XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

A Novel Admission Control Scheme for Network Slicing based on Squatting and Kicking Strategies

Ahmed El-mekkawi, Xavier Hesselbach, and Jose Ramon Piney
Dept. Network Engineering,
Universitat Politecnica de Catalunya
C/ Jordi Girona, 1-3 - Edif.C3 - Campus Nord - 08034 Barcelona - Spain
(ahmed.mohamed.abdelaty.elmekaw, xavier.hesselbach)@upc.edu, jpiney@entel.upc.edu

***Abstract*—New services and applications impose different quality of service (QoS) requirements on network slicing. To meet differentiated service requirements, current Internet service model has to support emerging real-time applications from 5G networks. The admission control mechanisms are expected to be one of the key components of the future integrated service Internet model, for providing multi-level service guarantees with the different classes (slices) of services. Therefore, this paper introduces a new flexible admission control mechanism, based on squatting and kicking techniques (SKM), which can be employed under network slicing scenario. From the results, SKM provides 100% total resource utilization in bandwidth context and 100% acceptance ratio for highest priority class under different input traffic volumes, which cannot be achieved by other existing schemes such as AllocTC-Sharing model due to priority constraints.**

***Index Terms*—SKM, Admission Control, Class of Service, Utilization Optimization**

I. INTRODUCTION

With the emergence of network slicing in 5G, and network virtualization embedding strategies, resource management models are required to provide 100% utilization in a multi-class context under bandwidth constraints [1]. Furthermore, as the demand for different types of services and applications increases, integrating the services into the Internet will have a profound influence on the future extension of Internet networking technologies. Hence, the diversified applications with different QoS requirements are considered to be the most important components of the future IP services under 5G networks [2] [3]. Under the motivation of the rapid growth of real-time service requirements, the current Internet is smoothly shifting from the best-effort network into an integrated services network, little by little. Recently, more and more emerging Internet real-time applications that, require more than best-effort service are increasingly being carried out on the Internet [4] [5]. Moreover, the network operator wants to maximize the revenue by increasing the number of users without compromising the promised Quality of service. This can only be achieved by efficient admission control model that directly controls the number of users admitted into the system. In this regard, Bandwidth Allocation Models (BAMs) that have been proposed in the past to set application requirements and priorities over a range of traffic classes, can serve as models for admission control. BAMs establish the amount of bandwidth per-class and any eventual sharing among them [6]. Moreover, BAMs can handle any type of resources allocation [7]. In the literature, several works deal with the dynamic bandwidth allocation for guaranteeing a given QoS level per class and optimizing the utilization. These contributions are based on the Maximum Allocation Model (MAM) [8], Russian Doll Model (RDM) [9], Generalized RDM (G-RDM) [10], AllocTC-Sharing model (AllocTC) [11] among others. Fig. 1 illustrates examples of MAM, RDM, G-RDM and AllocTC allocation algorithms for three CTs, where the RC (resource constraints) value corresponds to the bandwidth restriction (limit) imposed to one or more CTs. MAM is a strict

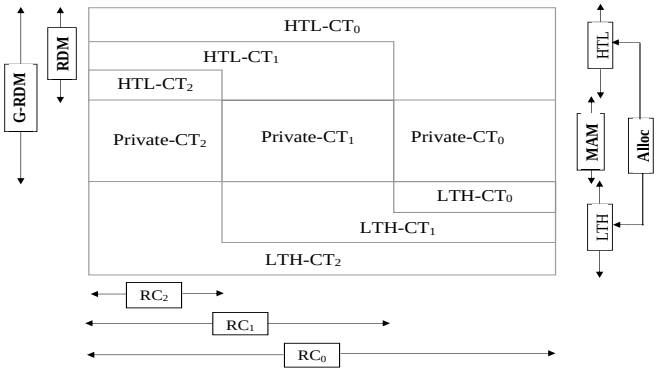


Fig. 1: BAMs and resource allocation strategies [12]

