSIONES

La experiencia ha cumplido las expectativas que se tenían en ella cuando se propuso, y el continuar por segundo año consecutivo nos permite ir aumentando los datos que tenemos y seguir mejorándola.

Los alumnos también se muestran satisfechos y creen que entienden mejor los conceptos trabajados de esta forma, no obstante, se sigue trabajando en esta línea para tener datos cuantitativos que soporten esta hipótesis.

Las reuniones periódicas que mantienen los docentes permiten tener puntos de vista muy amplios y diversos, y permite ir reflexionando sobre la experiencia antes de trasladarla a los alumnos, durante su desarrollo y a su conclusión. Las aportaciones del conjunto enriquecen notablemente el análisis del desarrollo del proyecto.

Se consideran valiosos los puntos de mejora que se han detectado y sobre todo las buenas prácticas, que podrían ser de gran utilidad para los docentes que estén interesados en poner en marcha experiencias similares.

De cara al futuro esperamos que se incorporen de manera completa al proyecto asignaturas de titulaciones que no sean de la rama de ingeniería, para poder comparar los resultados con otros escenarios. También

se contempla la posibilidad de utilizar los mejores vídeos como material docente para cursos futuros.

AGRADECIMIENTOS

Este proyecto ha sido parcialmente financiado por la convocatoria de innovación docente de la Universidad de Zaragoza para los cursos 2015/16 y 2016/17 (Proyectos PIIDUZ_15_411 y PIIDUZ_16_070), Ministerio de Economía y Competitividad de España (Proyecto TIN2015-70648-P y TIN2015-64770-R), Fondo Social Europeo y Gobierno de Aragón (grupo de investigación reconocido T98).

REFERENCIAS

[1] D. Fernandez *et al*, "Gamificación en el aula universitaria: Un caso práctico en una asignatura de redes," in *XII Jornadas De Ingeniería Telemática*, Palma de Mallorca, 2015, pp. 426-432.

[2] A. Fernandez, "Metodologías activas para la formación en competencias," *Educatio Siglo XXI*, vol. 24, pp. 35-56, 2006.

[3] J. Vassileva, "Toward Social Learning Environments," *IEEE Trans. Learn. Technol.*, vol. 1, pp. 199-214, 2008.

[4] M. E. Sousa-Vieira *et al*, "Aprendizaje Social y Gamificación en una Asaignatura de Redes de Ordenadores," *Actas De Las Jornadas De Ingeniería Telemática 2013*, pp. 509-513, 2013.

[5] J. Cross, *Informal Learning: Rediscovering the Natural Pathways that Inspire Innovation and Performance.* John Wiley & Sons, 2011.

[6] Q. Wang, M. D. Myers and D. Sundaram, "Digital natives and digital immigrants: Towards a model of digital fluency," *Busin. Info. Sys. Eng.*, vol. 5, pp. 409-419, 2013.

[7] L. Palmgren-Neuvonen, M. Jaakkola and R. -. Korkeamäki, "School-context videos in Janus-faced online publicity: Learner-Generated Digital Video Production Going Online," *Scan. J. Educ. Res.*, vol. 59, pp. 255-274, 2015.

[8] L. Palmgren-Neuvonen and R. -. Korkeamäki, "Teacher as an orchestrator of collaborative planning in learner-generated video production," *Learn. Cult. Soc. Interact.*, vol. 7, pp. 1-11, 2015.

[9] L. Palmgren-Neuvonen and R. -. Korkeamäki, "Group interaction of primary-aged students in the context of a learner-generated digital video production," *Learn. Cult. Soc. Interact.*, vol. 3, pp. 1-14, 2014.

[10] B. Ryan, "A walk down the red carpet: Students as producers of digital video-based knowledge," *Int. J. Technol. Enhanced Learn.*, vol. 5, pp. 24-41, 2013.

[11] M. Kearney, "A learning design for student-generated digital storytelling," *Learn. Media Technol.*, vol. 36, pp. 169-188, 2011.

[12] C. Orús *et al*, "The effects of learner-generated videos for YouTube on learning outcomes and satisfaction," *Comput. Educ.*, vol. 95, pp. 254-269, 2016.

[13] J. Pirhonen and P. Rasi, "Student-generated instructional videos facilitate learning through positive emotions," *J. Biol. Educ.*, pp. 1-13, 2016.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6579

Desarrollo de un laboratorio abierto de enjambres de robots autónomos de limpieza

Laura Pozueco, José Antonio Sánchez, Alejandro G. Tuero, David Melendi, Roberto García, Xabiel G. Pañeda, Noemí Asenjo, Oscar Quintana, Javier Viñuela, Pedro B. López, Adrián Santinho

Departamento de Informática

Universidad de Oviedo

2.7.4, Edificio Polivalente, Campus de Xixón, s/n, Asturies, España

pozuecolaura@uniovi.es, sanchezsjose@uniovi.es, garciatalejandro@uniovi.es, melendi@uniovi.es, garciaroberto@uniovi.es, xabiel@uniovi.es, UO232277@uniovi.es, UO187253@uniovi.es, UO217178@uniovi.es, UO83450@uniovi.es, UO212296@uniovi.es

***Resumen*- En el corto plazo, la industria del robot autónomo y de los drones será un factor de desarrollo importante. Tanto la fabricación de los propios dispositivos, como del desarrollo del software que los hace funcionar serán actividades empresariales de importancia en el sector TIC. No obstante, el futuro de los robots de trabajo autónomo es colaborar unos con otros y con el entorno que los rodea (adaptación al contexto). Para ello, deben de ser capaces de interactuar con otros equipos mediante protocolos de comunicación y sistemas de razonamiento. Por ello, en este trabajo se presenta una experiencia encaminada a la obtención de un sistema de trabajo colaborativo para que un grupo/enjambre de robots autónomos de limpieza puedan trabajar de forma conjunta. La experiencia se basa en un laboratorio abierto que permite a los alumnos proponer y realizar sus propios desarrollos. En el trabajo se presentan tanto los aspectos metodológicos de la experiencia, como los avances que se han conseguido realizar hasta la fecha.**

***Palabras Clave*- robots autónomos, sistema colaborativo, project oriented learning, laboratorios abiertos**

I. Introducción

Los robots de limpieza suelen trabajar de forma individual. Mediante unos sensores, se desplazan por una superficie que puede tener pequeños obstáculos. Internamente, disponen de una serie de algoritmos que les permiten maximizar el área de limpieza. No obstante, cuando se plantea su despliegue para cubrir una gran superficie, puede llegar a ser necesario utilizar varios de estos robots. En una situación ideal, los robots serían capaces de colaborar unos con otros con la finalidad de repartirse la superficie a recorrer o trabajar por turnos. El funcionamiento coordinado en forma de enjambre, no solo les permite ser muy eficientes, sino que también viene acompañado de otras ventajas. Algunas pueden ser la fiabilidad, debido a la redundancia, o la capacidad de disponer de un sistema con altas capacidades de cómputo, gracias a la paralelización [1] No obstante, hasta donde los autores pueden saber, no existen productos comerciales de limpieza capaces de operar en forma de enjambre.

Con la finalidad de disponer de robots que operen de forma coordinada, el grupo de investigación DMMS de la Universidad de Oviedo ha diseñado un laboratorio que se centra en el desarrollo de un sistema de trabajo colaborativo para robots autónomos de limpieza comerciales. El desarrollo de este sistema parte de la necesidad de ser capaces de ejercer un control sobre el funcionamiento de los robots que vaya más allá de funcionamiento convencional. Adicionalmente, es necesario incorporar funciones básicas de comunicación. Idealmente, estas comunicaciones serán inalámbricas, utilizando mecanismos ad-hoc y/o elementos de infraestructura. Además, es necesario crear un protocolo de comunicaciones de alto nivel. Este protocolo, sobre la infraestructura inalámbrica subyacente, permitiría el intercambio de mensajes necesario para llevar a cabo el trabajo colaborativo. Finalmente, deberá crearse un sistema de razonamiento que les permita analizar los mensajes recibidos y la información capturada de un conjunto de sensores, para

reaccionar en consecuencia. Todo esto permitiría la implementación de aplicaciones finales.

Toda la experiencia se ha enmarcado en el paradigma de los espacios de trabajo abierto, siguiendo el planteamiento de [2], y el paradigma del *project oriented learning* (POL). De esta forma, se permite a los estudiantes utilizar los materiales del laboratorio para desarrollar sus propios proyectos. La única restricción es que todos deberán tener un objetivo general definido: conseguir que lo que diseñen limpie de la forma más eficiente posible. Cada estudiante o equipo podrá trabajar definiendo su contexto de limpieza, tipo de local o zona, si requieren comunicaciones entre los robots, sensores, inteligencia, etc. De esta forma, el laboratorio del proyecto se ha configurado como un entorno de aprendizaje cooperativo entre los estudiantes, en el que proyectos anteriores podrán ser utilizados como base de conocimiento para los nuevos diseños. Así se definirá un proceso evolutivo dentro del propio laboratorio en el que generación tras generación de proyectos se evolucione la tecnología, experimentando los alumnos la situación real de los desarrollos tecnológicos en la industria. Los aspectos pedagógicos del laboratorio se describen en [3].

El resto del artículo se ha estructurado de la siguiente forma. En la Sección II se describen los aspectos metodológicos del proyecto. En la Sección III se presenta el sistema de base que ha sido utilizado. En la Sección IV se describen los distintos proyectos que han sido completados hasta la fecha. Finalmente, la Sección V presenta las conclusiones y algunos trabajos futuros.

II. Aspectos Metodológicos del Laboratorio

La metodología seguida en el proyecto se ha basado en ciclos evolutivos de creación y evaluación de prototipos centrados en robots autónomos. Es decir, se van construyendo nuevos proyectos tomando como partida los proyectos que han sido desarrollados con anterioridad. La idea es la de crear un prototipo inicial o sistema base, sobre el cuál se vayan aplicando diferentes funcionalidades para conseguir una evolución del mismo. Esta evolución puede ir encaminada en la mejora del intercambio de información entre los robots, en el desarrollo de métodos de comunicación o en la obtención de un sistema final.

Desde un punto de vista pedagógico, los alumnos desarrollan sus proyectos partiendo del trabajo de otros alumnos. Por ello, la única restricción que tienen es la de construir módulos que puedan ser reaprovechables. En este sentido, es muy importante el diseño del software desarrollado y la calidad de la documentación generada por los alumnos. Esta forma de trabajar, se ilustra en la Fig. 1.

Además de un correcto diseño y de una buena documentación, los prototipos desarrollados por los alumnos deberían pasar un pequeño control de calidad. Tras las pruebas realizadas en un entorno de desarrollo convencional, se realizarían pruebas de campo en un laboratorio diseñado al efecto. El laboratorio es un aula vacía en la que se han instalado dos cámaras que permiten la monitorización remota de los experimentos, y su potencial control mediante técnicas de visión por computador. Una cámara fija obtiene un plano general del laboratorio, y la cámara robotizada cuyo interfaz se muestra en la Fig. 2 realiza el seguimiento del robot.

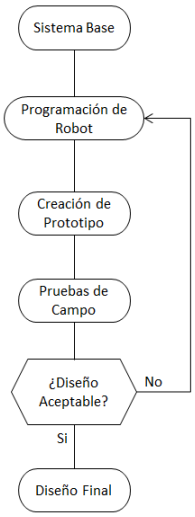


Fig. 1. Flujo de tareas del proyecto.

Para evitar problemas con el seguimiento del robot por parte de la cámara robotizada, se han instalado vinilos en las ventanas delo laboratorio. Antes de la instalación, la cámara detectaba movimiento fuera del laboratorio y se perdía el rastro del robot. Adicionalmente, las pruebas siempre se realizan bajo las mismas condiciones de iluminación para evitar problemas con el tratamiento de las imágenes capturadas con la cámara.

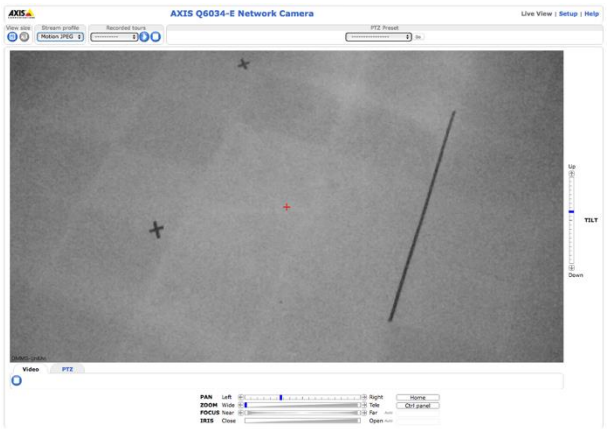


Fig. 2. Interfaz de usuario de la cámara robotizada instalada en el laboratorio de experimentación.

III. Descripción del Sistema Base

El laboratorio ha utilizado como punto de partida el kit de desarrollo que la empresa iRobot (spin-off del MIT) dentro de su programa *Create 2* [4]. Este programa se basa en los robots que fabrica esta

empresa, dentro de la serie Roomba. El kit incluye el robot de prototipado que se muestra en la Fig. 3, y permite la incorporación de nuevo hardware con su software correspondiente. No en vano, los robots Roomba han sido utilizados en entornos educativos desde hace tiempo [5], por tratarse de un recurso de bajo coste aplicable en el aprendizaje y la investigación en el campo de la robótica.

Fig. 3. iRobot *Create 2*.

Dado que el robot Roomba funciona de forma autónoma e independiente, lograr el trabajo cooperativo implica dotar a este dispositivo de algún tipo de elemento capaz de ejecutar programas y de permitir comunicaciones inalámbricas entre los robots, o con algún tipo de infraestructura. Por lo tanto, los inicios del proyecto consistieron en el establecimiento de un sistema base que contemplase estos requisitos. Para ello, seleccionamos equipos Rapsberry Pi debido a su flexibilidad, a su coste reducido y al hecho de que disponen de comunicaciones inalámbricas (integradas en las versiones más recientes y a través de un adaptador USB en las versiones anteriores). Estos aspectos han sido considerados para usos similares en trabajos anteriores como [6].

La interconexión entre el robot y la Rapsberry Pi se realiza mediante comunicaciones serie. A través de una conexión de este tipo, la Rapsberry Pi puede acceder a la información de los sensores del robot y enviarle comandos para que haga algo (por ejemplo, moverse en una dirección concreta). Esta conexión se realiza utilizando un cable de comunicación serie-USB. El cable se proporciona, dentro del proyecto *Create 2*, junto con el robot.

Durante el desarrollo del proyecto, también descubrimos que el interfaz serie presente en el robot incorporado en el kit del programa *Create 2* también está presente en otros robots del mismo fabricante. Esto nos permitió disponer de dispositivos adicionales, más sencillos de adquirir. En particular, se consiguieron varios robots Roomba modelo 631. Para acceder al interfaz serie, es necesario efectuar un orificio en la cubierta del robot.

Por otro lado, los robots disponen de una batería que les permite trabajar de forma autónoma. No obstante, surge el problema de alimentar eléctricamente la Rapsberry Pi en movilidad. Por ello, y dado que el cable provisto en el kit no posee sistema de alimentación, ha sido necesario emplear baterías externas portables que se conectan a la entrada MicroUSB de las Raspberry Pi. El prototipo completo se muestra en la Fig. 4.

IV. Proyectos del Laboratorio

En esta sección se describen algunos trabajos que han sido desarrollados en el marco del laboratorio abierto, siguiendo el marco metodológico expuesto con anterioridad. En particular, han participado alumnos de Grado en Ingeniería, en las disciplinas Industrial/Mecánica, de las Telecomunicaciones y de la Informática.

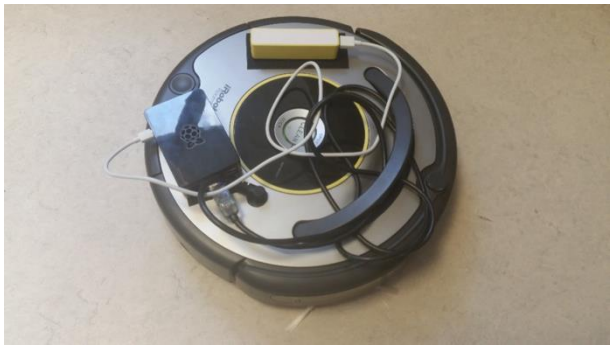


Fig. 4. Prototipo desarrollado.

A. *Desarrollo de un interfaz de control de robots*

Más allá de los aspectos físicos del sistema base, era necesario disponer de algún mecanismo que permitiese el control del robot a través de comunicaciones serie. En los primeros prototipos, este control se efectuaba mediante un programa sencillo que se ejecutaba en la Rapsberry Pi. Este programa se basaba en la información indicada en la especificación del robot.

La especificación del robot incluye el conjunto de instrucciones que permiten acceder a los valores de los sensores y provocar el desplazamiento del robot [7]. Su funcionamiento se basa en el modelo de estados que se muestra en la Fig. 5.

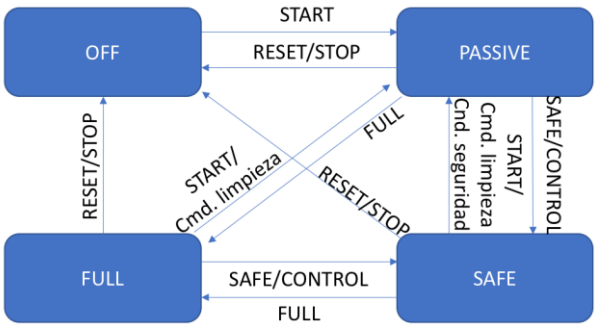


Fig. 5. Modelo de estados del interface del robot.

Los estados se describen a continuación:

- OFF: En este modo la interfaz de entrada no está activa hasta que se envía un comando START, tras el cual ya es posible el envío de otros comandos al robot. Por otro lado, en este modo el robot es capaz de enviar información de los eventos que puede detectar, como una pulsación

del botón CLEAN, la separación del robot del suelo, etc.

- PASSIVE: En este modo se entra tras el envío de un comando START o cualquiera de los comandos de limpieza disponibles en la interfaz (CLEAN, SPOT, etc.). No obstante, no es posible cambiar los parámetros de los comandos de actuación (MOTORS, DRIVE, etc.). Para ello es necesario hacer una transición a los modos SAFE o FULL.
- SAFE: En este modo se entra tras un comando SAFE, que proporciona un control total sobre el robot, a excepción de algunas condiciones de seguridad. Algunas de estas condiciones son la detección de algún desnivel durante el movimiento hacia delante o hacia atrás, la detección de las ruedas levantadas o la conexión con la estación de carga. Si ocurre alguna de estas condiciones, el robot realiza una transición al modo PASSIVE.
- FULL: En este modo se entra tras la recepción de un comando FULL, que proporciona un control completo del robot sin ningún tipo de restricción.

Teniendo en cuenta el funcionamiento del robot, se hicieron pruebas sencillas de funcionamiento como las siguientes:

- Con el envío de comandos de actuación al robot se comprobó que se movía hacia delante o atrás, que giraba hacia un lado u otro, etc.
- Con los comandos de limpieza se confirmó que limpiaba, encendía los motores de aspiración, etc.
- Con la recepción de información de los múltiples sensores hubo que estudiar cada uno en particular para interpretar correctamente la información.

Una vez completadas las pruebas con la aplicación de prototipado, la experiencia conseguida permitió el desarrollo de una librería de programación reutilizable. Esta librería encapsulaba todas las funciones de control del robot en unas clases, con la idea de poder ser utilizadas en programas planteados con un nivel de abstracción mayor.

B. Desarrollo de un subsistema de comunicaciones

Adicionalmente, para permitir el trabajo cooperativo, se desarrollaron mecanismos de comunicación basados en redes inalámbricas ad-hoc. Los dispositivos Rapsberry Pi instalados en los robots, son capaces de comunicarse entre sí e intercambiar información sin necesidad de ningún tipo de infraestructura inalámbrica. Además, para permitir la comunicación efectiva entre las aplicaciones a desarrollar, se ha definido un protocolo de comunicaciones de alto nivel. Respetando la filosofía de laboratorio, el diseño se ha planteado de forma que se obtenga un middleware reutilizable que permita abstraer a los desarrolladores de detalles relacionados con la comunicación, como en [8].

En nuestro caso, el protocolo se basa en documentos XML y datagramas UDP y permite intercambiar información entre los dispositivos respecto a diferentes situaciones que se pueden producir. El protocolo contempla dos tipos de mensajes:

- NetMsg: se utilizan por los procesos de mantenimiento y configuración de la red, pero en ningún caso se proporcionan a la capa de aplicación. Un ejemplo se muestra en la Fig. 6.
- RooMsg: Contiene información de lo que se desea enviar al resto de robots de la red.
- Ack: Son mensajes de confirmación de mensajes RooMsg
- Err: Son mensajes de error en el procesamiento de algún mensaje RooMsg

```
<?xml version="1.0" encoding="UTF-8"?>
<NetMsg len="41" cksum="123" id="42c1304f1">
    <source>0000000042c1304f</source>
    <msg>HI</msg>
</NetMsg>
```

Fig. 6. Ejemplo de mensaje NetMsg.

C. Sistema de inteligencia del robot

El dispositivo Rapsberry Pi también se ha dotado de un programa que hace las funciones de cerebro del sistema. El funcionamiento de este programa se basa en el concepto de "situación". Una situación es una herramienta que permite modelar un escenario determinado en un documento XML. Está formada por uno o varios sensores, una o varias acciones y una o varias señales que deben enviarse al resto de robots. El administrador puede configurar estas situaciones para implementar escenarios concretos. Un ejemplo de situación se muestra en la Fig. 7.

```
<situacion id="robo">

            <sensor name="RIGHT_WHEEL">1</sensor>
            <sensor name="LEFT_WHEEL">1</sensor>

            <externa>LED 0,0,0,1,128,128</externa>

            <actuacion tipo="SONG">40,50</actuacion>

</situacion>
```

Fig. 7. Ejemplo de situación.

El diseño de alto nivel del sistema de inteligencia se muestra en la Fig. 8.

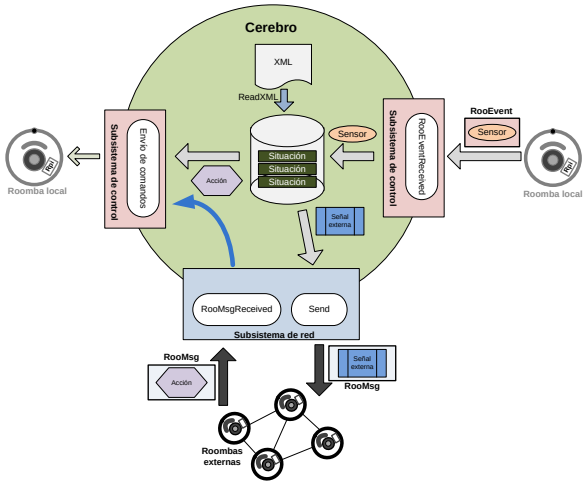


Fig. 8. Cerebro del sistema.

Básicamente, se examina el comportamiento del robot en el que se ejecuta la aplicación, y se generan eventos que contienen la información leída de los sensores. Estos eventos se comparan con la información de la base de datos local, de forma que si alguno de los valores de un sensor, o sensores, coincide con lo establecido en una situación determinada, se obtienen las acciones y señales externas. Las primeras se ejecutan en el robot y las segundas se envían al resto de robots Roomba. En caso de que se reciba un mensaje de otro robot, éste contiene la acción que debe ser realizada.

La combinación de los tres primeros proyectos, es la base para desarrollar nuevas aplicaciones colaborativas. La filosofía es la misma que la de los casos anteriores. Los desarrollos se construyen de forma incremental y se encapsulan en librerías de programación para futuros proyectos.

D. Banco de pruebas para sistemas Roomba

Los robots utilizados en el proyecto se desplazan gracias a una pareja de ruedas motrices que giran de forma independiente. Este diseño permite al robot hacer giros de 360 grados sin necesidad de desplazarse. Debido a este diseño, fue necesario desarrollar un banco de pruebas que permitiese operar el robot en un espacio reducido (por ejemplo, encima de una mesa). Este banco de pruebas consistiría en un sistema de rodillos que permitiría fijar el robot en una posición estable.

El primer prototipo desarrollado para la ejecución de las pruebas se muestra en la Fig. 9. Para implementar el prototipo, se ha partido de un banco de rodillos comercial. Se han aprovechado la parte superior del banco y los rodillos. Como se puede ver en la figura, se cortaron los carriles laterales y se agujerearon de nuevo para introducir los rodillos con la distancia adecuada para las ruedas de los robots. Las piezas fueron montadas sobre una tabla de madera, que actuaba como bastidor.

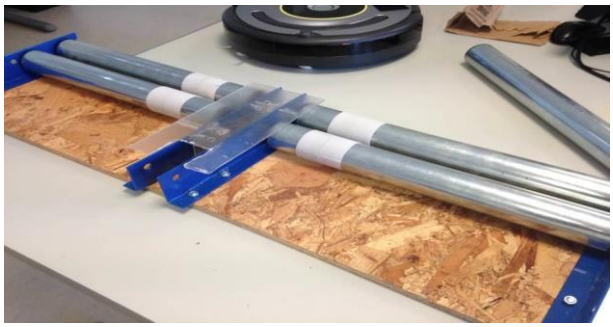


Fig. 9. Primera versión del banco de pruebas.

Tras las primeras pruebas con el prototipo, quedó de manifiesto que el banco de ensayos podía establecer unas condiciones de funcionamiento muy similares a la realidad. No obstante, se constató que se producían vibraciones en el robot. Los problemas se debían al diseño del soporte central del robot, a la excentricidad de los rodillos y a la falta de rodamientos. Por todo lo anterior, se decidió realizar un nuevo diseño más elaborado. El resultado final se muestra en la Fig. 10.

El banco se complementa con una tapa de acero dotada de una goma sobre la que se deposita el robot. Adicionalmente, tiene dos aberturas que permiten depositar las ruedas del robot sobre los rodillos. Adicionalmente, las instrucciones de desplazamiento y de giro que se dan al robot se pueden controlar realizando mediciones sobre los rodillos. Para ello, el banco se ha equipado con una placa Arduino conectada a un circuito con sensores de efecto Hall. En concreto, para cada pareja de rodillos se instaló un imán y 3 sensores separados 180º.

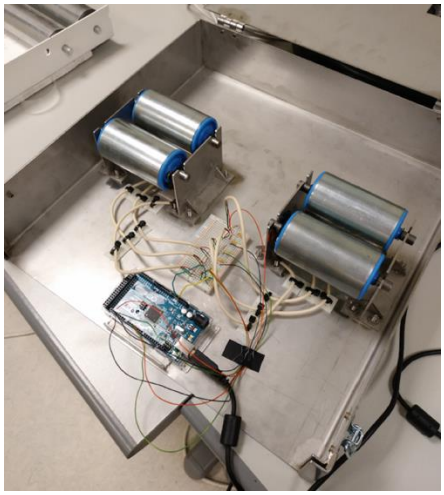


Fig. 10. Banco de pruebas desarrollado.

E. Aplicación RooBra

Esta aplicación es un sencillo sistema de alarma que se activa cuando un robot se separa del suelo. Su funcionamiento se puede observar en la Fig. 11. Cuando la lectura de los sensores del robot indica que éste equipo se ha separado del suelo, se genera una alarma que se propaga hacia el resto de elementos de la red. La propagación se realiza mediante comunicaciones ad-hoc utilizando el paradigma *store*,

carry and forward. Cuando los robots reciben este mensaje de alarma, simplemente encienden un led de color rojo.

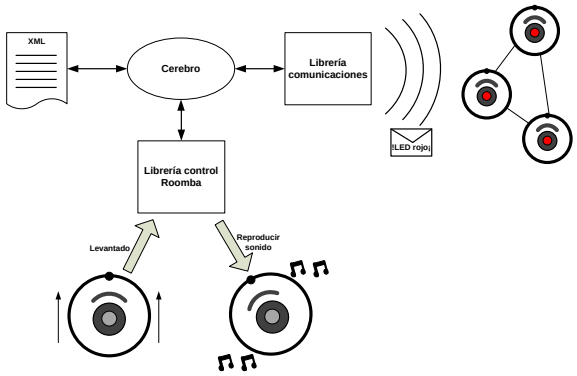


Fig. 11. Aplicación RooBra.

F. *Sistema de carga colaborativa*

Los robots están equipados con una batería que es necesario recargar de forma periódica. Cuando un robot detecta que su batería está a punto de vaciarse, deja el trabajo de limpieza para dirigirse a un punto de recarga. El problema surge cuando hay varios robots en una misma zona, que comparten una única estación de recarga. Para permitir una utilización solidaria del punto de recarga, esta aplicación permite que varios robots se coordinen. El comportamiento normal del robot es el de permanecer en la estación de carga mientras no se encuentra limpiando. En este contexto sería imposible que varios robots pudiesen usar la misma estación. Esta aplicación permite liberar la estación de carga para que sea utilizada por el robot que tiene menos capacidad almacenada. Un escenario de los producidos en esta aplicación se muestra en la Fig. 12.

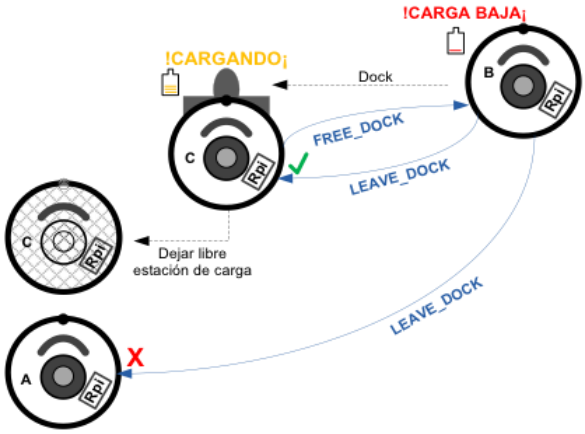


Fig. 12. Aplicación de carga colaborativa.

G. *Sistema de trabajo dirigido*

Esta aplicación permite dar instrucciones a los robots en función de las imágenes que se toman desde una cámara. Utilizando técnicas de visión por computador y las imágenes capturadas por las cámaras del laboratorio, la aplicación es capaz de detectar puntos de suciedad, tal y como se observa en la Fig. 13. Adicionalmente, la aplicación es capaz de ubicar al robot, dentro de la zona observada. De esta forma, el sistema de trabajo dirigido es capaz de proporcionar instrucciones de desplazamiento al robot, relativas a su posición actual. De esta forma, se consigue dirigir al robot para limpiar las zonas en las que se ha detectado suciedad.

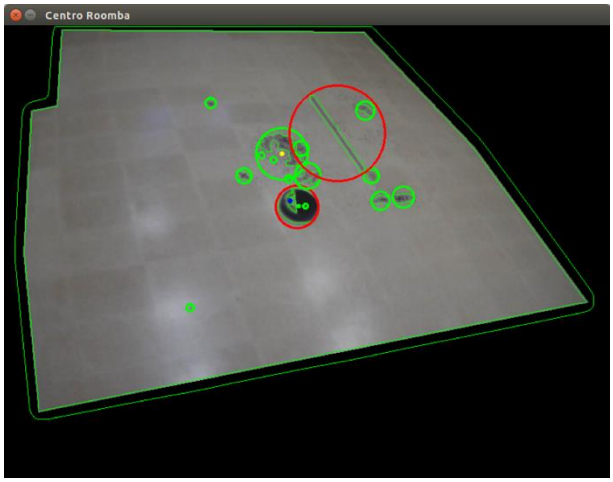


Fig. 13. Sistema de trabajo dirigido.

H. *Sistema de control basado en Arduino*

En este proyecto, se pretende obtener un sistema equivalente al sistema básico, pero utilizando una placa Arduino. Estos dispositivos son un alternativa a las Rapsberry Pi. En concreto, se ha utilizado una placa Arduino UNO.

Tal y como se observa en la Fig. 14, la implementación de la comunicación serie con el robot se ha realizado a través de los pines de entrada salida de la placa, dejando libre el puerto USB para otros usos. Para las tareas de comunicaciones inalámbricas se emplea un chip ESP8266. En este momento se ha implementado un módulo para el control de robot y se han realizado pruebas de concepto del protocolo de comunicaciones, pero no se han iniciado los trabajos de desarrollo del subsistema de inteligencia.

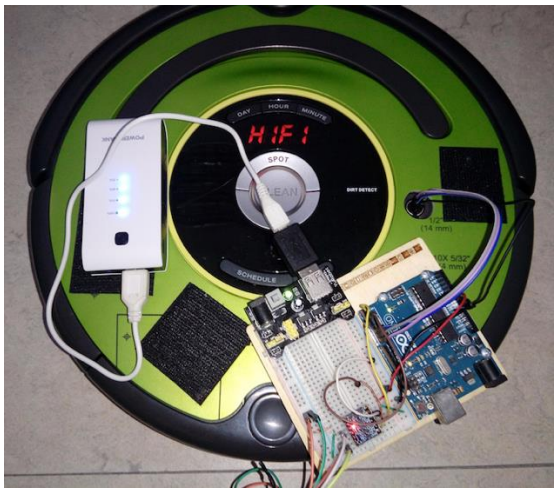


Fig. 14. Prototipo de sistema de control basado en Arduino.

V. CONCLUSIONES

En este artículo se han presentado distintos aspectos y el estado de evolución del proyecto de desarrollo de un laboratorio abierto de enjambres de robots autónomos de limpieza, emprendido por el grupo de investigación DMMS de la Universidad de Oviedo.

En un proceso de evolución continua, se han llevado a cabo distintos proyectos encaminados a la obtención de un sistema colaborativo de robots. Se han creado mecanismos de control de los robots y funciones de comunicaciones que permiten el intercambio de información entre los miembros del enjambre de robots. Por otro lado, los robots son capaces de procesar la información que recogen de sus sensores y combinarla con los mensajes que reciben de otros robots para tomar decisiones.

El enfoque pedagógico que se ha dado al proyecto tampoco está ausente de algunos aspectos innovadores. La concepción del laboratorio de base como un espacio abierto de trabajo, permite a los estudiantes disponer de material y equipos sobre los que implementar sus propios diseños. Al mismo tiempo que se enriquece el proyecto, se crea un espacio para la innovación. Todo esto ha permitido la defensa de distintos trabajos fin de estudios, tanto a nivel de grado como de máster.

El sistema de inteligencia que se ha implementado en los robots, les permite tomar decisiones previstas con anterioridad. No obstante, nos gustaría dotar al sistema de mecanismos de aprendizaje automático que le permitan mejorar sus resultados. Esto podríamos enmarcarlo dentro de los trabajos futuros que se plantea el grupo de investigación. La ubicación en interiores también es algo que necesita cierto desarrollo. Ser capaces de ubicar los distintos robots que operan en un área concreta, nos permitiría distribuir el trabajo de limpieza de forma más eficiente. Finalmente, sería necesario diseñar un sistema que permita alimentar las Rapsberry Pi desde la propia batería del robot Roomba, evitando la necesidad de instalar una batería adicional para esta función.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Instituto Universitario de Tecnología Industrial de Asturias (IUTA) a través del proyecto "Desarrollo de un sistema de trabajo colaborativo de robots autónomos de limpieza".

REFERENCIAS

[1] Gerardo Beni, “From Swarm Intelligence to Swarm Robotics”, en Lecture Notes in Computer Science, vol 3342. Springer, 2005.

[2] University of Dayton, “Open Lab: University of Dayton, Ohio” [En línea]. Disponible en: https://www.udayton.edu/artssciences/about/casit/labs/index.php [Accedido por última vez: 10-may-2017].

[3] José Antonio Sánchez, Laura Pozueco, Alejandro G. Tuero, Noemí Asenjo, David Melendi, Roberto García, Xabiel G. Pañeda y Gabriel D. Orueta, “Laboratorio abierto para el desarrollo de proyectos con robots de limpieza autónomos” En actas del Congreso Tecnología, Aprendizaje y Enseñanza de la Electrónica, pp. 363-368, Sevilla, 2016.

[4] iRobot, “iRobot-Create 2” [En línea] Disponible en: http://www.irobot.com/About-iRobot/STEM/Create-2.aspx [Accedido por última vez: 10-may-2017].

[5] Ben Tribelhorn y Zachary Dodds, “Evaluating the Roomba: A low-cost, ubiquitous platform for robotics research and education”, en actas del IEEE International Conference on Robotics and Automation, Roma, 2007.

[6] Viseslav Celan, Ivo Stancic y Josip Music, “Cleaning up Smart Cities - Localization of Semi-Autonomous Floor Scrubber”, en actas del International Multidisciplinary Conference on Computer and Energy Science (SpliTech), Split, Croacia, 2016.

[7] iRobot, “iRobot Create 2 Open Interface (OI)”. 2015.

[8] Danilo H. F. Menezes, Marco T. Chella y Hendrik T. Macedo, “A Client/Server Message Oriented Middleware for Mobile Robots”, en la revista Journal of Software, vol. 7, no. 5, 2012.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6590

Evaluación Abierta y Transparente en Tiempo Real de Asignaturas de Ingeniería Telemática

Elsa Macías, Alvaro Suárez
Departamento de Ingeniería Telemática
Universidad de Las Palmas de Gran Canaria (ULPGC)
Edificio de Electrónica y Telecomunicación – Campus universitario de Tafira – 35017 Las Palmas de G.C.
elsa.macias@ulpgc.es, alvaro.suarez@ulpgc.es

***Resumen*- Existe una cantidad enorme de métodos de evaluación de la docencia universitaria a día de hoy (tan grande como la cantidad de enseñantes). Las asignaturas de Ingeniería Telemática en España tienen distintas metodologías docentes y métodos de evaluación, pero en su mayoría tienen docencia de teoría, problemas en el aula y prácticas de laboratorio. Cada una de esas metodologías tiene sus propios métodos de evaluación particulares. Normalmente, se suelen marcar exámenes de evaluación de competencias que se realizan en una fecha determinada; elaboración de problemas o trabajos teóricos a desarrollar en un plazo prefijado; trabajos en grupo… que se evaluán después del plazo de entrega. Para este tipo de evaluación no conocemos ninguna iniciativa en la que tanto los profesores como los alumnos puedan recibir realimentación en tiempo real de la evaluación que se está realizando por parte de los profesores y comentarios de los otros alumnos. En este artículo planteamos un sistema de evaluación que permite que los alumnos y profesores puedan tener información, en tiempo real sobre la evaluación que se está realizando puntualmente. A partir de nuestra experiencia práctica, afirmamos que este sistema de evaluación ha demostrado ser muy efectivo y ahorra tiempo de revisión, a la vez que aumenta la calidad de las tareas y evaluaciones realizadas.**

***Palabras Clave*- técnicas de evaluación individual y grupal, modelos de evaluación continua**

I. INTRODUCCIÓN

Internet y la Web han influenciado enormemente la docencia en los últimos años. Desde el inicio de los modelos básicos de teleenseñanza, hasta los modernos modelos de enseñanza abiertos (*open learning*) [1] o aprendizaje sofisticado (*smart learning*) [2] pasando por los entornos de aprendizaje personalizados (*Personalized Learning Environment* (*PLE*)) [3] se han visto nacer, crecer y morir una gran cantidad de metodologías y plataformas diferentes. Destaca de todas estas tendencias la enseñanza abierta cuyo objetivo es que las universiades abran sus cursos a cualquier persona. En este modelo se enmarcan los cursos masivos ofertados por entidades y empresas que no son empresas nativas de educación como por ejemplo Movistar [4]. Cada uno de estos modelos y plataformas suelen llevar asociado uno o varios modelos de evaluación en los que se hacen pruebas en un momento determinado o se da un plazo de entrega de tareas y después se evalúa de forma personalizada al alumno. Por otro lado, es usual que la evaluación de los trabajos a realizar se hagan en evaluación por pares en tiempo diferido (la corrección se hace entre alumnos). Estos modelos de evaluación producen un elevado número de quejas que se producen al no existir una verificación de los trabajos que se entregan ni tampoco de una supervisión por parte de los profesores del curso.

El *learning analytics* [5] se suele usar como parte del modelo de aprendizaje sofisticado para recabar información del alumno. Junto con técnicas de procesado de datos masivo (*Big Data*) [6] puede servir para mejorar la experiencia posterior del alumno en nuevos cursos relacionados. Hasta donde alcanza nuestro conocimiento, no se suele informar al alumno ni a otros profesores del curso del conocimiento obtenido del alumno para mejorar la calidad del aprendizaje ni de la evaluación. Y mucho menos se hace esta evaluación personalizadamente ni en tiempo real.