.

allocation model in which another class type (CT) cannot share (private resources) the unused bandwidth of a given CT. On the other hand, RDM is a nested allocation model where non-utilized bandwidth allocated to the higher hierarchical CTs might be used by lower priority CTs temporarily (High to Low loan - HTL loan). Moreover, AllocTC model allows an opportunistic sharing of the bandwidth between the different classes. It is considered as an enhancement of the RDM model because it not only allows an HTL loan but Low to High loan

(LTH loan) as well. G-RDM is a hybrid model in which the "HTL loan" strategy of RDM incorporates the private resource strategy defined by MAM. However, these models do not take into account various Service Level Agreements (SLA) such as latency, packet loss, and jitter to adjust bandwidth, and they can not guarantee higher admission for high priority classes after network congestion. Therefore, the main contribution of this paper proposes to integrate all of these models in a single admission control model, in multi-class networks being able to provide 100% total resource utilization based on squatting and kicking strategies that can work under offline and online scenarios. In offline scenario, all demands are known in advance without lifetime constraint, while in online scenario, demands arrive on a real-time basis with a specific lifetime. SKM, guarantees high admission for QoS of higher priority classes under different input traffic volumes, especially in congested scenarios (i.e. such as video, if it is more important than others in a network, then by using the SKM, a network administrator can prioritize video traffic to ensure that the service remains uninterrupted, while the other traffic may be suspended or even dropped). On the other hand, for the case of uncongested scenarios, the SKM behaves similar to MAM, RDM and AllocTC.

Moreover, SKM is a suitable strategy for emerging technologies that are charactered by diverse QoS requirements and prioritized admission control. The concept of QoS allows certain types of traffic to be prioritized in the network. A case at hand will be network slicing scenario, where the different slices have varying priorities in terms of admission and resource allocation.

The remainder of this paper is organized as follows: In section II, related works are listed. In section III, we present the definition and the description of SKM proposal, including SKM scenarios in offline and online mode. Section IV describes performance evaluation issues. Section V presents the obtained results and discussion. Section VI concludes the paper and presents future work.

II. RELATED WORKS

BAMs are of great value in the context of efficient and customized use of resources management. Moreover, BAMs can work as admission control models. Several works based on BAMs dealt with the dynamic bandwidth allocation for guaranteeing a given QoS level per CT and optimizing the utilization. In [13], the authors propose a method of dynamic and hierarchical allocation of the bandwidth using RDM strategy. This method is based on the classification and the prioritization of services. The algorithm provides the bandwidth required for the demands based on fairness factor and services priority.

The general problem of the algorithms based on RDM is that the resource reservation is carried out from bottom to top; the lower priority traffic shares its resources with higher priority traffic and not the inverse. Several works have been carried out proposing new dynamic bandwidth sharing algorithms by adopting the RDM strategy [14] [15].

To make the reservation from top to bottom and from bottom to top, the AllocTC [11] initiated two-way algorithm of dynamic bandwidth sharing, where unused bandwidth of high priority CTs can be shared with low priority CTs. In [7] the authors studied the behaviour and resource allocation characteristics of the BAMs, then they compared distinct BAMs using different traffic scenarios. The authors proved by simulation that AllocTC is more efficient in terms of optimizing the utilization of the link and that it is better suited for elastic traffic and high bandwidth utilization. The authors in [12] propose a new approach with a combination of (MAM, RDM, G-RDM, and AllocTC) models based on a controller by using different metrics to switch from one model to another one in order to improve the efficiency of the performance for instance link utilization, blocking probability, and packet number. In [16], the authors proposed a new model called (smart AllocTC), which runs on a controller to manage the QoS and routing with QoS constraints. The model applies RDM and AllocTC strategies to classify demands based on their threshold severity (high, medium, and low). Whenever the priority of demand is of the high threshold, the (smart AllocTC) benefits from other categories bandwidth and calculates the fairness index of the categories to allocate resources precisely to all demands taking into account their priorities.