En este artículo presentamos una experiencia (modelo) en el marco de la evaluación del apendizaje [7] para asignaturas presenciales del área de conocimiento de Ingeniería Telemática en España, que hemos venido utilizando en los últimos años en asignaturas de master y doctorado. También se ha utilizado en cursos on line. Hasta donde alcanza nuestro conocimiento no se conoce una experiencia similar. Esta experiencia está enmarcada en nuestra trayectoria de propuestas sobre enseñanza abierta en un contexto global y no solamente abriendo los contenidos de la enseñanza a cualquier persona [8]. La idea básica es que los profesores de una asignatura evalúan de forma abierta y transparente a los alumnos (el resto de profesores puede acceder a esas evaluaciones en tiempo real); pero también los alumnos pueden acceder a esas evaluaciones. Así cualquier alumno puede mejorar su tarea en tiempo real. Este proceso se lleva a cabo dentro de un plazo marcado inicialmente por los profesores de acuerdo con los alumnos. Los resultados obtenidos son muy alentadores puesto que fomenta una elaboración de trabajos de forma competitiva (logrando una calidad superior a los trabajos que se realizan sin este tipo de evaluación), y además se logra una evaluación más completa que la tradicional y de mayor calidad fomentando la competitividad sana entre los alumnos y profesores. Aunque el tiempo invertido en la evaluación por parte de los profesores es mayor que en la tradicional; este método de evaluación permite un ahorro de tiempo considerable en la fase de revisión de las evaluaciones que se reduce a cero prácticamente.

La estructura de este artículo es la siguiente: en la sección II se presenta el marco de aplicación de este nuevo método de evaluación, la idea básica, sus ventajas e inconvenientes. En la sección III se presenta la aplicación de este método a asignaturas de nuestro entorno académico y en la sección IV se presentan las principales conclusiones.

II. Evaluación del aprendizaje abierta, transparente y en tiempo real

En este apartado presentamos, en primer lugar el contexto de aplicación de nuestro método, después revisamos brevemente las ideas básicas del modelo de evaluación tradicional, a continuación revisamos brevemente las ideas principales de nuestro nuevo modelo de evaluación del aprendizaje.

A. Metodologías docentes de las asignaturas de Ingeniería Telemática

En [9] se presentan algunas metodologías adaptadas al Espacio de Educación Superior Europeo. Muchas universidades en España, basándose en este trabajo han elaborado un catálogo de metodologías docentes y de evaluación para sus estudios. En el caso de la *Universidad de Las Palmas de Gran Canaria* (*ULPGC*), se estipula qué metodologías docentes y evaluación puede llevarse a cabo en sus estudios [10]. Las metodologías docentes recogidas en este Reglamento que se suelen aplicar a las materias de Ingeniería Telemática se recogen en los documentos VERIFICA de los títulos que se imparten en la *Escuela de Ingeniería de Telecomunicación y Electrónica* (*EITE*) [11], que por comodidad se resumen brevemente:

a) *Modalidades de enseñanza presencial*:
- *Teórica*: clase teórica, seminario, taller-trabajo en grupo, clase teórica de problemas o casos, evaluación.
- *Práctica*: laboratorio (grupos pequeños o medianos) y evaluación,
- *Común*: Tutoría (grupos pequeños o medianos).

b) *Modalidades de trabajo autónomo del alumno*:
- *Trabajos teóricos*.
- *Estudio teórico*.
- *Trabajos prácticos*.
- *Estudio práctico*.
- *Actividades complementarias*.

En cada universidad también existe reglamentación específica sobre el tipo de evaluación y actividades evaluables que se pueden llevar a cabo. En el caso de la ULPGC se especifica en el artículo 14 del Reglamento de Evaluación de los Resultados de Aprendizaje y de las Competencias adquiridas por el alumnado en los Títulos Oficiales, Títulos Propios y de Formación Continua de la ULPGC [12], el tipo de actividades evaluables, de ellas nuestro método se puede aplicar especialmente para:

a) *Actividades de evaluación con soporte virtual.*

b) *Actividades virtuales*.

c) *Presentaciones o exposiciones individuales o en grupo*.

d) *Trabajos individuales o en grupo* (*el uso fraudulento del trabajo de otros como si se tratara del de uno mismo y con la intención de aprovecharlo en beneficio propio acarreará las responsabilidades previstas en el artículo 30 del presente Reglamento*).

Destacar que *la valoración de los trabajos encargados al estudiante, de forma individual o grupal, estarán orientados a la comprobación de las competencias adquiridas por los estudiantes*.

Nuestro método se puede aplicar tanto para evaluación individual como para evaluación en grupo.

En el artículo 18 se regula el tipo de evaluaciones posibles:

a) *Pruebas o exámenes escritos*.

b) *Pruebas o exámenes orales*.

c) *Actividades de laboratorio, clínicas o de campo, prácticas, seminarios o talleres*.

d) *Trabajos, para cuya ejecución será necesario la previa determinación por el profesorado de las condiciones de realización, de la exposición y de la puntuación que se otorgue, respetando lo aprobado en el proyecto docente*.

e) *Prácticas externas*.

f) *Proyecto o Trabajos Fin titulación*, y

g) *Otras actividades que se detallen en el Proyecto Docente*.

Nuestro método no se puede aplicar para la evaluación de tipo: *b* (porque al ser oral no hay posibilidad de realimentación durante un tiempo), *e* (porque cada empresa trata a los alumnos de manera específica) ni *f* (porque sólo hay cooperación entre el tutor y el alumno generalmente).

Finalmente subrayar que el mencionado Reglamento prevé que:

a) *El estudiante tiene derecho a solicitar sus resultados en toda prueba, trabajo o examen realizado, de acuerdo con el sistema de evaluación previamente establecido en el proyecto docente de la asignatura*.

b) *Los trabajos y memorias de prácticas, una vez calificados, se devolverán a los interesados siempre que lo soliciten por escrito*.

c) *Junto con los resultados de las evaluaciones, los profesores deberán hacer público el horario, el lugar y la fecha en que tendrá lugar su revisión, con un mínimo de dos sesiones que no podrán coincidir en el mismo día*, y

d) *Las calificaciones finales pueden ser objeto de reclamación por los estudiantes*.

Uno de los objetivos de nuestro método es precisamene ahorrar tiempo del profesor y el alumno en estos menesteres.

B. El método de evaluación tradicional

En la Fig. 1 se esquematiza el proceso tradicional de evaluación de actividades:

a) El profesor coordinador (de acuerdo con el Proyecto Docente de la asignatura), avisa a los alumnos que tienen un plazo determinado para realizar una actividad docente, establece los requisitos para realizarla, recuerda los criterios de evaluación y marca las pautas para el qué y cómo se debe hacer.

b) El alumno realiza la actividad y la envía a su profesor para que éste la evalúe. Normalmente el alumno suele procastinar [13] y la envía en el límite del plazo, aunque existan aplicaciones móviles para evitarlo [14].

c) El profesor evalúa las actividades de los alumnos (individualmente si la actividad es individual) o a cada grupo por separado (si la actividad es grupal). Normalmente, en nuestra experiencia, los alumnos no acceden a las actividades del resto (incluso habiendo trabajado en grupo) porque se reparten las tareas y las hacen de forma independiente, y procastinan y entregan las actividades justo en el plazo marcado.

d) El profesor informa de las notas obtenidas por cada uno de los alumnos; pero la evaluación de cada actividad sólo la conoce cada alumno por separado (o en grupo; pero no entre grupos). Algunos profesores también procastinan y cuando el alumno recibe la evaluación ya ha pasado el tiempo suficiente como para que el alumno pierda el interés en la evaluación o incluso ya ni se acuerde de los contenidos evaluados.

e) Se establece un plazo reclamaciones para respetar la reglamentación pertinente.

Este proceso lleva su tiempo y si es largo puede desmotivar al alumno en su proceso enseñanza-aprendizaje, y al profesor por verlo como una carga de trabajo desmotivante [15].

C. Ideas básicas del nuevo método de evaluación

El proceso de evaluación tradicional es directo, en una sola pasada, no permite realimentación directa al alumno y tampoco colectiva.

El modelo de manejo de procesos que nos ha inducido a emplear nuestro nuevo modelo de evaluación son las técnicas ágiles de desarrollo de proyectos [16]. Porque por un lado, a día de hoy el desarrollo de proyectos se lleva a cabo con participación directa de los clientes (en nuestro caso los profesores), atendiendo a los cambios (razonables) de análisis de requisitos que hacen los clientes y aceptan los desarrolladores (alumnos en nuestro caso). Por otro lado, esta técnica supone un entrenamiento para el desarrollo profesional del futuro ingeniero o doctor.

El objetivo principal es que toda la información de evaluación (y el proceso de evaluación en sí mismo) fluya de manera transparente y en tiempo real entre profesores y alumnos y se minimicen las reclamaciones finales de los alumnos.

El método que hemos venido usando en los dos últimos años consiste en (Fig. 2):

Fig. 1. Método tradicional de evaluación al final del plazo de entrega de tareas marcadas por el profesor

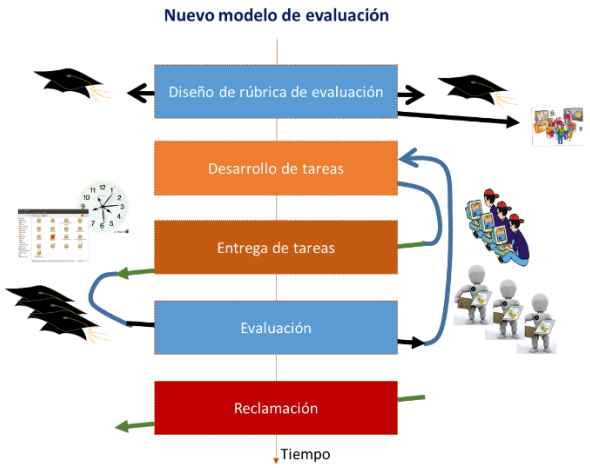


Fig. 2. Nuevo método de evaluación durante el plazo de entrega de tareas marcadas por el profesor

a) Los profesores de la asignatura establecen una rúbrica [17] para establecer criterios comunes de corrección y que se da a conocer a los alumnos.

b) Se lleva a cabo la fase (a) del proceso tradicional, con una variación: los profesores evalúan la actividad cada vez que detecten cambios significativos (que incluyan cierta envergadura a considerar por parte del profesor), o a petición del alumno o simplemente una vez cada cierto número de días. El objetivo es que no haya una única evaluación al finalizar el plazo de entrega.

c) Las actividades se depositan en un lugar conocido por todos los alumnos (por ejemplo una carpeta compartida en La Nube si se trata de la memoria escrita o software), y todos los profesores.

d) Se implanta un sistema de versiones para controlar las correcciones que han ido haciendo cada uno de los profesores (visibles al resto de profesores y a todos los alumnos). De esta manera cada alumno puede recibir sugerencias o ayudas por parte de otros alumnos, para que mejore sus trabajos en función de las correcciones que observa del profesor y su punto de vista de cómo mejorar la actividad. La calificación final premiará a aquellos alumnos que realicen sugerencias o ayudas a los demás.

e) Se implanta un registro personalizado de mejoras realizadas por el alumno en respuesta a las correcciones de los profesores y a comentarios de sus compañeros. Este registro debe estar anotado con valores de tiempo que indiquen cuando se han hecho esas mejoras.

f) En base a esas anotaciones de tiempo se premiará a quellos alumnos que hayan realizado las mejoras antes que otros y también a los que hayan aconsejado mejoras (en especial a aquellos que hayan aportado información de revistas, libros o páginas Web que ayuden a mejorar la actividad). También se hace pública la información relativa a cuando los profesores han hecho sus correcciones y por supuesto qué correcciones han propuesto.

De esta manera se evita la procastinación de los alumnos, porque de manera implícita, aquel que no realice mejoras rápidamente va a ir viendo reducida su capacidad de superar la evaluación al ir obteniendo notas muy bajas (apartado e). También logra que el profesor realice de forma rápida sus evaluaciones para que los alumnos tengan la mayor probabilidad de obtener buena nota y mantener su "buen hacer" frente al resto de profesores y alumnos (que le evalúan en las encuestas oficiales). Esto es, se fomenta una competencia sana entre profesores y entre alumnos.

También se permite mejorar la calidad de las actividades realizadas por los alumnos puesto que todos pueden ver las que realiza el resto. Y se fomenta que los alumnos cooperen entre ellos porque reciben mejor calificación si se ayudan entre ellos (apartado d). Un efecto colateral es que todos los alumnos tienen información de las actividades realizadas por todos, y no sólo de la que ellos han realizado.

Se consigue una normalización de la forma de corregir que no es posible conseguir con una rúbrica que sólo indica qué y cómo se debe evaluar; pero no permite observar cuál es la evaluación realizada. Además, todos los profesores pueden observar en tiempo real cuales son las correcciones que están haciendo el resto de profesores así como los comentarios de alumnos. De esta forma se puede minimizar la cantidad de reclamaciones que se suelen recibir por parte de los alumnos. Además, como los alumnos han tenido información en tiempo real de sus

evaluaciones pueden ir observando cómo se les ha ido corrigiendo (a todos), minimizando sospechas de que a otros se les haya tratado de forma diferente en la evaluación. Esto también minimiza las reclamaciones. En cualquier caso, si existieran reclamaciones más allá del ámbito de actuación del profesor coordinador (a la *Comisión de Asesoramiento Docente* o a instancias superiores), toda la información compatida se haría accesible a esas instancias para que tengan información detallada de lo ocurrido. Es decir, el sistema permite por sí mismo la auditoría detallada de la actuación de profesores y alumnos en el proceso de evaluación. Por último destacar que esta información quedaría visible a los evaluadores de las titulaciones de la *Agencia Nacional de Evaluación de la Calidad y Acreditación* (*ANECA*), como evidencias de la evaluación realizada.

Este método permite emplear técnicas de learning analytics, y aplicarla con una diferencia fundamental: todos los datos que se pueden recoger de la evaluación están disponibles a todos los profesores y alumnos en tiempo real (en realidad cualquiera de ellos podría emplear las técnicas de learning analytics porque los datos están compartidos).

El mayor inconveniente de este método es que el alumno puede verse forzado a responder de forma inmediata a las correcciones y sugerencias de mejora de sus compañeros, empleando un tiempo del que no dispondría. Por ese motivo, se aconseja relajar el criterio de tiempo de respuesta, consensuando con los alumnos los plazos de tiempo en los que se puede responder (sin ser penalizados) a las correcciones y sugerencias. Dejando libertad al alumno para que lo haga o no según su criterio y disponibilidad temporal.

Otro inconveniente importante de este método es la proporción de alumnos por profesor. Si esa proporción es elevada (más de 10) podría darse problemas para manejar los tiempos de respuesta a las actualizaciones de las actividades por parte de los alumnos.

Otro detalle importante es la labor de motivación que debe llevar a cabo el profesor para crear un ambiente de colaboración e interés por los trabajos realizados por los alumnos (entre los profesores y entre los alumnos). Sin esta motivación esta técnica es bastante difícil que de buenos resultados. Por ello se aconseja que se utlice técnicas de motivación e incluso técnicas de gamificación [18] conducentes a mejorar este aspecto parcial de la evaluación.

III. APLICACIÓN DEL MÉTODO EN ASIGNATURAS DE INGENIERÍA TELEMÁTICA

Nosotros utilizamos la potencia y simplicidad de Google Drive para implantar nuestro sistema. En aquellas universidades en las que es obligatorio usar las cuentas de correo electrónico para la comunicación con ellos se puede sincronizar estas cuentas para permitir el uso (a través de ellas del Google Drive). Para aquelas universidades que tengan suscrito acuerdos con otras plataformas en La Nube (*Office 365* u otras) este método también es implementable. Este sistema también se puede usar en plataformas como *Moodle* y otros similares; pero su manejo sería bastante más tedioso tanto por parte del alumno como del profesor.

Nosotros hemos aplicado este método de evaluación en: a) La asignatura *Tecnologías de Internet de Nueva Generación* del *Master de Ingeniería de Telecomunicación* (*MUIT*) [19] de la EITE durante los 2 últimos cursos. b) El *Seminario de Introducción a la Investigación* (en los 2 últimos cursos) del Programa de doctorado *Empresa, Internet y Tecnologías de las Comunicaciones* (*EmiTIC*) [20]. c) El módulo *Despliegue de redes, protocolos y servicios telemáticos en la Empresa 2.0* (en este curso) del *Seminario de investigación específica* del *EmiTIC*). En la Fig. 3 se muestran las distintas carpetas (se han puesto nombres no legibles fácilmente deliberadamente) de Google Drive compartidas para la asignatura, el seminario y el módulo en este curso académico. El uso de esta carpeta de Google Drive es muy simple tanto para el profesor como para el alumno. Configurar el acceso compartido a esas carpetas, por parte del profesor coordinador se puede llevar a cabo, de forma muy simple en minutos.

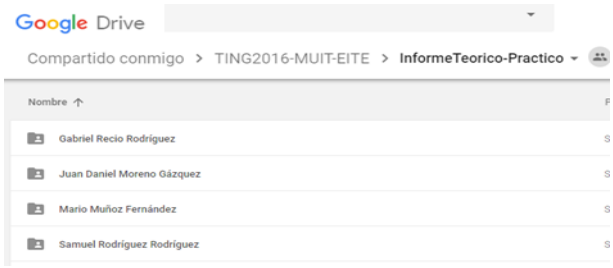
En esas carpetas, compartidas por todos los profesores y alumnos, el alumno deposita las memorias de sus tareas. Cada alumno tiene su propia carpeta (si la tarea es compleja, sino basta con un simple archivo). Cada alumno, y todos los profesores, tienen derecho a escritura en su propia carpeta y derecho a comentar en las carpetas de los otros alumnos. De esta forma se evitan posibles escrituras o modificaciones accidentales que provocan malestar entre los alumnos. Dentro de su carpeta el alumno debe ir actualizando la memoria de su tarea.

Automáticamene, Google Drive actualiza información de las modificaciones realizadas en ese archivo por parte de cualquier persona que pueda editarlo o comentarlo. Además activa marcas de tiempo de cuando se realizaron esos cambios. También permite realizar comentarios (sugerencias de los alumnos o del profesor). Esos comentarios se pueden resolver o se pueden contestar, dando lugar a conversaciones sobre cada uno de los comentarios. Un detalle importante es que Google drive permite enviar un correo electrónico por cada comentario que se haga, con lo cual se instruye a los alumnos que lo mejor es que se hagan conversaciones de comentarios para que sea posible tener un registro inmediato de cada vez que se haga un comentario en el correo electrónico de las personas que comparten el archivo. De esta manera es posible recolectar toda la información necesaria para hacer la calificación de la actividad llevada a cabo así como de

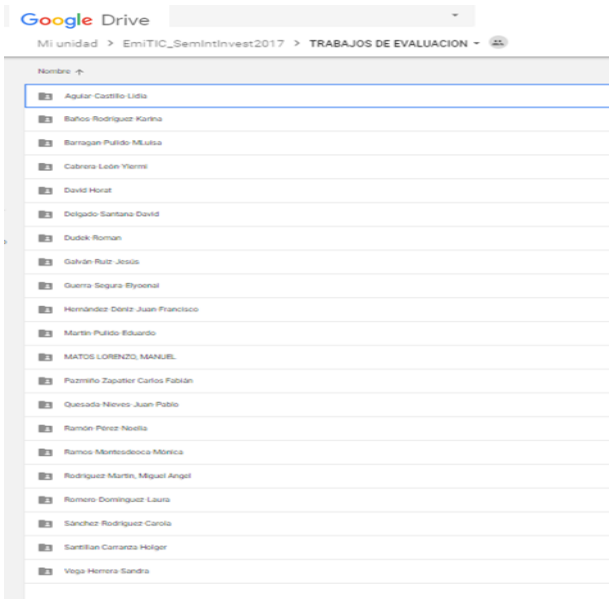
los tiempos en los que se hicieron los comentarios y correcciones. Con el sistema de versiones de Google Drive es posible conocer exactamente la traza de actuación de cada alumno y profesor.

En la asignatura Tecnologías de Internet de Nueva Generación, obligatoria, del primer curso del MUIT, se matriculan muy pocos alumnos, y un detalle importante es que estos alumnos provienen del Grado en Ingeniería en Tecnologías de Telecomunicación que es un grado generalista con 4 intensificaciones (Sonido e Imagen, Sistemas de Telecomunicación, Electrónica e Ingeniería Telemática). En estos dos últimos años no ha cursado esta asignatura ningún alumno de Ingeniería Telemática, lo que podría llevar a pensar que los alumnos tienen menos experiencia utilizando servicios de Google. Pero en la práctica no ha sido así y los alumnos inmediatamente han dominado dichos servicios. En la Fig. 4.a se muestra el número de alumnos que ha comenzado y finalizado la asignatura en los dos últimos cursos. Se muestra que el número de alumnos es muy reducido. La cantidad de profesores que atiende a esta asignatura son dos, sin embargo el profesor encargado de las prácticas es uno sólo y el de los problemas en el aula es también uno sólo con lo cual el ratio de alumnos por profesor es reducido. La particularidad de esta asignatura en cuanto a evaluación es que al existir un solo profesor evaluador de tareas, se hace innecesario coordinar la rúbrica de evaluación.

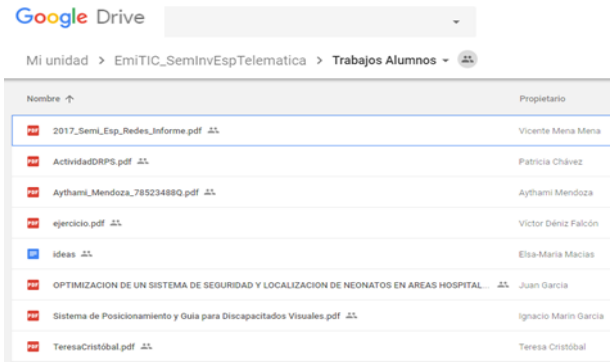
En el Seminario de Introducción a la Investigación (EmiTIC), obligatorio, se han matriculado del orden de 20 alumnos (Fig. 4.b) y tiene una duración de 12 horas. Un detalle muy importante es que los alumnos de este seminario provienen de Ingeniería de Telecomunicación, Informática, de Arquitectura de Edificios, Ciencias Económicas y Empresariales y en menor proporción de Administración de Empresas. Es decir, la heterogeneidad del perfil de los alumnos es muy elevado. Se observa una mayor dificultad para elaborar las memorias de las tareas marcadas (básicamente un trabajo de seminario en el que deben elaborar un borrador de proyecto de investigación en su línea de investigación e innovación) entre aquellos alumnos que no son de Ingeniería de Telecomunicación ni de Informática. Sin embargo, en la práctica, todos los alumnos han podido trabajar de forma eficaz y sin mayores problemas con las carpetas compartidas y el correo electrónico para gestionar comentarios. El número de profesores del seminario es 4 y las tareas se reparten de forma equitativa entre todos ellos. Por tanto es necesario acordar una rúbrica de evaluación que incluye aspectos de la profundidad con la que se ha elaborado la memoria, la cantidad de referencias bibliográficas usadas con índice de impacto y la capacidad de síntesis de esas referencias bibliográficas, entre otros aspectos. Un parámetro importante es la cantidad de ideas que se han cruzado entre alumnos que se evalúa de forma positiva. En la práctica, la colaboración entre alumnos es muy elevada y entre ellos se cruzan mucha información proporcionándose ideas valiosas. La evaluación se va haciendo cada 3 días (el plazo es de 15 días) y en general un 60% de los alumnos suele mejorar considerablemente sus trabajos con las revisiones del profesor y comentarios de otros alumnos. Además, la homogeneización en la evaluación siempre ha sido muy elevada, dado que todos los profesores comparten sus evaluaciones.



(a)



(b)



(c)

Fig. 3. Carpeta de google Drive compartidas para la asignatura, el seminario y el módulo: a) Tecnologías de Internet de Nueva Generación, b) Seminario de Introducción a la Investigación (EmiTIC) y c) Módulo Despliegue de redes, protocolos y servicios telemáticos en la Empresa 2.0

En este curso, en el módulo Despliegue de redes, protocolos y servicios telemáticos en la Empresa 2.0 (EmiTIC), optativo, se han matriculado 7 alumnos del total de 10 que suelen asistir a las clases del segundo curso (Fig. 4.c). Todos tienen un perfil de Informática o Ingeniería de Telecomunicación, por lo que podemos decir que son homogéneos. La ventaja de este módulo es que los alumnos ya conocen bien el método de evaluación del curso anterior. En este módulo de 6 horas lo imparten tres profesores, se invita a expertos de otras universidades y la Empresa a que impartan charlas muy breves (*story telling* de 20 minutos) pero sólo uno de los profesores es el encargado de evaluar las memorias de los trabajos que versan sobre la optimización del despliegue de un servicio telemático innovador en la Empresa. Se pacta con los alumnos que cada 3 días, dentro del plazo de 15 días, se hacen revisiones por parte del profesor. El 60% de los alumnos suele trabajar al día a día en la mejora de sus trabajos. A los alumnos que hacen la tesis doctoral y además trabajan (alumnos a tiempo parcial), se les facilita otras temporizaciones.

En la Fig. 5 se muestran las notas obtenidas por los alumnos. El objetivo no es comparar datos cuantitativos entre distintas asignaturas (ni entre cursos d ela misma asignatura), sino demostrar que en todos los cursos, por asignatura, las notas son bastante buenas. Destacar que el tiempo dedicado a las reclamaciones de notas es nulo en todos los casos.

IV. Conclusiones

La evaluación del aprendizaje del alumno es una cuestión importante en la docencia universitaria. Nosotros hemos presentado una experiencia docente en varias asignaturas con baja proporción de alumnos (menos de 10) por profesor en las que hemos aplicado nuestro método de evaluación abierta en tiempo real. La característica más importante de este método es que permite aumentar la competencia y la cooperación entre los alumnos y homogeniza el trabajo de evaluación de los profesores, minimizando la procastinación de los alumnos y reclamaciones a la calificación obtenida. Queremos destacar que en las asignaturas en las que hemos aplicado el método se ha fomentado la competitividad sana, alcanzándose en todas un grado considerable de cooperación entre profesores, entre alumnos y entre profesores y alumnos. Aunque no se ha graficado los valores, para la asignatura Tecnologías de Internet de Nueva generación, los resutlados obtenidos son mejores que en los cursos previos a los dos últimos.

Al finalizar la escritura de este artículo Google ha decidido abrir la plataforma *Google Classroom* (*classroom.google.com*) a todo aquel que tenga una cuenta activa en *gmail* sin necesidad de tener cuenta corporativa en *Google (G) Suite*. Las primeras pruebas realizadas sobre dicha plataforma aconsejan migrar todo el sistema de notificación a los alumnos y manejo de carpetas compartidas y calendarios, a esta plataforma. Por tanto, como trabajo futuro pretendemos migrar toda nuestras ideas de docencia abierta a esta plataforma.

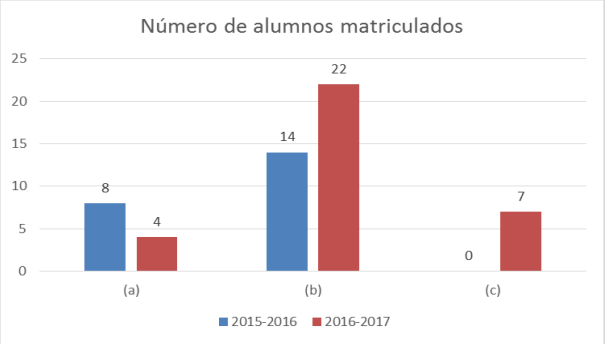
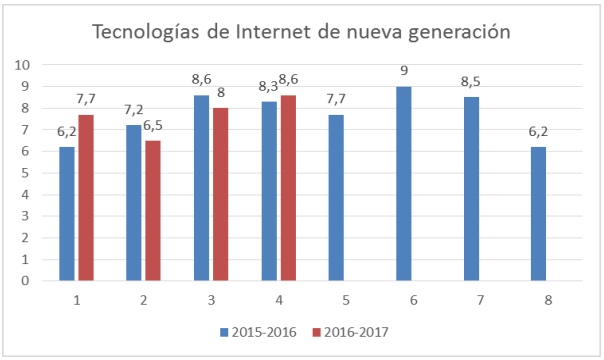
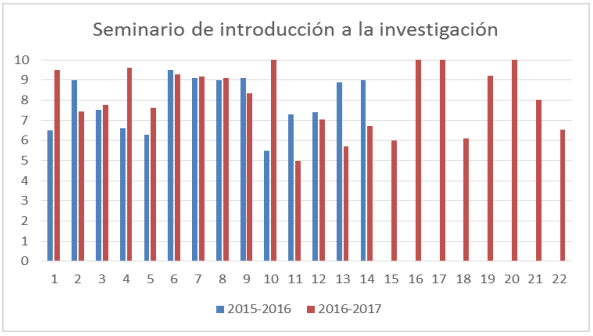


Fig. 4. Número de alumnos de los dos últimos cursos en: a) Tecnologías de Internet de Nueva Generación, b) Seminario de Introducción a la Investigación (EmiTIC) y c) Módulo Despliegue de redes, protocolos y servicios telemáticos en la Empresa 2.0 (sólo se ha impartido el último curso)





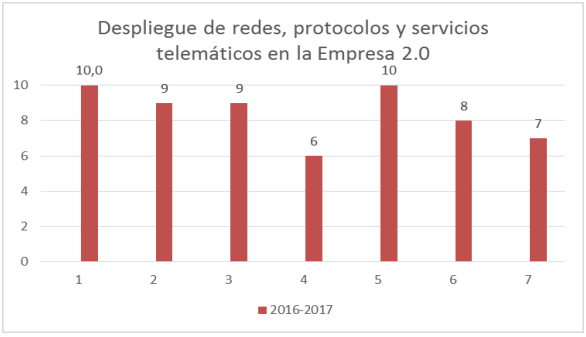


Fig. 5. Notas obtenidas por los alumnos

AGRADECIMIENTOS

Los autores quieren agradecer a los alumnos de estas materias por haber participado de la motivación que se les trató de contagiar.

REFERENCIAS

[1] OpenLearning: [Online]. Available: https://www.openlearning.com/

[2] SmartLearning: [Online]. Available: http://www.smart-learning.co.uk/

[3] S. Leone, *Characterisation of a Personal Learning Environment as a Lifelong Learning Tool*. New York: Springer-Verlag. 2013. ISBN 978-1-4614-6273-6.

[4] MiríadaX: [Online]. Available: https://miriadax.net/home

[5] M. Khalil, M. Ebner. (2016, June). What is Learning Analytics about? A Survey of Different Methods Used in 2013-2015 [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1606/1606.02878.pdf

[6] L. Joyanes, *Big Data. Análisis de Grandes Volúmenes de Datos en Organizaciones*. Marcombo. 2013. ISBN 978-8-4267-2081-8.

[7] B. Salinas-Fernández, C. Cotillas-Alandí (coordinadores), *La evaluación de los Estudiantes en la Educación Superior. Apuntes de Buenas Prácticas.* Servicio de formación permanente de la Universidad de Valencia, DL: V-4566-2007, 2007. Available: http://www3.uji.es/~betoret/Formacion/Evaluacion/Documentacion/La%20evaluacion%20estudiantes%20en%20la%20ESuperior%20UV.pdf

[8] Elsa Macías, Alvaro Suárez, "Organización de docencia cooperativa usando Google Drive ", En Actas de las III Jornadas de Innovación Educativa Telemática, 2013.

[9] M. de Miguel, "Cambio de paradigma metodológico en la Educación Superior: Exigencias que conlleva", Cuadernos de integración europea (Fundació General de la Universitat de València: Centre de Documentació Europea), ISSN 1885-1754, no. 2, 2005.

[10] Reglamento de planificación académica de la ULPGC, BOULPGC Año VII, nº 1, 14 de enero de 2014. [Online]. Available: https://www2.ulpgc.es/hege/almacen/download/7107/7107474/reglamento_de_planificacion_academica.pdf

[11] Documento de Verificación del Título de Master en Ingeniería de Telecomunicación de la Escuela de Ingeniería de Telecomunicación y Electrónica (EITE). Available: http://www.eite.ulpgc.es/images/directorio_master/2015-02-24%20Documentos%20MUIT/10.%20Memoria%20de%20verificaci%C3%B3n%20del%20t%C3%ADtulo_ANECA.pdf.

[12] Reglamento de Evaluación de los Resultados de Aprendizaje y de las Competencias adquiridas por el alumnado en los Títulos Oficiales, Títulos Propios y de Formación Continua de la ULPGC. Available: https://www2.ulpgc.es/hege/almacen/download/7115/7115248/14_enero_2014__reglamento_de_evaluacion.pdf.

[13] E. Herrero-Curiel, Llegan los exámenes y con ellos las ganas de procrastinar. Available: http://www.cop.es/colegiados/m-13106/images/PrensaDigital3-1209.pdf.

[14] ¡ Deja de procrastinar! El mejor 'software' para concentrarse mientras trabajas. Available: http://www.elconfidencial.com/tecnologia/2016-05-05/software-aplicaciones-productividad-windows-mac_1194862/.

[15] A qué dedican el tiempo los profesores universitarios, Available: http://nadaesgratis.es/admin/a-que-se-dedican-los-profesores-de-universidad.

[16] R. C. Martin, *Agile Software Development, Principles, Patterns, and Practices*. Prentice Hall, 2002. ISBN: 0-13-597444-5.

[17] J. Alsina-Masmitjà (coordinador), *Rúbricas Para La Evaluación De Competencias*. ICE y ediciones octaedro, 2013. ISBN: 978-84-9921-476-4, 2013. Available: http://www.ub.edu/ice/sites/default/files/docs/qdu/26cuaderno.pdf.

[18] R. S. Contreras, J. L. Eguia (Eds.), *Gamificación en las Aulas Universitarias*. Instituto de la Comunicación, Unversidad Autónoma de Barcelona, 2016. ISBN: 978-84-944171-6-0. Available: http://incom.uab.cat/download/eBook_incomuab_gamificacion.pdf.

[19] Asignatura: Tecnologías de Internet de Nueva Generación (proyecto docente). Available: http://www2.ulpgc.es/aplicaciones/proyectosdocentes/pdf.php?id_proyecto=47946&NUEVA=1.

[20] Memoria de Verificación el Programa de doctorado Empresa, Internet y Tecnologías de las Comunicaciones. Available: http://edulpgc.ulpgc.es/sites/default/files/IMCE/Programas_doctorado/memorias_verificacion/Memoria%20Definitiva%20ANECA%20PD%20EmITIC.pdf.

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6596

Uso de Software-Defined Radio en la enseñanza de sistemas de comunicaciones

Jaume Segura-Garcia, Antonio Soriano-Asensi, Carmen Botella-Mascarell
Santiago Felici-Castell, Miguel García-Pineda
Departament d'Informàtica,
Universitat de València
Avda de la Universitat s/n - 46100 Burjassot - València.
jaume.segura@uv.es, antonio.soriano-asensi@uv.es,carmen.botella@uv.es
santiago.felici@uv.es, miguel.garcia-pineda@uv.es

***Resumen*—En la docencia de sistemas de comunicación hay una componente teórica elevada. La percepción de los estudiantes al estudiar estas asignaturas es negativa. Este trabajo trata de explicar la motivación y la estrategia seguida para reorientar esta percepción a partir de la introducción de elementos de "Software-Defined Radio" (SDR) y "Universal Software Radio Peripherals" (USRP) en diferentes asignaturas del Grado de Ingeniería Telemática de la Universitat de València.**

***Palabras Clave*—Software-Defined Radio, telemática, USRP, RTL-SDR, RDS, DVB-T**

I. INTRODUCCIÓN

La enseñanza en el Grado de Ingeniería Telemática (GIT) tiene una notable carga docente en sistemas de comunicación. Algunos autores [1] han tratado la integración curricular de las tecnologías de la comunicación en las aulas, estableciendo así una serie de criterios para esta integración, entre ellos encontramos:

- el análisis de la calidad de los recursos, que siendo de diferentes niveles permiten interaccionar con los sujetos, ya que de acuerdo con estos autores lo importante es esta interacción entre sujetos y recursos/medios;
- la inserción de los contextos metodológicos adecuados, ya que un potente medio puede tener menos potencialidad si el método en el que se incluye no es acorde a los objetivos buscados;
- la identificación de los destinatarios adecuados, ya que los recursos deben estar adaptados a las necesidades y capacidades de los estudiantes. De esta forma puede entenderse que a determinados niveles de maduración sea más viable el uso de recursos que otros;
- la conducción del profesor, ya que los estudios empíricos demuestran que el conocimiento y la implicación de éste es uno de los factores decisivos para determinar la bondad del recurso.

Sin embargo, se debe tener en cuenta una serie de condiciones que comprenden que los recursos no sustituyen al profesor y que requieren un uso reflexivo, crítico y adaptado a la realidad de los estudiantes. Con todo ello, se debe contemplar la explotación de los recursos de innovación para que los estudiantes consigan un aprendizaje significativo [2].

De acuerdo con lo anteriormente expuesto, nuestro propósito en este artículo es explicar la motivación y el desarrollo metodológico basado en el uso de plataformas Software-Defined Radio (SDR) para la docencia en el Grado de Ingeniería Telemática de la Universitat de València, así como explicar una prueba piloto desarrollada e implementada en una asignatura del grado (Fundamentos de Sistemas de Telecomunicación) basada en SDR y hacer una propuesta para otra (Transmisión de Datos) basada en DVB-T. Estos desarrollos y propuestas se han realizado en el marco de un proyecto de innovación educativa que ha sido financiado en parte por la Universitat de València. El resto del artículo se desglosa de forma resumida en las siguientes secciones: metodología en la que se desarrolla el marco en el que se desarrolla el proyecto, desarrollos del proyecto, la descripción del material sobre el que se desarrolla el cambio de paradigma en las asignaturas de Fundamentos de Sistemas de Telecomunicación y una propuesta para la asignatura de Transmisión de Datos del Grado de Ingeniería Telemática de la ETSE de la Universitat de València.

II. METODOLOGÍAS PARA LA ENSEÑANZA DE COMUNICACIONES

El desarrollo metodológico que se ha usado está basado en la apreciación que se tenía en los últimos años sobre el desarrollo teórico-práctico de los laboratorios

de las asignaturas de Fundamentos de Sistemas de Telecomunicación, Transmisión de Datos, Teoría de Comunicación, Comunicaciones Inalámbricas y Movilidad, del Grado de Ingeniería Telemática y de asignaturas que desarrollan contenidos más avanzados en el marco del Máster de Ingeniería en Telecomunicaciones. Los contenidos que se distribuyen en las diferentes asignaturas se desarrollan progresivamente especificando: diferentes modulaciones digitales (en banda base y pasabanda), diferentes técnicas de ecualización de canal para evitar interferencias intersimbólicas (ISI) monoportadora y multiportadora, técnicas de codificación de canal, técnicas de sincronización, técnicas de diversidad y de espectro ensanchado, etc.

Esta apreciación estuvo refrendada por la encuesta que se realizó a estudiantes de estas asignaturas en el curso 2015-16, en la que se les preguntó sobre la visión de las comunicaciones que ofrecían estas asignaturas en el marco del grado/máster y que trata de valorar la necesidad de una mayor carga experimental en la docencia sobre comunicaciones.

Se planteó el desarrollo de esta carga experimental basada en Software Defined Radio (SDR). El concepto de SDR fue introducido por Joseph Mitola [3] y establece un nuevo paradigma educativo en el ámbito de las telecomunicaciones que permite implementar mediante software muchos componentes de sistemas de radiocomunicaciones y éstos pueden ser reconfigurados en línea. Con ello se consiguen plataformas hardware inalámbricas multi-estándar, multi-banda y multifuncionales.

Durante el curso 2016-2017 se ha iniciado este cambio de paradigma que se espera tenga un impacto notable en la evolución de futuras comunicaciones inalámbricas y sistemas en red.[4]

Algunas herramientas basadas en esta tecnología, como son las "Universal Software Radio Peripherials" (USRP) o RTL-SDR, son opciones abiertas, económicas y que ofrecen una versatilidad suficiente, que con un enfoque pedagógico adecuado pueden ser muy útiles para el estudio de estándares tecnológicos actualmente vigentes. También, el uso de estándares como Radio Data System (RDS) [5] o Digital Video Broadcasting - Terrestrial (DVB-T) [6] resultan muy útiles para este desarrollo metodológico, ya que tienen un buen solape con los contenidos desarrollados en las asignaturas mencionadas (al menos en las asignaturas que se van a utilizar como piloto en el desarrollo de este proyecto).