However, all these models cannot give 100% total resource utilization and guarantee higher admission for higher priority classes at same time.

III. SQUATTING AND KICKING MODEL (SKM) PROPOSAL

The need for network slicing and network virtualization for 5G networks requires an admission control model that can support fast and dynamic discovery of the resources that will often be heterogeneous in type, implementation, and independently administered. Thus, the main idea of our proposed admission control model exploits resources partition and reservation, according to different priority classes with the flexibility of using the full amount of resources when other CTs do not demand them. Furthermore, SKM provides a smoother BAM policy transition among existing policy alternatives resulting from MAM, RDM, AllocTC adoption independently in a single solution, to improve the utilization and to guarantee high admission for the higher priority CTs. This strategy is used as an admission control function for highly congested scenarios, with strict constraints for the higher priority CTs. On the other hand, for the case of uncongested scenarios, then the SKM behaves similar to classical BAM techniques.

A. Definitions

Traffic Classes - TC (also CT or class or class of service COS) according to RFC 4127: is a logical group of demands that meet a given resources constraint, such as equal value in a specific header field (e.g. source-destination) [9]. TC populate the so-called multi-class networks.

Squatting: action of occupying resources allocated to other (higher or lower) classes when their holders are not using them.

Kicking: action of expelling a lower priority class from its allocated resources, either partially or totally [17].

B. Assumptions and Notations

The goal of the auto-provisioning, SKM model is to achieve more efficient admission control mechanism for prioritized

user demands. The proposed model jointly considers the priorities of both admitted and arriving demands and the current resource utilization in the system. In this work, the contested resources of single link can support up to R, which represents the capacity of the resource of the system; the size of the R can be discrete or continuous. R is partitioned in classes, N is the number of classes defined in the link, and where RCe is the maximum reservable resources in class e, as shown in Fig. 2.

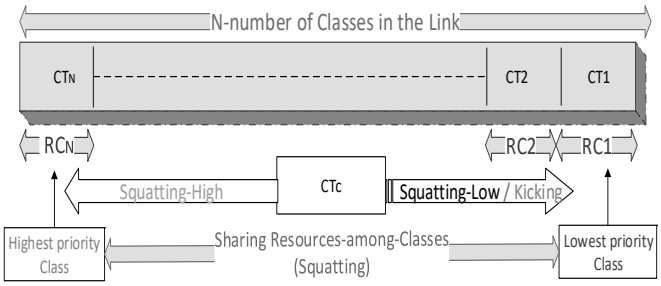


Fig. 2: SKM-Strategy

C. Algorithm Setting

A description of all parameters and decision variables used in our admission control is provided in Table I and Table II respectively.

TABLE I: Parameters of the Model

Abbreviation	Explanation
RC_c	Resource Constraints for class c also equal to maximum reservable resources for class c
CT_c	Class of priority c where $c \in [1, N]$ and CT_N is the highest priority class and CT_1 is the lowest priority class.
R	Maximum allocable resources for all classes together and is equal to link capacity
$d_j(CT_c)$	The amount of resources (size) of demand j belonging to class c where $j \in [1, D]$

TABLE II: variables of the Model

Abbreviation	Explanation
D	Total Number of demands by all classes
D_c	Total Number of demands by class c
S_c	The actually allocated resources to class c
BD	Number of blocked demands by all classes
BD_c	Number of blocked demands by class c
AD	Number of accepted demands by all classes
AD_c	Number of accepted demands by class c
P_{LTH}	The number of preemption of higher priority traffic by lower priority traffic
P_{HTL}	The number of preemption of lower priority traffic by higher priority traffic
SH_i	Squatted resources from higher priority class i
SL_i	Squatted resources from lower priority class i
K_i	Kicked resources from lower priority class i