Una Universal Software Radio Peripheral (USRP)[7] es una plataforma diseñada por Ettus Research (y actualmente vendida por National Instruments), basada en una arquitectura con FPGA y una capa de comunicaciones (daughter-board) que es intercambiable. A este respecto, tomamos como referentes los casos de la Universidad de Cantabria, la Universidad de Sevilla [8] y la Universidad de Washington [9], que ya han incorporado estos elementos a su docencia.

La aproximación metodológica seguida se ha basado en el diseño de una serie de sesiones prácticas usando RTL-SDR o USRPs y GNU Radio-Companion que desarrollan diversos conceptos de las dos asignaturas implicadas. Estas sesiones prácticas, basadas en el desarrollo conceptual del estándar RDS[5] y del estándar DVB-T [6], han sido introducidas en el curso 2016-2017 en los grupos de laboratorio de la asignatura de Fundamentos de Sistemas de Telecomunicación y se prevé que durante el curso 2017-2018 se introduzca en la asignatura de Transmisión de Datos, con sesiones de laboratorio basadas en DVB-T [6].

III. SDR para los "Fundamentos de Sistemas de Telecomunicación"

La asignatura de Fundamentos de Sistemas de Telecomunicación (FST) está planteada como una introducción a los sistemas de telecomunicaciones. Tiene carácter obligatorio y se imparte en el segundo cuatrimestre del segundo curso del Grado de Ingeniería Telemática. En la asignatura se introducen las bases y fundamentos de los sistemas de telecomunicaciones: el soporte físico de las comunicaciones, el uso del espectro electromagnético y la implementación física de los canales de radio. También se aborda el problema de la representación de la información en banda base y su transformación a pasa banda empleando diferentes tipos de modulación, tanto lineales como no lineales.

El laboratorio de FST está planteado de forma que proporciona a los estudiantes una aplicación práctica de los conceptos de comunicaciones vistos en el aula. En las tres primeras sesiones del laboratorio de FST se estudian diferentes tipos de modulación y demodulación analógica. En particular, se estudian las modulaciones en amplitud (AM): AM–convencional, Dual Side Band (DSB), Single Side Band (SSB); y la modulación analógica en frecuencia (FM) y en fase (PM). En estas tres sesiones, basadas en Matlab, los alumnos tienen ocasión de estudiar cómo se relacionan las señales modulada, portadora, moduladora y demodulada para cada uno de los tipos de modulación mencionados. También tienen ocasión de comprobar las ventajas e inconvenientes de cada tipo de modulación.

Tras las tres primeras prácticas, realizadas en Matlab, en las que se estudian los diferentes tipos de modulación analógica se emplean tres sesiones de laboratorio en las que se particularizan los conocimientos aprendidos al caso de un receptor de radiodifusión de FM (basado en el estándar RDS [5]). Estas tres sesiones son la primera toma de contacto de los estudiantes del Grado de Ingeniería Telemática con el entorno GNURadio.

A. Analizador de espectros basado en SDR

La primera de las sesiones basadas en SDR pretende ser una introducción al entorno de desarrollo GnuRadio. En esta primera sesión, los estudiantes implementarán un analizador de espectros de la señal de FM adquirida por el RTL-SDR. Al finalizar la práctica, el estudiante debe ser capaz de entender el funcionamiento del entorno de desarrollo GnuRadio y debe saber configurar correctamente el bloque *osmocom source*, encargado de

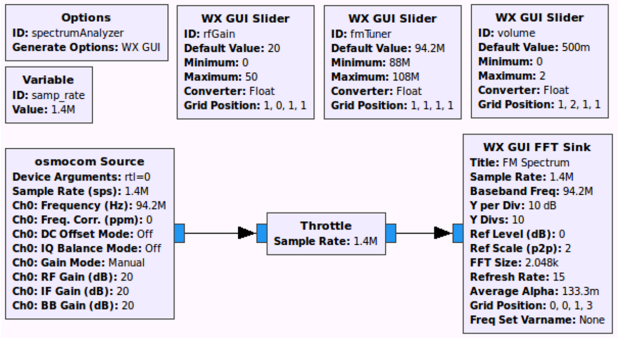


Fig. 1. Diagrama de bloques del analizador de espectros implementado en la primera sesión de laboratorio.

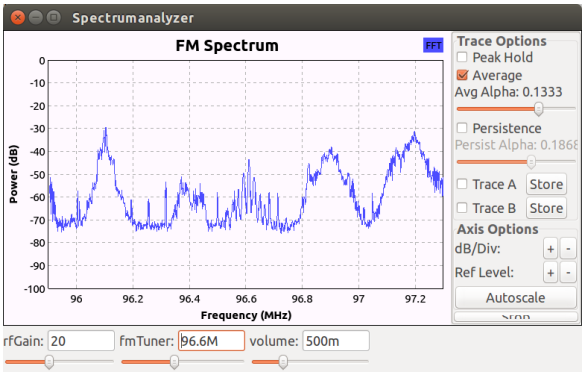


Fig. 2. Resultado de la ejecución del diagrama de flujo implementado.

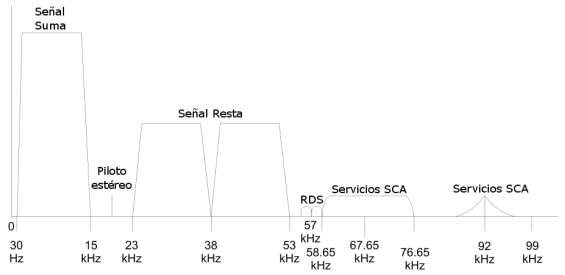


Fig. 3. Estructura de la información en cada emisora de FM.

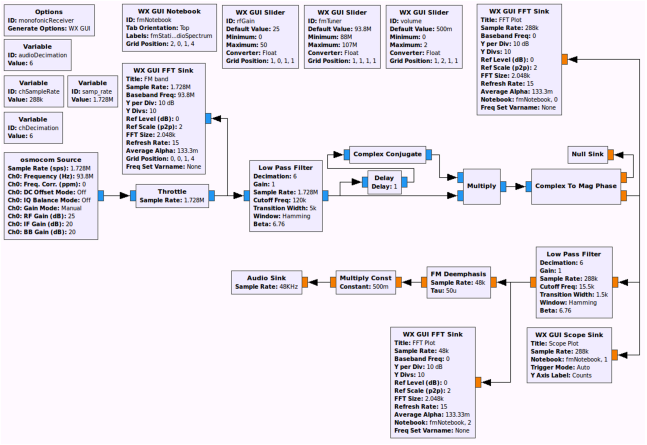


Fig. 4. Diagrama de bloques de un receptor de FM monofónico.

actuar como interfaz entre el entorno de desarrollo y el hardware de RF.

En la Figura 1 se muestra el diagrama de flujo del analizador de espectros que se implementará en esta práctica. El bloque *osmocom source* se encarga de tomar las muestras de fase (I) y cuadratura (Q) digitalizadas por el RTL2832U. Se emplea una frecuencia de muestreo de 1,44 MHz para poder visualizar de forma simultánea varias estaciones de radio y facilita la adecuada sintonia de cada emisora.

Además de los bloques estrictamente relacionados con el proceso de adquisición y demodulación de la señal de FM, se fomenta que el alumno se familiarice con otros bloques propios de GnuRadio empleados para la representación gráfica de señales. A lo largo de esta sesión se trabaja con el bloque *wx gui fft sink*, empleado para representar el espectro de cada señal. El diagrama de bloques propuesto en Figura 1 emplea diferentes tipos de variables para facilitar la sintonización de las diferentes estaciones de radio (*fmTuner*) y ajustar la ganancia de la etapa de RF del RTL–SDR (*rfGain*) de forma dinámica sin necesidad de detener el proceso de toma de datos. Además, a lo largo de toda esta sesión, se trabaja la forma en que deben situarse los diferentes controles en la ventana en que se muestran los resultados Figura 2.

Tras completar el montaje del diagrama de flujo propuesto en esta práctica, se propone a los estudiantes diferentes actividades para que se familiaricen con la configuración del bloque *osmocom source* en las que deben observar cómo varía el espectro recibido al cambiar los valores de la ganancia de la etapa de RF, la frecuencia de muestreo y la frecuencia de sintonización. También se les proponen dos actividades en las que deben determinar la relación señal ruido (SNR) y el ancho de banda de varias emisoras.

B. Receptor monofónico

La propuesta para la segunda sesión de laboratorio sobre GnuRadio es tomar el trabajo de la sesión anterior y añadir los bloques necesarios para realizar la demodulación FM, extraer la componente monofónica, y adecuarla para ser enviada a la salida de audio del PC.

En la Figura 3 se muestra cómo se organiza la información que transmite cada emisora de FM. En el rango de frecuencias inferiores a 15 kHz se encuentra la señal monofónica (L+R), y la diferencia entre ambos canales (L-R) se modula en AM empleando una portadora de 38 kHz, de forma que ocupa la banda entre 23 kHz y 53 kHz. Para poder reconstruir la portadora en el receptor, se envia una señal piloto formada por un armónico de 19 kHz. Haciendo uso de la portadora reconstruida a partir de la señal piloto es posible realizar la demodulación AM de la señal (L-R). Combinando las señales L-R y L+R es posible extraer las señales L y R, y así reconstruir la señal estereofónica.

En la Figura 4 se muestra el diagrama de flujo del receptor FM monofónico que se implementa en la segunda sesión de laboratorio. El diagrama de la Figura 4 está compuesto por tres etapas. La primera de ellas corresponde

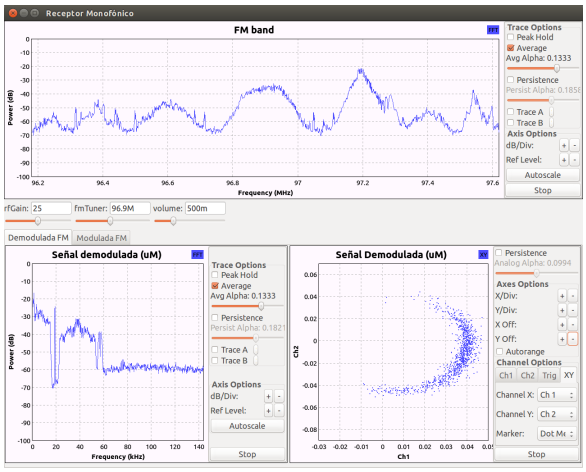


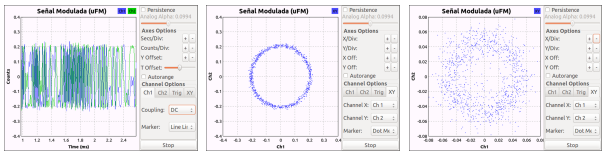
Fig. 5. Representación de la señal obtenida tras la demodulación FM. (Arriba) Espectro de la señal registrada por el RTL–SDR. (Inferior izquierda) Espectro de la señal obtenida tras la demodulación FM. (Inferior derecha) Representación I-Q de la señal demodulada.

al bloque *osmocom source* que recibe las muestras registradas por el RTL2832U, y cuyo funcionamiento ya se vió en la sesión anterior. La segunda etapa se encarga de la demodulación FM de la señal registrada. La tercera etapa extrae la componente monofónica (L+R) y la acondiciona para enviarla a la salida de audio del PC.

En Figura 5 se muestra la ventana que resulta de la ejecución del diagrama de flujo implementado en la práctica. En la parte superior se aprecia el espectro de la señal recibida por el RTL-SDR. En la parte inferior se aprecia la señal obtenida tras la demodulación FM, tanto en el dominio temporal (inferior derecha) como en el dominio de la frecuencia (inferior izquierda). En el espectro de la señal demodulada se aprecian claramente las señales L+R, piloto, L-R y RDS. Las actividades planteadas en esta segunda sesión de laboratorio son de tipo cualitativo. Se pide a los estudiantes que revisen la implementación del demodulador de FM que previamente completaron en una práctica teórica basada en Matlab, con la estructura de bloques empleada en este caso para demodular la señal de FM, y que evaluen las ventajas e inconvenientes de cada una de ellas. Al configurar en modo xy el osciloscopio en el que se muestra la señal temporal, es posible visualizar las muestras I-Q registradas, Figura 6(b). También se les pide que relacionen las representaciones I-Q antes y después de la demodulación Figura 5 Y que estudien cómo afecta del SNR (Figura 6(c)) y la potencia recibida en la representación I-Q.

C. Receptor estereofónico

De acuerdo con el diagrama mostrado en la Figura 3, para recomponer la señal estereofónica es necesario extraer la información de la resta de canales (L-R) y demodularla en amplitud para combinarla con la suma de canales (L+R) y así extraer las componentes derecha e izquierda. El primer paso es generar la señal de 38 kHz para realizar la demodulación en amplitud. Mediante un filtro pasa–banda,



(a) (b) (c)

Fig. 6. Representación de la señal modulada en FM. (a) Representación temporal de las componentes I (azul) y Q (verde). (b) Representación I-Q de la señal modulada. (c) I-Q registrada de una emisora con SNR bajo.

se selecciona la señal piloto de 19 kHz y se dobla la frecuencia. La señal de 38 kHz resultante se emplea para demodular en amplitud la diferencia de canales (L-R), que se ha seleccionado mediante un filtro pasa–banda entre 22 kHz y 54 kHz. Las componentes suma y diferencia se combinan para extraer los canales L y R. Tras aplicar el filtro de deénfasis, cada una de las componentes es enviada a uno de los canales de la salida de audio del PC. En la Figura 7 se presenta un diagrama de bloques completo del receptor de FM que implementan los alumnos en esta sesión de laboratorio.

Tras la ejecución del diagrama de bloques desarrollado en la presente práctica se observa la ventana mostrada en Figura 8. En la que se puede observar la señal (L-R) antes y después de la demodulación AM. En la parte inferior derecha de Figura 8 se aprecia en verde la señal (L-R) antes de la demodulación AM. En azul se presenta la imagen que resulta tras la demodulación. El diagrama de flujo Figura 7 que debe implementarse en esta tercera práctica es bastante más complejo que el de las dos prácticas anteriores, por lo que se les proporciona un diagrama que ya contiene la parte implementada anteriormente. Adicionalmente a la implementación del diagrama de flujo en esta sesión de laboratorio se propone que los alumnos completen el diseño de los filtros con que se seleccionan las componentes (L+R) y (L-R). También se emplea un pequeño transmisor de FM que emite la señal de la salida de audio del PC. El transmisor de FM permite visualizar cómo al aumentar la potencia del mensaje se ensancha o estrecha el ancho de banda de la señal de FM. En definitiva, les permite entender cómo afectan las variaciones en la potencia del mensaje en cada una de las etapas del receptor desarrollado.

IV. DVB-T para la "Transmisión de Datos"

La asignatura Transmisión de Datos pertenece a la materia de Comunicaciones Digitales, formada además por las asignaturas Fundamentos Matemáticos de las Comunicaciones (segundo cuatrimestre de segundo curso), Teoría de la Comunicación (primer cuatrimestre de tercer curso) y Procesado Digital de la Señal (segundo cuatrimestre de tercer curso). En la asignatura de Transmisión de Datos del Grado de Ingeniería Telemática, los contenidos que se imparten están ligados a la modulación y demodulación (o detección) digital, la ecualización como medio de mitigación de la ISI, la codificación convolucional (como otro tipo de codificación

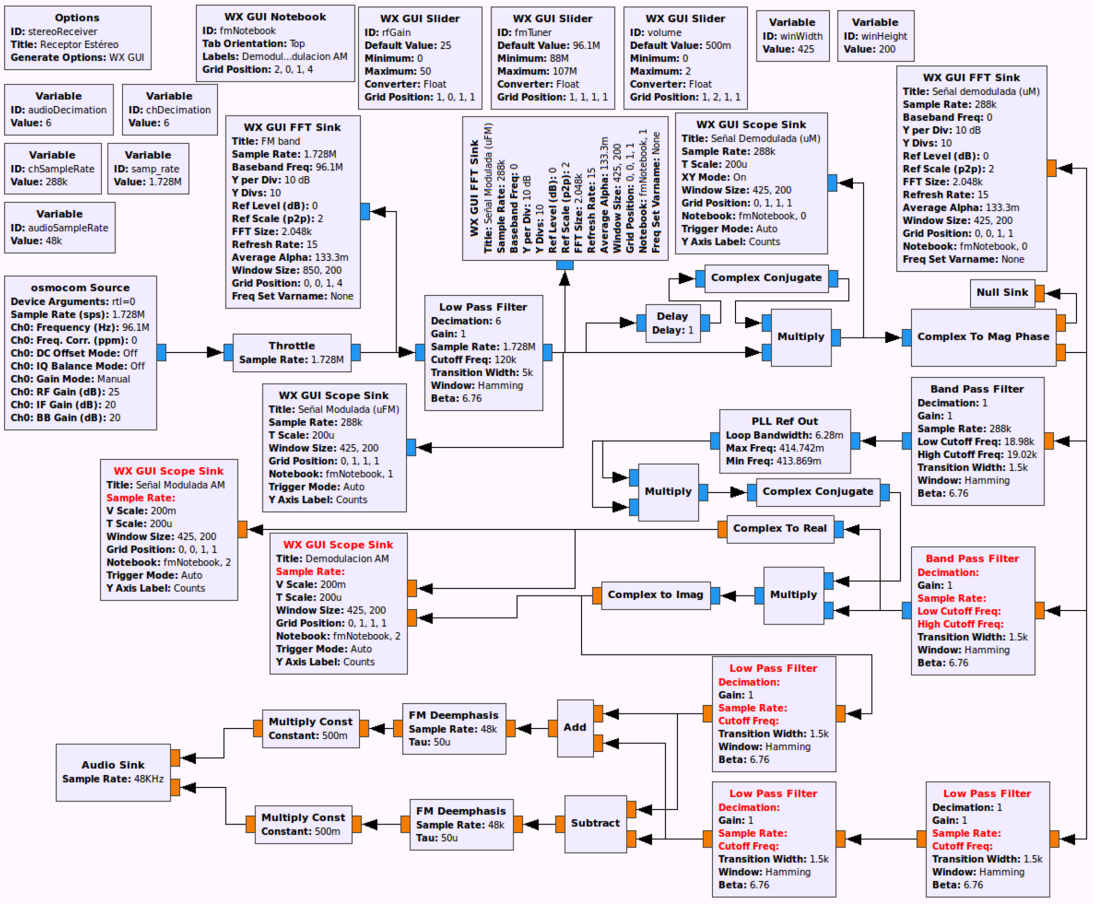


Fig. 7. Diagrama de bloques de un receptor de FM estereofónico.

Fig. 8. Resultado de la ejecución del diagrama de bloques del receptor estereofónico de FM.

de canal para la corrección de errores debidos al ruido del canal) y la sincronización.

El estudio del estándar ETSI 300 744 (DVB-T) [6] como medio que vertebre los contenidos de esta asignatura puede ser útil para la motivación, basándonos en el interés de las aplicaciones de éste, y para la comprensión de contenidos a partir de la experiencia. El estándar ETSI 300 744 DVB-T [6] está vinculado a la difusión de señal digital de TV. Para el curso 2017/2018 está previsto aplicar una serie de laboratorios basados conceptos de SDR y en dispositivos USRP y GNURadio-Companion. En las sesiones de laboratorio se desarrollarán los conceptos de la asignatura mediante la implementación de diferentes modelos de emisores y receptores DVB-T que permitan integrar los conocimientos adquiridos e ir acumulando elementos del sistema de difusión hasta completar el Tx/Rx de DVB-T. Además se propondrán trabajos de fin de grado para la implementación de algoritmos de sincronización, demodulación, ecualización, codificación, etc., que permita extender el conocimiento de los estudiantes y permita tener herramientas de test para una implementación real del estándar DVB-T en el entorno académico.