D. Conceptual model behavior

Different strategies such as Squatting model, MAM, RDM, GRAM, AllocTC and others can be considered, depending on performance and goals provided by each strategy. In our proposed model, the sharing approach already used by MAM, RDM, AllocTC and the SKM allows CTs with higher priority to use available resources allocated to lower priority CTs and vice versa. Unlike other BAMs, in SKM, if a given CT of service requires more resources than those allocated to it, the procedure of the model, for each demand, will be as follows:

- SKM starts working as a normal MAM algorithm (step 1).
- If resources are not enough, SKM check where resources are not used, starting with higher priority classes (Step 2). This is similar to squatting of the higher priority classes (Sq-H) or RDM style.
- Else, if more resources are required, SKM check where resources are not used from lower priority classes (Step 3). This is similar to squatting of the lower priority classes (Sq-L) or loan of lower priority traffic by higher priority traffic (AllocTC style).
- Else, try using Kicking and count the kicked class in the blocking probability for the same class.
- Else, the demand cannot be allocated.

Based on the service policy, the Squatting technique aided by its two priority classes (high and low) to be less aggressive than kicking technique, especially in case of the uncongested scenarios. Therefore squatting technique is generally preferred over kicking, if the class requires extra resource allocation, as shown in Alg 1.

Algorithm 1 Process Assignment algorithm for SKM

```
procedure PROCESS ASSIGNMENT(Loop D :Demands; Loop Demands)
    for Each Demand d_l = d_l(CT_i) ∈ D do
        if d_l ≤ RC_i then                                   ▷ Strategy MAM
            Allocate d_l resources from the class i
        else if ∃j s.t. j > i ∧ d_l ≤ CT_j Available resources then
    ▷ Strategy RDM or Squatting-High
            Allocate d_l resources from CT_j                 ▷ SH_j
        else if ∃j where j < i s.t. d_l ≤ CT_j Available resources
    then                                                     ▷ Squatting-Low
            Allocate d_l resources from (CT_j)               ▷ SL_j
        else
            found-kick=false
            for j=1 to i-1 do
                if ¬(found-kick) and (∃d_m(CT_n) ∈ (CT_j) , and , n<i)
    then
                    kick d_m(CT_n) from (CT_j)               ▷ found-kick=true
                end if
            end for
            if ¬(found-kick) then
                Reject d_l
            end if
        end if
    end for
end procedure
```

E. Offline and online scenarios

The proposed algorithm in this paper was designed to work as admission control for offline and online scenarios. In offline scenario, all demands are known in advance, and they do not have lifetime constraint (i.e. allocated without expiry limit). While in online scenario, demands arrive on real-time basis with specific arrival and expiry times. The following paragraphs introduce the overall idea of each scenario as follow:

1) Offline scenario: The goal of SKM performance is to make the best selection of user demands to be admitted considering, user priorities and available resources in the system.

The SKM offline behaviour introduces a new method for deciding the demands that can be admitted. In this scenario, we simplify the procedure of checking demands by arranging them according to their priorities and sizes. Based on that, the high priority classes will be allocated to the high priority demands first, and then low priority classes can be allocated to the remaining demands if there are enough resources.

2) Online scenarios: In the SKM performance of the online scenario, the traffic of the system can be distributed fairly according to the QoS policy. This provides efficient usage of system resources and solves the online allocation problems such as the rerouting of the demands according to the priority along the unit times. In the online mode, the demands are sorted according to size and priority to minimize the number of kicking operation. The difference between the SKM behaviour in offline mode and online mode is that in the offline mode the sorting process performed once before the allocation process. In online mode, the sorting is done before the process of the assignment of the demands in each unit time as in Alg 1.

IV. Performance evaluation

This section presents technical comparison of SKM against the state of the art algorithms. Also, an evaluation methodology is presented, which includes performance metrics and descriptions of simulation scenarios.