Como primer ejercicio, y previamente a la primera de estas sesiones de laboratorio, se les pedirá la búsqueda de información de implementaciones *open-source* del estándar DVB-T. Por ejemplo, entre ellas se puede proponer la búsqueda de la implementación de OpenDVB, la implementación de Giusseppe Barruffa (DVB-T simulator) o la de Vicenzo Pellegrini (SR-DVB).

El desarrollo de estas sesiones de laboratorio para Transmisión de Datos se hará mediante el uso de USRPs con una *daughterboard* WBX para Rx/Tx (o TVRx2 para

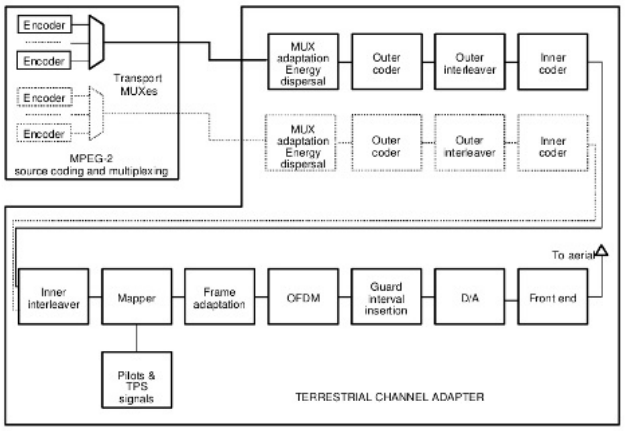


Fig. 9. Cadena del transmisor (tomada del ETSI EN 300 744)

el receptor) y sticks RTL2832u DVB-T como receptor (con frecuencias de operación entre 862 MHz y 1700 MHz).

El transmisor incluirá una aplicación de streaming, tipo VLC o Kaffeine, que se usará como fuente de video para comprobar el funcionamiento del sistema de radiodifusión.

A. Implementación del transmisor

El estándar ETSI EN 300 744 define ambas implementaciones, tanto desde el punto de vista del transmisor como del receptor. La Figura 9 presenta la cadena de transmisión que se implementará.

Por otra parte, existe documentación Doxigen sobre esta implementación en GNURadio recogida en la implementación del repositorio C-GRAN[1]. La implementación para GNURadio-Companion se muestra en la Figura 10. La mayoría de bloques están implementados como nuevos bloques GNURadio, por lo que se hace necesario la instalación del código de este repositorio.

Los módulos implicados en la implementación anterior son:

- Dispersión de energía: Con el fin de dispersar la energía en la fuente MPEG2-TS, este bloque realiza un xor de los datos reales con la salida del generador PRBS. También crea frames MUX compuestos de 8 paquetes MUX de 188bytes. La sincronización se realiza reemplazando el byte $0x47$ de sincronización con $0xb8$ al principio del frame MUX.
- Codificador externo (Reed-Solomon): Éste es un tipo de código de bloque. Este bloque implementa RS(188, 204), codificando con t=8 capacidad de corrección de errores.
- Entrelazador convolucional: La cuestión al usar un *codificador externo + entrelazador + codificador interno* es generar una palabra código larga. El entrelazador interno es de tipo Ramsey III.
- Entrelazador interno: El entrelazador interno está hecho a partir de un entrelazador de símbolos y un entrelazador bit a bit. El resultado debe prepararse para mapear el número de bits a la constelación de acuerdo con el tipo de modulación utilizado.
- Mapeador: Asigna los bits a una constelación y esto depende no sólo del tipo de modulación, sino también de si se utiliza la modulación jerárquica.
- Bloque de señales de referencia: Para soportar la sincronización y la ecualización en el lado del receptor, la norma añade tres tipos de pilotos: pilotos continuos, pilotos dispersos y *transmission parameter signaling* (TPS). El bloque TPS permite el envío los parámetros de transmisión y esto se hace usando la modulación diferencial que es inmune a las rotaciones de fase. La decodificación de TPS es lo primero a implementar en el receptor después de una sincronización inicial.
- FFT: La modulación COFDM (OFDM codificada) utilizada en DVB-T crea compartimientos de dominio de frecuencia y hace una I-FFT para convertirlos en dominio de tiempo. Se puede utilizar FFT de tamaño 2k, 8k y 4k.
- Prefijo cíclico: Para evitar ISI, ICI en el SFN (redes de frecuencia única) que se emplean normalmente en las implementaciones DVB-T, se utiliza un prefijo cíclico (CP) de varios tamaños. Consiste en copiar de la última parte de la señal en dominio del tiempo al principio de la señal. El hecho de que la señal sea periódica ayuda a la sincronización en el receptor. Muchos algoritmos para la corrección de tiempo y frecuencia se basan en CP.
- Resampler racional: El USRP N210 que usaremos como receptor y transmisor tiene un reloj de 100MHz. Por desgracia, en DVB-T el reloj en el que las muestras tienen un desplazamiento específico (9.14 Msps) y para eso necesitamos un resampler. Esto hace que el modelo se retrase ya que este bloque es el de mayor consumo de tiempo (tomado de GnuRadio). También se observa que el retraso de grupo tendrá variaciones, así que necesita ser investigado.

B. Implementación del receptor

La implementación del receptor DVB-T en GNURadio-Companion se describe en la Figura 11.

Los bloques utilizados para la recepción son básicamente los bloques transmisores en el orden inverso con una diferencia importante, aquí hay necesidad de bloques de sincronización para obtener una constelación limpia.

- Adquisición de símbolos OFDM: El objeto de este bloque es sincronizar el receptor de manera que se obtenga una señal en dominio del tiempo limpia antes del bloque FFT. Hay varias subtareas de las que este bloque se encarga: (1) usar Cyclic Prefix (CP) para obtener el inicio del símbolo OFDM en dominio del tiempo, (2) una vez empieza el CP, se aplica una de-rotación de la señal para obtener el símbolo OFDM correcto en dominio temporal y (3) se saca el CP delante del símbolo OFDM.

[1] http://www.cgran.org/pages/gr-dvbt.html

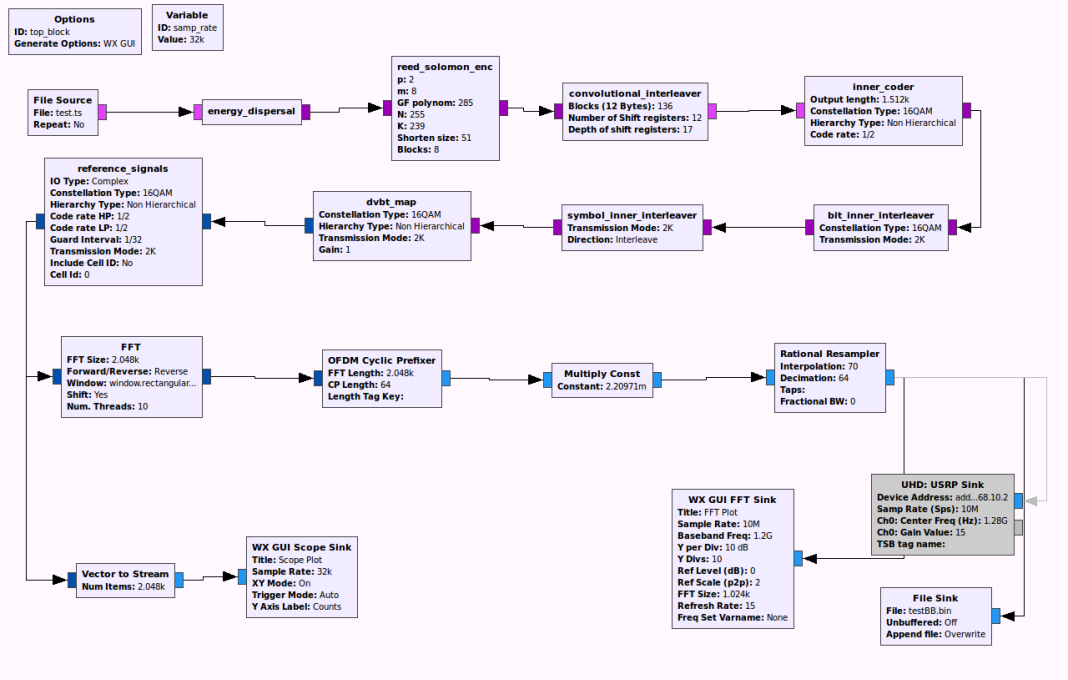


Fig. 10. Implementación del transmisor DVB-T en GNURadio-Companion

- FFT para la conversión de dominio de tiempo a frecuencia: Esto es simplemente conversión FFT para 2k, 4k o 8k dependiendo de los parámetros. Aquí se utilizará el bloque GNUradio. Haciendo una representación de la constelación después de la FFT, se puede ver que la constelación está girando y por lo tanto requiere más procesamiento para ser útil para el demapeo (ver Figura 12). Esta representación se realiza cuando se transmite con USRP N210 y se recibe con USRP N210, ambos equipados con *daughter-boards* WBX. Los parámetros para DVB-T son ancho de banda de 8MHz, FFT 2k, 16-QAM, FEC 1/2.
- Demodulación de señales de referencia: El estándar DVB-T utiliza varias subportadoras para insertar señales piloto que se utilizan para la sincronización, transmisión de parámetros de señal y ecualización. Como ya se ha explicado, hay tres tipos de señales piloto utilizadas en el estándar DVB-T: (1) señales piloto dispersas, (2) señal piloto continua y (3) TPS (Señales de Parámetros de Transmisión).

C. Tareas a realizar

Las tareas a realizar estarán basadas en el testeo del transmisor y el receptor DVB-T. El transmisor DVB-T soporta modos OFDM 2k/8k, constelaciones QPSK/QAM16/QAM64 y tasas de codificación 1/2, 2/3, 3/4, 5/6, 7/8.

En los ficheros demo que se presentan, se puede ejecutar codificación MPEG2-TS (para el transmisor, p.e. apps/dvbt_tx_demo.grc), se generan muestras de 10Msps en banda base. En la ejecución de este diagrama de bloques de ejemplo se tiene una configuración como: OFDM 2k, código FEC 1/2, modulación 16-QAM, e intervalo de guarda 1/32.

En el caso del receptor, la implementación soporta todos los tipos disponibles en el transmisor (i.e. QPSK/QAM16/QAM64) y tasas 1/2, 2/3, 3/4, 5/6, 7/8. Se probarán todas las tasas y constelaciones mediante el uso de OFDM 2k. El receptor implementado (para el receptor, usa apps/dvbt_rx_demo.grc) recogerá muestras en banda base en un fichero binario que se recodificará a *Transport-Stream* y se podría comparar con el fichero original usando Matlab.

V. RENDIMIENTO ACADÉMICO EN LA ASIGNATURA DE FST

Con el fin de evaluar el impacto que ha tenido la introducción de varias prácticas basadas en SDR en la asignatura de FST se ha aprovechado la última sesión de laboratorio para organizar un breve debate entre los estudiantes para recoger sus impresiones sobre el desarrollo del laboratorio. El sentir general de los estudiantes era que el hecho de realizar una prácticas más aplicadas hacía más ameno el proceso de aprendizaje. Si bien algunos reconocían que en algunas ocasiones les resultaba un poco más complicado entender los resultados que estaban obteniendo debido al hecho de que era la primera vez que trabajaban con GnuRadio. Aproximadamente el 70 % de los estudiantes eran partidarios de seguir empleando el GnuRadio, un 20 % proponían aumentar el número de sesiones de carácter más aplicado, y el 10 % preferían seguir empleando Matlab como entorno de trabajo para las sesiones de laboratorio.

La mayoría de los estudiantes coincidían en el hecho de que trabajar los mismos conceptos, primero de forma teórica con Matlab, les ayudaba a comprender lo que

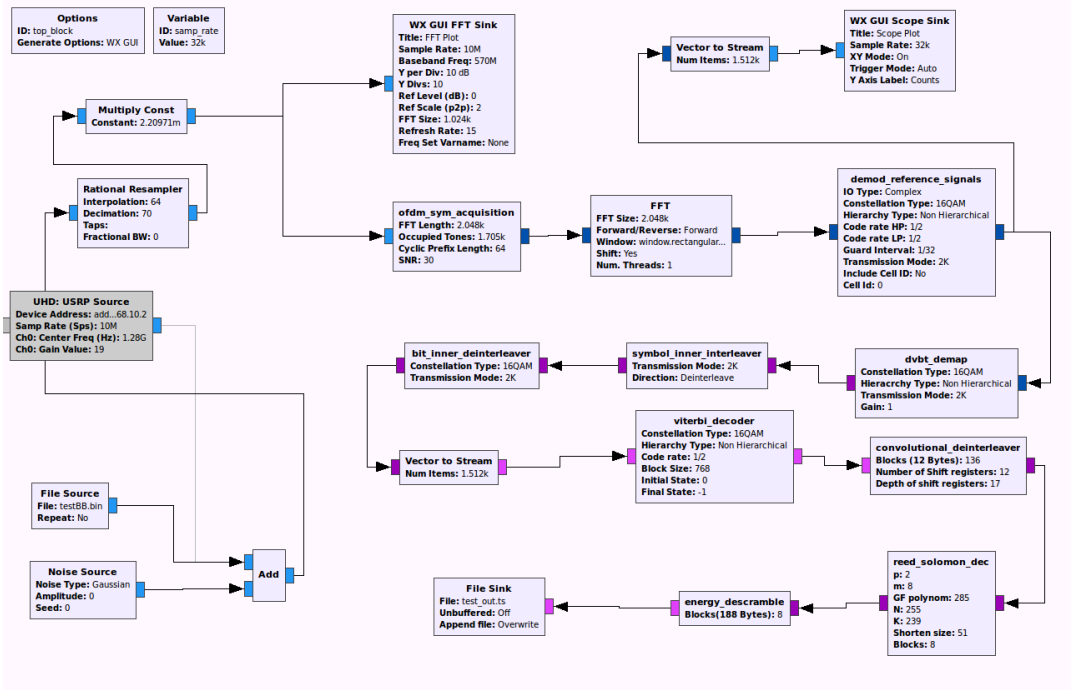


Fig. 11. Implementación del receptor DVB-T en GNURadio-Companion

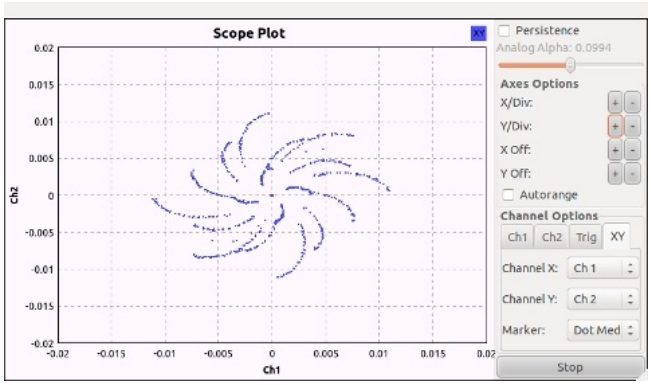


Fig. 12. Scope Plot en el receptor DVB-T.

Tabla I
NOTAS PROMEDIO EN CADA UNA DE LAS SESIONES DE LABORATORIOS EN LOS CURSOS 15/16 Y 16/17. EN LAS SESIONES L1, L2, L3 SE TRABAJAN LAS MODULACIONES ANALÓGICAS MEDIANTE SCRIPTS MATLAB, Y EN L4, L5 Y L6 INCORPORAN CONCEPTOS DE SDR.

	L1	L2	L3	L4	L5	L6	L7	Promedio
15/16	7.4	6.7	6.4	6.6	7.3	6.7		6.9
16/17	7.3	7.5	7.2	7.8	7.5	6.3	7.0	7.2

observaban con el GnuRadio. De los 37 alumnos que participaron en el laboratorio, 5 manifestaron que no les había ayudado el hecho de trabajar primero los conceptos de forma teórica con Matlab; y 2 de ellos propusieron realizar primero las sesiónes de tipo aplicado con GnuRadio, para posteriormente profundizar con más detalle en los aspectos teóricos.

Además de recoger las opiniones de los estudiantes en el debate, en las Tabla I y Tabla III se muestran, para los

Tabla II
RELACIÓN ENTRE EL VALOR NUMÉRICO DE LA NOTA (Q) Y LA CALIFICACIÓN.

Calificación	Valor numérico
A	$Q \geq 8.5$
B	$8.5 > Q \geq 7$
C+	$7 > Q \geq 6$
C-	$6 > Q \geq 5$
D	$Q \leq 4$

cursos 2015/2016 (sin SDR) y 2016/2017 en el que se ha introducido SDR, las notas promedio obtenidas en cada sesión de laboratorio y la frecuencia de las calificaciones obtenidas, respectivamente. Es necesario indicar que en el curso 2015/2016 se realizó una sesión menos de prácticas por problemas de horario. Al comparar las calificaciones promedio mostradas en la Tabla I se aprecia una ligera mejora de 0.3 puntos en el curso 2016/2017. Pero más significativo que esa mejora de la calificación promedio obtenida por los estudiantes es el hecho de que en el curso 2015/2016 se aprecia una caída en el rendimiento de los estudiantes en la segunda sesión de laboratorio, mientras que en el curso 2016/2017 esa caída se aprecia en la sesión L6, casi al final del semestre.

En el caso de los estudiantes de FST, en el segundo curso del grado de Telemática, la realización de las sesiones prácticas usando conceptos basados en SDR ha evidenciado una ligera mejora en el curso 16/17 como muestra la tabla I.

Para analizar la distribución de calificaciones conseguidas por los alumnos se han categorizado los valores numéricos de acuerdo con los valores indicados en la Tabla II. En la valoración del laboratorio se tienen en cuenta aspectos como la actitud o el esfuerzo

Tabla III
ANÁLISIS DE FRECUENCIAS DE LAS NOTAS DEL LABORATORIOS DE FST EN LOS CURSOS 15/16 Y 16/17.

Curso	A	B	C+	C-	D	NP
15/16	4	18	4	6	3	8
16/17	5	22	7	2	1	4

del alumno por lo que las notas suelen ser mejores que las de un exámen teórico. Si el trabajo del alumno es el adecuado debería obtener una calificación superior a 6, por ese motivo se ha fijado en el 6 la calificación C+.

En la Tabla III se muestra el número de alumnos que han obtenido cada calificación. Al comparar los resultados obtenidos en ambos cursos se aprecia cómo las prácticas de laboratorio basadas en SDR han favorecido a mejorar el número de alumnos en los tramos de calificaciones más altos. Especialmente significativo es la disminución de alumnos con calificaciones bajas. La calificación C- suele estar asociada a alumnos que les ha costado seguir el ritmo de la clase y la D suele corresponder a alumnos que inicialmente siguen las clases pero con el tiempo terminan abandonando.

En el caso del número de aprobados, también se evidencia una sensible mejora en el curso 16/17, como muestra la tabla III. Como mostró una discusión final con los estudiantes, esta mejora puede atribuirse directamente a la introducción de conceptos de SDR ya que les permitió una motivación añadida.

VI. CONCLUSIONES

En este artículo se ha explorado la situación docente de las asignaturas de comunicaciones en el Grado de Ingeniería Telemática y en el Máster de Telecomunicaciones de la ETSE de la Universitat de València. La mejora de esta situación pasa por un incremento de la experimentalidad en la docencia. Para el aumento de la carga experimental, se ha propuesto la introducción de algunas sesiones basadas en sistemas SDR en los laboratorios de estas asignaturas. Para el desarrollo de éstas, se han elegido las plataformas USRP y RTL-SDR por su versatilidad.

Este proyecto ha iniciado su aplicación en el curso 2016-2017 en la asignatura de Fundamentos de Sistemas de Telecomunicación, y aún no ha acabado, ya que en este curso se encuentra en fase de evaluación. Actualmente se están diseñando las sesiones prácticas para la asignatura de Transmisión de Datos (y Digital Communication Theory en el máster) que tendrán su aplicación en el curso 2017-2018. La respuesta del estudio final de los estudiantes del grado, que analiza la apreciación subjetiva después de la realización de las experiencias con USRPs/RTL-SDR, aún no ha sido recogida.

A partir de las opiniones de los estudiantes y del análisis de las calificaciones obtenidas se puede concluir que la introducción de las sesiones de laboratorio basadas en SDR han contribuido a mantener el interés del estudiante a lo largo del cuatrimestre. Esto ha facilitado que algunos estudiantes a los que habitualmente les supone un mayor esfuerzo seguir la asignatura hayan podido completarla con éxito. La introducción de SDR en el laboratorio ha mejorado el interés de los estudiantes en seguir la asignatura. El mayor interés no ha supuesto una gran mejora de la calificación promedio obtenida por los estudiantes. También ellos han manifestado en el debate mantenido al final del periodo lectivo que trabajar los diferentes tipos de modulación primero de forma teórica y posteriormente de forma aplicada les ha ayudado a entender los conceptos planteados. Por este motivo, para cursos sucesivos se prevee seguir utilizando éstas técnicas de SDR, pero no incorporar nuevas prácticas en el laboratorio de FST. Para intensificar el uso de SDR en el grado de Ingeniería Telemática se trabajará en la incorporación de este tipo de prácticas en otros laboratorios de la titulación como Transmisión de Datos, Teoría de las Comunicaciones, o Comunicaciones Inalámbricas.

AGRADECIMIENTOS

Los autores quieren agradecer al Servei de Formació Permanent i Innovació Educativa - Centre de Formació i Qualitat "Manuel Sanchis Guarner" de la Universitat de València por la ayuda concedida para el soporte parcial de este proyecto de innovación docente (ref: UV-SFPIE_RMD15-314373) y la Ministerio de Economía e Innovación por la ayuda con referencia TEC2013-47141-C4-4-R con la que se ha financiado parcialmente este trabajo.

REFERENCIAS

[1] I. Blázquez. “Propósitos formativos de las nuevas tecnologías de la información y la comunicación en la formación de maestros”, en «Nuevas tecnologías de la información y la comunicació», Editores F. Blázquez, J. Cabero, F. Loscertales. Publ. Alfar (Sevilla), pp 257-268. 1994.

[2] J.I. Aguaded-Gómez. “Aprender y enseñar con las tecnologías de la comunicación”. Ágora digital 1, 2001.

[3] J. Mitola. “Software radios-survey, critical evaluation and future directions”. Proc. IEEE National Telesystems Conference (NTC’92), pp.15–23, Washington DC, USA, May 1992.

[4] Sh. Mao, Y. Huang, Y. Li, Pr. Agrawal. “Introducing Software Defined Radio into Undergraduate Wireless Engineering Curriculum through a Hands-on Approach”. Proc. of the 120th ASEE Annual Conference & Exposition, June 23-26, 2013.

[5] EN 50067:1998. “Specification of the Radio Data System (RDS) for VHF/FM sound broadcasting in the frequency range from 87.5 to 108.0 MHz’. http://www.interactive-radio-system.com/docs/EN50067_RDS_Standard.pdf (visitado en: 20/04/2017)

[6] ETSI EN 300744:2009. “Digital Video Broadcasting (DVB): Framing structure, channel coding and modulation for digital terrestrial television”. http://www.etsi.org/deliver/etsi_en/300700_300799/300744/01.06.01_60/en_300744v010601p.pdf (visitado en: 20/04/2017)

[7] Ettus Research. *Universal Software Radio Peripheral (USRP)*, 2015. https://en.wikipedia.org/wiki/Universal_Software_Radio_Peripheral (visitado en: 20/04/2017).

[8] I. Pinar-Domínguez, J.J. Murillo-Fuentes. *Laboratorio de Comunicaciones Digitales Radio Definida por Software*. TSC - Ed Universidad de Sevilla, 2011. http://personal.us.es/murillo/docente/Libros/LibroSDR.htm (visitado en: 20/04/2017).

[9] A. M. Wyglinski, D. Pu, D. J. Cullen. “Digital Communication Systems Education via Software-Defined Radio Experimentation”. Proceedings of the 118th ASEE Annual Conference and Exposition Vancouver, BC, Canada, 26-29 June, 2011. http://courses.washington.edu/ee420/index.html (visitado en: 20/04/2017)

JITEL 2017
XIII Jornadas de Ingeniería Telemática
27-29 Septiembre 2017, Valencia (España)

Actas de las XIII Jornadas de Ingeniería Telemática (JITEL 2017), Valencia (España), 27-29 de Septiembre de 2017.

ISBN: 978-84-9048-595-8
DOI: http://dx.doi.org/10.4995/JITEL2017.2017.6571

Equipamiento de laboratorio para mejorar el aprendizaje en comunicaciones móviles

Almudena Díaz Zayas, Pedro Merino Gómez
Departamento de Lenguajes y Ciencias de la Computación,
Universidad de Málaga, Andalucía Tech,
Edificio de Investigación Ada Byron, Málaga, España, 29071.
almudiaz@lcc.uma.es, pedro@lcc.uma.es

F. Javier Rivas Tocado
Keysight Technology
Málaga, Spain
Email: javi_rivas@keysight.com

***Resumen*—Se espera que las generaciones venideras de graduados e investigadores de ingeniería desarrollen un conjunto integral de habilidades para adaptarse a la industria de las comunicaciones móviles y a las expectativas del mercado. Para satisfacer estas demandas las universidades tendrán que colaborar e innovar en sus estilos de educación, herramientas de investigación y procesos de aprendizaje. La evolución de la experimentación remota y la disponibilidad de sistemas avanzados de pruebas para comunicaciones móviles ofrecen nuevas oportunidades, como la experimentación con la clase Gigabit LTE-A. En este artículo discutimos cómo un emulador de LTE-A podría ser adoptado con fines educativos y cómo se pueden usar tecnologías de acceso a entornos de investigación federados para proporcionar un acceso remoto y controlado.**

***Palabras Clave*—comunicaciones móviles, docencia, emulador de estación base, laboratorios remotos, experimentación**

I. INTRODUCCIÓN

La rápida evolución y adopción de LTE y la nueva especificación 5G New Radio (NR) son indicativas de la velocidad con la que se avanza en el sector de las tecnologías móviles. Cada nuevo desarrollo tecnológico trae consigo nuevos desafíos que las nuevas generaciones de ingenieros tendrán que afrontar. Por ejemplo, los dispositivos móviles, las aplicaciones y los protocolos tienen que lidiar con velocidades de datos muy elevadas (superior a 10GHz) que pueden requerir nuevos diseños y arquitecturas, tanto hardware como software. En este sentido, los recursos educativos disponibles en las universidades deben permitir desarrollar estas nuevas competencia. En un campo en el que la tecnología evoluciona de una forma tan rápida es fundamental que los futuros ingenieros reciban una formación lo más actualizada posible y que los investigadores tengan acceso a instrumentos de vanguardia para obtener resultados relevantes.

En [1] Castro et al. se lleva a cabo un análisis de la evolución y avances que han tenido lugar en las técnicas educativas utilizadas en el ámbito de las titulaciones de ingeniería en la última década. En su estudio destacan que, en el pasado, las Universidades habían exhibido un cierto inmovilismo en atender las necesidades específicas de la industria. Sin embargo, se está produciendo un cambio de paradigma debido al uso de Internet y de las tecnologías móviles para proporcionar oportunidades de aprendizaje ubicuas y a lo largo de toda la vida profesional centradas en el estudiante.

La necesidad de conseguir una educación actualizada y orientada a las necesidades vigentes en la industria no es específica del dominio de las telecomunicaciones. TATU [2] es un ejemplo de un proyecto del programa TEMPUS de la Unión Europea que apoya la modernización de la educación superior en paises socios de la Unión Europea y pertenecientes a zonas de Europa Oriental, Asia Central, los Balcanes occidentales y la región mediterránea. El proyecto TATU tiene por objetivo mejorar la empleabilidad de los graduados universitarios a través del establecimiento, en las universidades, de centros de entrenamientos especializados equipados con modernos laboratorios de tecnologías punteras en el sector de la automoción.

Por otro lado, y ya en el contexto de las comunicaciones inalámbricas, el trabajo presentado en [3] es un ejemplo significativo de la integración de sistemas de gestión de aprendizaje tradicionales con el acceso remoto a instrumentación real ubicada en laboratorios de comunicaciones inalámbricas.

La experimentación remota a distancia está recibiendo cada vez más atención ya que maximiza el uso de los equipos, aumenta la disponibilidad de los laboratorios y permite mejorar el aprendizaje a distancia proporcionando entornos para la ejecución de prácticas [1][3][4][5][6]. Existen diferentes tipo de laboratorios remotos, muchos de ellos basados en experimentos simulados mediante entornos web virtuales. El uso de simuladores es útil para reducir costes, pero existen argumentos sobre su

Fig. 1. UXM wireless test set

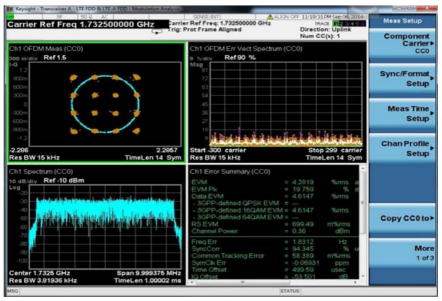


Fig. 2. Herramientas de medidas XApps

efectividad comparada con el acceso a equipamiento real [7], por lo que debería combinarse con laboratorios reales para conseguir entornos de aprendizaje para estudiantes más efectivos [4].

En ese sentido existen también laboratorios de experimentación remota que incluyen instrumentos reales como generadores de señal y analizadores de señal que permiten generar y medir señales de RF [8]. En [5] se resaltan los beneficios de los laboratorios que proporcionan acceso desde Internet para conseguir un eficiente uso de los equipos de radiocomunicación. Como parte de este trabajo los autores han realizado también una evaluación de los efectos positivos que tienen sobre el aprendizaje el uso de laboratorios remotos y los beneficios que tienen para los estudiantes en términos de flexibilidad y mayor disponibilidad.

Sin embargo, los autores del presente trabajo no tienen constancia de la existencia de ningún laboratorio de educación remota que proporcione acceso a emuladores de redes móviles que permiten configurar y analizar de cerca el funcionamiento real de los teléfonos móviles en un entorno realista y a la vez controlado.

El resto del artículo está organizado de la siguiente forma. La sección II proporciona una descripción detallada de la principales características proporcionadas por el UXM Wireless Test Set, y cómo éstas pueden ser utilizadas para mejorar el aprendizaje de los alumnos en asignaturas de comunicaciones móviles. En la sección III se muestra cómo la Universidad de Málaga ha integrado dicho sistema en el testbed PerformNetworks el cual puede ser accedido remotamente, a través de tecnologías y protocolos FIRE (Future Internet Research y Experimentación), tanto para experimentación como para educación.

II. Emulación de redes móviles

En este artículo planteamos el uso de sistemas avanzados de pruebas para comunicaciones móviles como el UXM Wireless Test Set de Keysight Technologies, que se muestra en la Figura 1 en labores docentes y de investigación. Estos equipos son, tradicionalmente, usados por fabricantes de terminales móviles, operadores y laboratorios de certificación para ejecutar las pruebas de conformidad definidas por organismos de estandarización como el 3GPP (3rd Generation Partnership Project) y que deben pasar los terminales móviles antes de ser lanzados al mercado. Estos equipos se comportan como estaciones base reales desde el punto de vista de los teléfonos móviles que se están probando, además de proporcionar una gran variedad de funcionalidades destinadas a proporcionar un entorno de pruebas de I + D potente y flexible.

A. Uso del UXM en actividades de laboratorio

El UXM es un instrumento extremadamente versátil. Tiene la capacidad de emular varias estaciones base con diferentes tecnologías de acceso radio, incluyendo LTE/LTE-A, WCMS/HSDPA+, GSM/GPRS/EGPRS y TD-SCDMA/HSPA. También puede operar simultáneamente como un emulador de canal radio,ruido y generador de formas de onda arbitrarias para generar interferencias y degradación controlada y como un analizador de señal.

Una de las ventajas más destacadas del UXM es la facilidad de uso del E7530A y E7630A LTE/LTE-A Test and Lab Applications [9] que se ejecutan en el propio instrumento. Para el acceso en modo local el instrumento ofrece un interfaz de usuario basado en ventanas diseñado para ser usado desde la pantalla táctil del instrumento. También se puede acceder a la interfaz gráfica del instrumento remotamente usando el protocolo RDP (Remote Desktop Protocol). Para controlar remotamente el equipo también se puede recurrir a la interfaz SCPI (Standard Commands for Programable Instruments) que ofrece.

A continuación introduciremos brevemente las características claves del UXM y el potencial que éstas tienes para su uso en tareas educativas y de investigación en comunicaciones móviles. Cada característica podría ser la base para el diseño de diferentes laboratorios experimentales.

1) Análisis de transmisión de RF y banda base: Entre sus muchas capacidades, el UXM proporciona soporte nativo de las soluciones de medida Keysight, que son el estándar de-facto para análisis de señal. A través de ellas es posible introducir a los estudiantes a multitud de conceptos, como la medida de Magnitud del Vector Error (EVM), análisis de espectro, ecualización y planicidad del canal, análisis de modulación IQ y muchos más. La Figura 2 muestra una medida de análisis de modulación de una señal de 20MHz con modulación 16QAM en el enlace

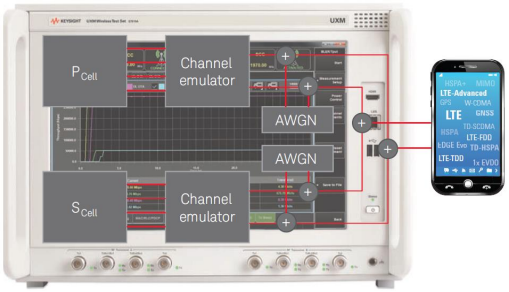


Fig. 3. Emulation de canal digital

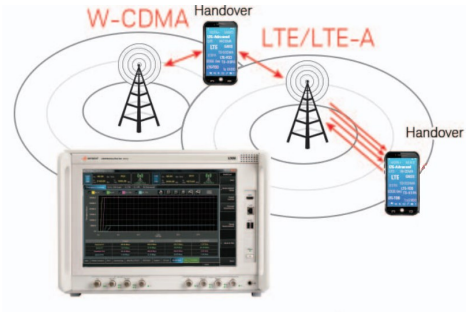


Fig. 4. Procedimientos de movilidad

ascendente.

2) Emulación de canal y Análisis de rendimiento de recepción: Los receptores de los dispositivos móviles se evalúan típicamente en términos de la probabilidad de recepción de datos en presencia de interferencia y degradación. El porcentaje de error resultante se chequea contra umbrales de evaluación definidos.

Para replicar el efecto de las condiciones de propagación radio, el UXM incorpora un emulador de canal digital. Esta funcionalidad proporciona gran facilidad de uso y mejora la precisión evitando las contribuciones de incertidumbre típicamente asociadas a un número elevado de interconexiones RF.

Además de la emulación de escenarios de propagación multicamino, se pueden generar degradaciones adicionales de la transmisión incluyendo generación de señales de continua (CW), ruido blanco gaussiano (AWGN) y formas de onda arbitrarias.

La Figura 3 muestra un diagrama lógico de la emulación de canal y la generación de interferentes para una configuración de dos celdas portadoras (CC) con una configuración de antenas MIMO 4x2. Representando 4 antenas de transmisión en la estación base y 2 antenas de recepción en el terminal móvil. Otros escenarios posibles en un mismo UXM incluyen hasta 4CC con 2x2 MIMO o 2CC con 4x4 MIMO.

3) Procedimientos de movilidad: Como se muestra en la Figura 4, es posible generar múltiples celdas en un mismo UXM, incluyendo diferentes tecnologías de acceso radio (RAT), proporcionando un interfaz táctil intuitivo y fácil de usar.

Entender los entresijos de los protocolos de señalización de las redes móviles, puede ser una tarea compleja si no se se cuenta con ejemplos reales de funcionamiento. El uso de un equipo como el UXM puede ser la forma ideal de mejorar la curva de aprendizaje en este tipo de escenarios.

Además de introducir conceptos básicos como la planificación del espectro radio, las bandas de frecuencia y los distintos anchos de canal, es extremadamente sencillo realizar y supervisar procedimientos de movilidad como el registro de red, el deregistro y la reselección de celda. Otros escenarios más avanzados de interés pueden ser los diferentes tipos de traspasos en la misma o distintas bandas de frecuencia, o incluso entre distintas tecnologías radio. Como se ha comentado anteriormente, es posible analizar el impacto del canal radio y de las interferencias de forma realista pero manteniendo control de las mismas.

4) Análisis de rendimiento de comunicaciones IP extremo a extremo: Los teléfonos inteligentes o smartphones hace tiempo que dejaron de usarse únicamente para realizar llamadas de voz. En su lugar predominan actualmente las comunicaciones multimedia basadas en el protocolo IP. Otra característica clave de los usos actuales es la amplia variedad de perfiles de usuarios y de patrones de tráfico.

En este escenario, es importante asegurar que los dispositivos móviles puedan transmitir datos a su máxima capacidad. La capacidad del UXM evoluciona continuamente para ayudar a investigadores e ingenieros a entender las interacciones entre aplicaciones, sistemas operativos y las pilas de protocolos.

Asímismo, es posible entender en profundidad las dinámicas de tráfico E2E mediante el uso de gráficas de monitorización de tiempo real. Este tipo de representaciones permite observar de forma simultánea la evolución tanto de la velocidad de las transmisiones IP como de las tasas de transmisión de capa MAC. La Figura 5 contiene gráficas de transmisión para un escenario de 3CC con 2x2 MIMO y una modulación 256QAM en el enlace descendente. Como resultado, los 6 flujos MIMO resultantes pueden transportar cerca de 600 Mbps.

Este tipo de representación resulta particularmente útil para identificar artefactos de tráfico que pueden relacionar interacciones entre capas que afectan al rendimiento de las comunicaciones a nivel de aplicación.

5) Monitorización y análisis de protocolos: El UXM incorpora una potente herramienta software de monitorización y análisis [10], la E7515A-L01. Este software permite a los usuarios del UXM controlar la generación y filtrado de trazas a través de un interfaz integrado con wireshark.

Además de analizar los protocolos de señalización de capas altas, como los de gestión de recursos radio (RRC) y del núcleo de red (NAS), es posible acceder a la información de control y a los bloques de datos (PDU) del protocolo de acceso al medio (MAC).

La Figura 6 proporciona un ejemplo de intercambio de mensajes RRC, como los que se usan para difundir la información de celda y para establecer las conexiones radio, así como información detallada sobre la operativa

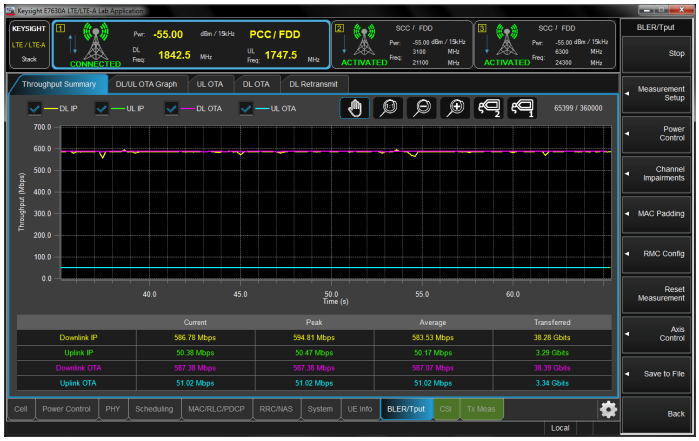


Fig. 5. E2E IP and MAC level throughput graphs

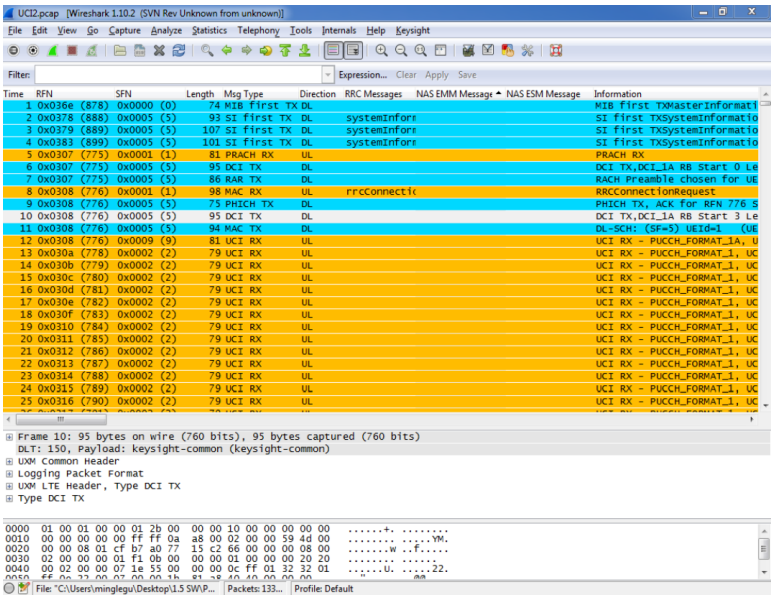


Fig. 6. Software de logging y análisis de protocolos

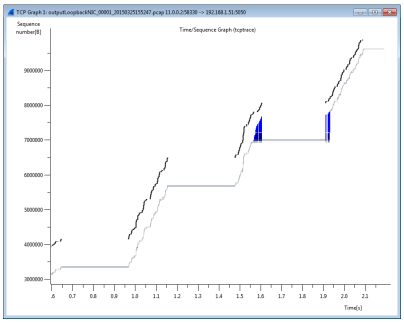


Fig. 8. Integración con herramientas de Wireshark para el análisis de flujos de datos

de los canales de control.