A. Technical behavior and other operational characteristics

Table III shows a set of possible behaviours and operational characteristics adopted to manage system resources for admitting user demands. In other words, to obtain expected use and accept demands depending on available resources and traffic load using SKM and other comparative models.

B. Offline evaluation metrics

The evaluated metrics for permanent demands addressed in this paper is the total acceptance ratio (AR), total utilization (U), acceptance ratio per class (AR_c) and utilization per class (U_c) according to Table IV as below:

TABLE IV: Offline metrics definitions

Abbreviation	Explanation
Acceptance ratio AR	Is the ratio between the number of accepted demands and the total number of demands Eq. 1
Acceptance ratio per class AR_c	Is the ratio between the number of accepted demands by ($Class_c$) and the total number of demands by this ($Class_c$) Eq. 2
Blocking probability Bp	The ratio between the number of blocked demands (rejected) and the total number of demands
Blocking probability per class Bp_c	The ratio between the number of blocked demands by ($Class_c$) and the total number of demands by this ($Class_c$) Eq. 3
Total Utilization U	The ratio between the accepted resources and the total capacity of resources Eq. 5
Utilization per class U_c	The ratio between the accepted resources by ($Class_c$) and the total capacity of resources by this($Class_c$) Eq. 6

$$AR = AD/D \tag{1}$$

$$AR_c = AD_c/D_c \tag{2}$$

$$Bp = BD/D \tag{3}$$

$$Bp_c = BD_c/D_c \tag{4}$$

$$U = \frac{\sum_{j=1}^{D} d_j(CT_c)\, I_{A(j)}}{R} \tag{5}$$

Where $I_{A(j)}$ is an indicator function equal to 1 if j belongs to A and 0 otherwise. The set A(j) corresponds to total accepted demands.

$$U_c = \frac{\sum_{j=1}^{D_c} d_j(CT_c)\, I_{A_c(j)}}{RC_c} \tag{6}$$

Where $I_{A_c(j)}$ is an indicator function equal to 1 if j belongs to A_c and 0 otherwise. The set $A_c(j)$ corresponds to accepted demands by class c.

1) Example of Proposed Off-line SKM Algorithm: SKM was compared to RDM and AllocTC, in terms of user priorities and available resources in the system. In this example, the resources capacity of the system equal to 40 units and divided into four priority classes. Each class has the same amount of resources equal to 10 units. Nine demands to use available resources (i.e. 10, 10, 10 and 10) must be admitted into the system as follows:

#1: From S to D, 8 units priority 3
#2: From S to D, 4 units priority 3
#3: From S to D, 7 units priority 4
#4: From S to D, 7 units priority 4
#5: From S to D, 9 units priority 1
#6: From S to D, 6 units priority 2
#7: From S to D, 6 units priority 3
#8: From S to D, 7 units priority 2
#9: From S to D, 12 units priority 4

The overall performance of SKM in this example as shown in Table V demonstrates the performance of SKM in the offline case for the demands to be admitted on the given classes of the link. For example, the demand $\#9 : 12_4$ is admitted on the system where it used all resources from its priority class and borrowed two unused resources from class 3. Table VI shows the link load by TC, U_c, U, AR_c and AR results by using offline SKM. Which means, after the admission of the demands, we can calculate the link load for each class, the utilization of each class, and how many admitted or rejected demands in the system.

C. Online evaluation metrics

The metrics for the finite duration demands considered in our work, as defined in Table VII can be evaluated as follows:

TABLE VII: Online metrics definitions

Abbreviation	Explanation
Acceptance ratio AR(T)	The ratio between the number of accepted demands and the total number of demands until time T. Where the observation time (total consumed time by simulation) from t_0 until T Eq. 7
Acceptance ratio per class $AR_c(T)$	The ratio between the number of accepted demands by each class separately and the total number of demands by the same class until time T Eq. 8
Total Utilization: U(T)	The ratio between the accepted resources in all classes within a time duration T_j and the total capacity of resources at the time of observation Eq. 9
Utilization per class: $U_c(T)$	The ratio between the accepted resources by each class separately within T_j and the total capacity of resources of the same class at the time of observation Eq. 10