Se pueden analizar escenarios de tráfico complejos, y entender los mecanismos de detección de error y retransmisión de las distintas capas de comunicación, forzando la aparición de errores de comunicación mediante interferentes y canales variantes.

Resulta particularmente interesante el uso de las capacidades integradas de análisis de flujos para estudiar la dinámica de las conexiones TCP a partir de los mensajes MAC_LTE, como se muestra en la Figura 7.

6) VoLTE y Multimedia: El UXM proporciona múltiples funciones de acceso radio que son necesarias para la transmisión de voz sobre LTE mediante el perfil VoLTE [11], incluyendo la planificación semi-persistente (SPS), el uso de múltiples portadoras radio (DRBs), la recepción discontinua (DRX) y otras más.

LTE se basa en una red IP extremo a extremo y hace uso del Subsistema Multimedia IP (IMS) para mantener un registro de terminales (UEs) compatibles con IMS y para establecer sesiones multimedia.

El UXM integra un servidor IMS-SIP [12] que puede combinarse con un cliente software IMCS-SIP para probar escenarios VoLTE y otros servicios multimedia. La Figura

Fig. 7. Herramientas IMS-SIP y VoLTE

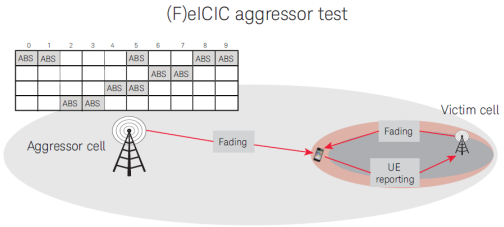


Fig. 9. Redes heterogéneas con eICIC

8 ilustra las diferentes herramientas para prueba de audio sobre comunicaciones IP.

7) Redes heterogéneas: Las redes heterogéneas han sido un campo de investigación con gran actividad en los últimos años. Estos escenarios combinan típicamente múltiples celdas con diferentes potencias, de forma que las distintas celdas se coordinan entre sí. De esta forma, en distintos instantes temporales, unas celdas transmiten mientras que las otras reducen al mínimo la interferencia que generan. Así se consigue mejorar la calidad de las conexiones especialmente en las fronteras entre celdas. Este tipo de técnicas se conoce como Coordinación de Interferencia Entre Celdas Mejorada (eICIC).

Para este fin, la red configura a los dispositivos móviles para que midan la potencia y calidad de señal de distintas celdas en diferentes instantes de tiempo. Se usa un patrón de transmisión especial de baja potencia, también llamado de subtramas casi vacías (ABS), para reducir la interferencia causada a las celdas vecinas tanto en la recepción de datos como en la estimación de canal. Con un correcto alineamiento de los períodos de medida a los patrones de transmisión ABS, los terminales pueden solicitar dos esquemas de modulación y codificación (MCS) diferentes para que la red se ajuste a los distintos niveles de interferencia. Así pues, en las subtramas en las que las celdas vecinas reducen su interferencia, la calidad del canal observado será mejor y será posible utilizar una modulación de mayor capacidad de transmisión.

La Figura 9 muestra un ejemplo de un patrón de interferencia de 40 ms, donde una celda agresora intercala períodos de transmisión a potencia completa con subtramas ABS. La transmisión completa causa una mayor interferencia desde la agresora, mientras que las subtramas ABS permiten al teléfono recibir la señal de la celda víctima con una menor interferencia.

Las especificaciones de prueba del 3GPP definen un conjunto de patrones de referencia y de configuraciones de celda, pero el UXM proporciona un interfaz altamente flexible para explorar el rendimiento de escenarios de

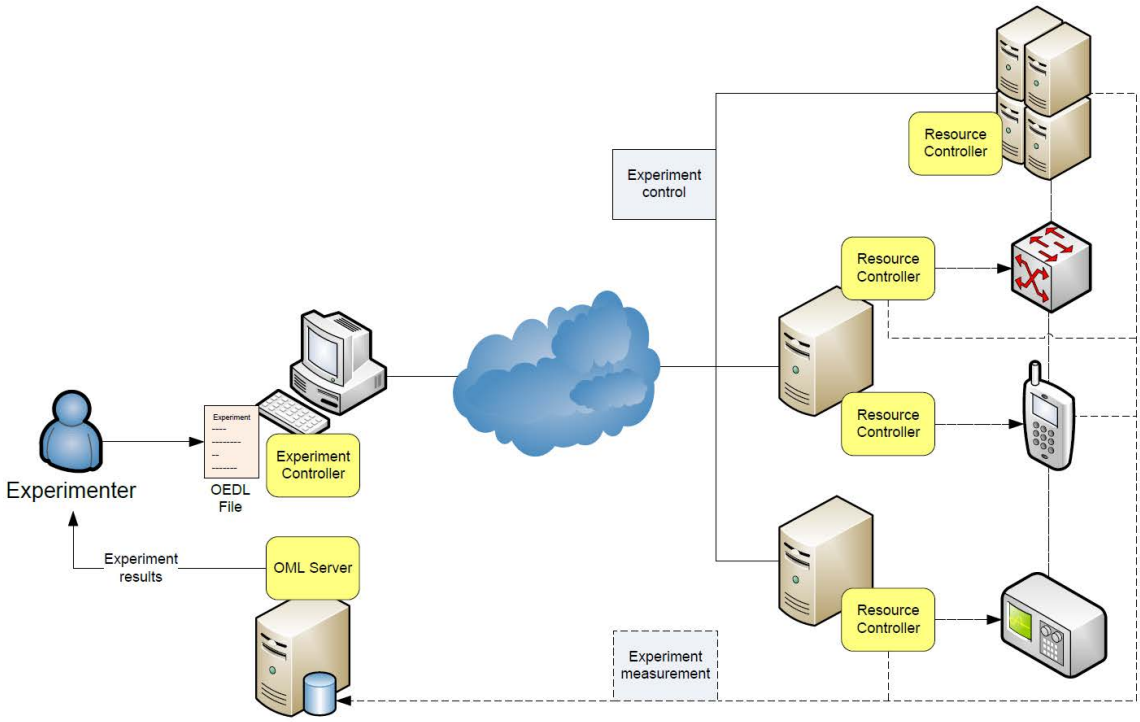


Fig. 10. Arquitectura de un testbed GENI/FIRE

agresión complejos. Es posible definir distintos escenarios ABS para reducir la potencia, donde típicamente se transmite al menos las señales de referencia específicas de la celda (CRS) para que los móviles acampados en la celda puedan funcionar.

III. PerformNetworks, un testbed de experimentación remota

En esta sección se introduce el testbed PerformNetworks [13] del grupo MORSE de la Universidad de Málaga. Dicho testbed está enfocado en la experimentación entorno a las comunicaciones móviles. PerformNetworks proporciona un entorno real y controlado compuesto por el emulador de estación base introducido en la sección anterior, teléfonos móviles comerciales, analizadores de potencia, un núcleo de red LTE y small cells, como elementos más destacados. Se puede consultar la descripción detallada y actualizada del testbed en http://morse.uma.es/performnetworks.

En esta sección se ilustra cómo se ha usado la tecnología FIRE para dar acceso remoto al testbed. La adopción de esta tecnología ha permitido que incorporar el testbed en la infraestructura de investigación europea promovida por el proyecto europeo Fed4Fire. También se explica el flujo de trabajo de la orquestación de un experimento.

A. Tecnologías para el control de recursos de laboratorio

GENI y FIRE son dos iniciativas que tratan de crear un entorno de experimentación común, la primera en Estados Unidos y la segunda en Europa. Ambas iniciativas se basan en la misma idea: la evaluación experimental es un requisito indispensable para proporcionar resultados significativos que puedan ser aplicados en el mundo real. Así, el concepto de testbed experimentales es fundamental tanto en la comunidad GENI como en la comunidad FIRE. Cabe destaca que aunque GENI y FIRE surgieron como dos iniciativa independientes, el proyecto Fed4Fire ha adoptado las tecnologías GENI, haciéndolas compatible con las tecnologías FIRE.

La investigación experimental requiere la capacidad de realizar experimentos de una manera fácil y repetible. En el testbed PerformNetworks se ha adoptado OMF y OML como soluciones para proporcionar control y medición de experimentos, respectivamente. La arquitectura de un testbed genérico basado en estas tecnología se muestra en la Figura 10. Un experimento se define como un archivo escrito en el Lenguaje de descripción de experimentos de OMF (OEDL). Un archivo OEDL declara los recursos que utilizará el experimento, los eventos a los que reaccionará y las acciones a realizar en respuesta a dichos eventos.

El controlador del experimento (EC) interpreta los scripts OEDL y coordina la ejecución del experimento. Cada recurso del testbed es administrado por un controlador de recursos (RC), que puede alojarse en un ordenador externo o en el propio recurso. Los RCs y los ECs intercambian información de control utilizando el Protocolo de Control de Recurso Federado (FRCP), que puede ser transportado a través del protocolo XMPP (Extensible Messaging and Presence Protoco) o del protocolo AMQP (Advanced Message Queuing Protocol).

El servidor OML recopila y almacena las medidas del experimento. Cada instrumento, o un RC en nombre de un instrumento, puede enviar las medidas a través de un cliente OML. Para la implementación de dicho cliente

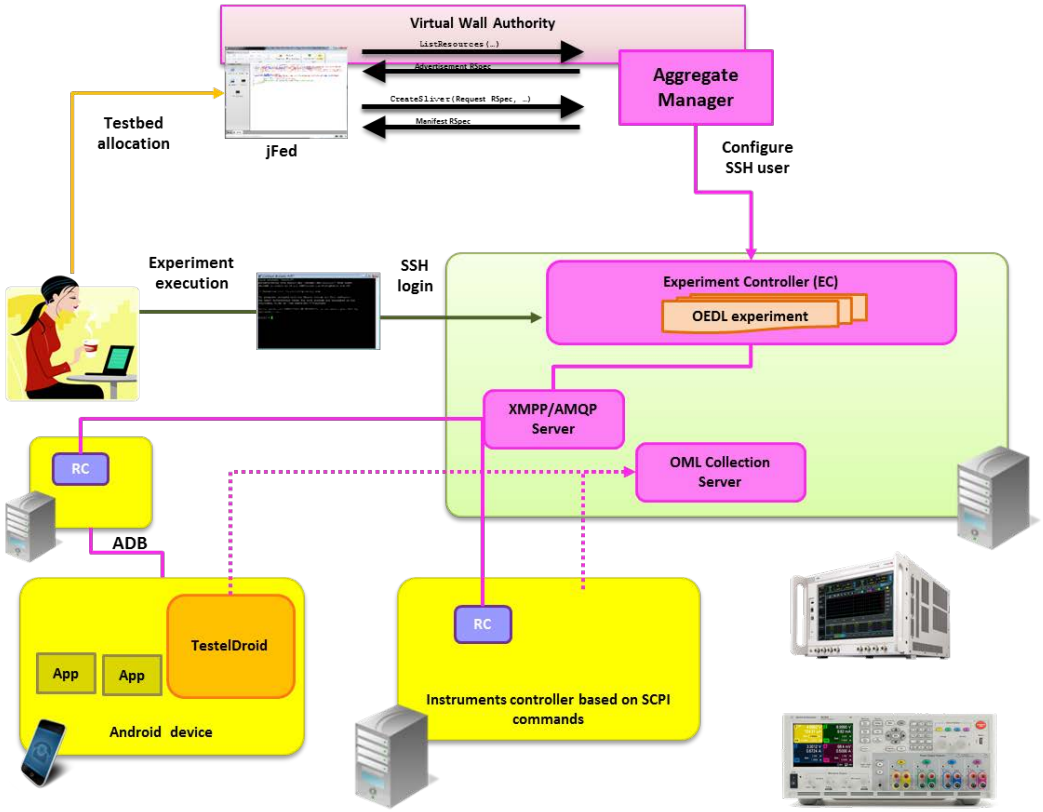


Fig. 11. Acceso remoto y control de la experimentación en el testbed PerformNetworks

Código 1. Ejemplo de script OEDL

```
after 100 seconds do
   info "Consultar la potencia de transmisión"

   group("RC").exec(":echo "CNTRL:LTE:CONNECTion:UL:ULPOWer?" | netcat -q 2
      10.102.81.31 55000")
end

after 110 seconds do
   info "Iniciar una aplicación en CSipSimple en el dispositivo Android
      conectado al emulador de estación base"

    group("RC").exec("C:\\Triangle\\Android\\platform-tools\\adb shell am start
      -n com.csipsimple/com.csipsimple.ui.SipHome")
end
```

se proporcionan librerías en distintos lenguajes: C, Ruby, Java, etc. El servidor OML tiene como backend una base de datos PostGreSQL de forma que las medidas almacenadas pueden ser consultadas utilizando herramientas SQL estándar.

Una vez que se ha definido el experimento, se envía al EC. El EC interactúa con los RCs de los recursos implicados en el experimento, por ejemplo un móvil Android o el emulador de estación base, para ejecutar en ellos las acciones definidas en el script. En el script OEDL también se definen los puntos de medidas del experimento, las medidas serán almacenadas en el servidor OML.

El descubrimiento, la reserva y el aprovisionamiento de los recursos disponibles en el testbed se realiza a través de SFA (Slice-based Facility Architecture) [14]. Para iniciar los procesos de descubrimiento y reserva necesitamos proporcionar un especificación de nuestro recurso denominada RSpec. La entidad definida por el SFA para gestionar las consultas sobre los recursos es el Aggregate Manager (AM). El AM es consultado por una herramienta cliente que implementa el API de comunicación con el AM, en nuestro caso usamos jFed [15].

Código 2. Ejecución de un experimento

```
omf_ec -c config.yml script.rb
```

B. Flujo de trabajo de experimentos ejecutados remotamente en PerformNetworks

El flujo de trabajo para acceder remotamente al testbed se muestra en la Figura 11. El primer paso es obtener un certificado X.509 a través de la autoridad de autenticación de Fed4Fire, accesible en https://authority.ilabt.imind.be. Este certificado es proporcionado a la herramienta jFed para que pueda establecer una comunicación con el AM y reservar un intervalo temporal para acceder al testbed. Una vez realizada la reserva del testbed el AM de PerformNetworks proporciona un acceso SSH federado al EC que gestiona los recursos del testbed.

Se proporcionan scripts de referencia que muestran cómo controlar el UXM y las aplicaciones que se ejecutan en el móvil (ver extracto de script que se muestra en Código 1). En los scripts se usan los procedimientos disponibles en el RC asociado a cada recurso del testbed para enviarle, en el caso de los instrumentos, los comandos SCPI definidos en el manual de usuario, o , en el caso de los dispositivos móviles Android, comandos ADB, que permiten configurar el móvil y lanzar aplicaciones.

Una vez editados los experimentos pueden ser ejecutados usando el procedimiento omf_ec (ver Código (2)) disponible en el EC al cual se tiene acceso mediante SSH.

IV. CONCLUSIONES

Para mantener el ritmo al que las tecnologías móviles evolucionan se necesitan nuevas herramientas y nuevos enfoques en las Universidades y en los centros de investigación. La utilización de equipos como el UXM Wireless Test Set permitirán a los estudiantes desarrollar y mantener sus competencias actualizadas.

En el presente artículo se ha ilustrado la forma de integrar este equipo en un testbed más amplio con acceso remoto. También se ha demostrado su uso remoto a través de los protocolos de experimentación promovidos por la comunidad de testbed europea FIRE.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio Español de Economía y Competitividad (TIN2015-67083-R), FEDER y el programa de investigación e innovación Horizonte 2020 de la Unión Europea (grant agreement No 688719).

REFERENCIAS

[1] Castro M.; Tawfik M.; Tovar E., "Digital and Global View of Engineering Education Using Remote Practical Competences", IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (RITA) (Volume:10 , Issue: 3) pp 126-133. doi: 10.1109/RITA.2015.2452651

[2] Workshop: "The TATU Lab & Smart Education", 2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV), pp 400-402

[3] Kafadarova N.; Mileva N.; Stoyanova S., "Remote Wireless Communications lab in real time". 2013 IEEE Global Engineering Education Conference (EDUCON), pp 69-74

[4] Tawfik M.; Sancristobal E.; Martin S.; Gil R.; Pesquera A.; Albert M.J.; Peire J.; Milev M.; Mileva N.; OSuilleabhain G.; Tzanova S.; Kreiner C.; Hormann L.B; Castro M., "Labor-Oriented Online Master Degree Program", 2013 IEEE Global Engineering Education Conference (EDUCON), pp 1098-1102

[5] Gampe, A; Melkonyan, A; Pontual, M;Akiopian D.; "An Assessment of Remote Laboratory Experiments in Radio Communications", 2014, IEEE Transactions on Education, Volume:57 , Issue: 1 , pp 12- 19

[6] Nassar, A.;, Mohammed, M.; Elrashidi, A.; Elleithy, K.; "Virtual Wireless and Mobile Communication Laboratory, Education", Vol. 2 No. 1, 2012, pp. 19-24. doi: 10.5923/j.edu.20120201.04.

[7] Feisel, L. D.; and Rosa, A.J.; "The role of the laboratory in undergraduate engineering education", Journal Eng. Edu., vol. 94, no. 1, pp. 121-130

[8] Kara, A.; Aydin, E.U.; Oktem, R.; Cagiltay, N., "A Remote Laboratory for Training in Radio Communications: ERRL," in Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on , vol., no., pp.1-5, 3-7 Sept

[9] Keysight Technologies. E7515A UXM Wireless Test Set [Online]. Available: www.keysight.com/find/uxm, visited on 1 January 2016.

[10] Keysight Technologies, E7515A-L01 Protocol logging and analysis software, User's guide. 2014

[11] Official Document IR.92 - IMS Profile for Voice and SMS. GSM Association. 2015.

[12] Keysight Technologies,E6966B IMS-SIP network emulator, Technical Overview. 2015

[13] Díaz Zayas, A.;Recio Perez A.M.; García Perez, C.A.; Merino P, PerformLTE: a Testbed for LTE testing in the Future Internet, 13th International Conference on Wired & Wireless Internet Communications. 2015

[14] Slice-Based Federation Architecture, Version 2.0, GENI Initiative, Tech. Rep., Jul. 2010. [Online]. Available: http://groups.geni.net/geni/ wiki/SliceFedArch

[15] B. Vermeulen, W. Van de Meerssche, T. Walcarius, jFed toolkit, Fed4FIRE, Federation; GENI Engineering Conference (GEC) , 2014

Índice de Autores

A

B

C

D

E

F

G

H

I

J

L

M

N

O

P

Q

R

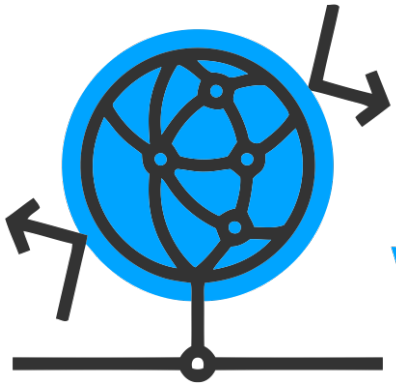
S

T

U

V

Z



JITEL

Valencia, 2017

Colaboradores

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE UPV INGENIEROS DE TELECOMUNICACIÓN

Organizadores

IGIC Instituto de investigación para la gestión integrada de zonas costeras

at

Instituto ITACA Tecnologías de la Información y Comunicaciones