TABLE III: Technical behavior and operational characteristics comparison matrix

Behavioral characteristics	**MAM**	**RDM**	**AllocTC**	**SKM**
Efficient Resource utilisation with high traffic load of lower priority classes	Low	High	High	High
Efficient Resource utilisation with high traffic load of higher priority classes	Low	Low	High	Very High
Resource utilisation along the link	Low	Low (but better than MAM)	High	High
Accepted demands of higher priority classes along with the link	Low	Low	Low	Very High
Traffic classes isolation	High	Medium	Low	Low
Operational characteristics	**MAM**	**RDM**	**AllocTC**	**SKM**
P_{HTL}	No	Yes	Yes	Yes
P_{LTH}	No	No	Yes	No
K_i	No	No	No	Yes

TABLE V: SKM example (Off-line)

# of demand : d_p 4 priority classes	**Avialable Resources**	**SKM-Allocation**
$\#9:12_4$	(10,10,10,10)	(10,10,8,0) SL_3
$\#3:7_4$	(10,10,8,0)	(10,10,1,0) MAM
$\#4:7_4$	(10,10,1,0)	(10,4,0,0) SL_2
$\#1:8_3$	(10,4,0,0)	(6,0,0,0) SL_1
$\#7:6_3$	(6,0,0,0)	(0,0,0,0) SL_1
$\#2:4_3$	(0,0,0,0)	Rejected
$\#8:7_2$	(0,0,0,0)	Rejected
$\#6:6_2$	(0,0,0,0)	Rejected
$\#5:9_1$	(0,0,0,0)	Rejected
# of demand : d_p 4 priority classes	**Avialable Resources**	**AllocTC-Allocation**
$\#1:8_3$	(10,10,10,10)	(10,10,2,10)
$\#2:4_3$	(10,10,2,10)	(10,8,0,10)
$\#3:7_4$	(10,8,0,10)	(10,8,0,3)
$\#4:7_4$	(10,8,0,3)	(10,4,0,0)
$\#5:9_1$	(10,8,0,3)	(1,4,0,0)
$\#6:6_2$	(1,4,0,0)	(1,0,2,0) P_{LTH}, $\#2:4_3$ Rejected
$\#7:6_3$	(1,0,2,0)	Rejected
$\#8:7_2$	(1,0,2,0)	(0,0,0,3) P_{LTH}, $\#4:7_4$ Rejected
$\#9:12_4$	(0,0,0,3)	Rejected
# of demand : d_p 4 priority classes	**Avialable Resources**	**RDM-Allocation**
$\#1:8_3$	(10,10,10,10)	(2,10,10,10)
$\#2:4_3$	(2,10,10,10)	(0,8,10,10)
$\#3:7_4$	(0,8,10,10)	(0,1,10,10)
$\#4:7_4$	(0,1,10,10)	Rejected
$\#5:9_1$	(0,1,10,10)	(0,0,2,10)
$\#6:6_2$	(0,0,2,10)	(0,0,0,6)
$\#7:6_3$	(0,0,0,6)	Rejected
$\#8:7_2$	(0,0,0,6)	Rejected
$\#9:12_4$	(0,0,0,6)	Rejected

TABLE VI: SKM example (Off-line) Results

SKM Strategy	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Link**
Load by priority	10	10	10	10	40
Utilization (U)	U_1=0/10 =0%	U_2=0/10 =0	U_3=8+6/40 =35%	U_4=12+7+7/40 =65%	U=40/40 =100%
Blocking probability (Bp)	Bp_1=1/1	Bp_2=2/2	Bp_3=1/3	Bp_4=0/3	Bp=4/9
Acceptance ratio (AR)	AR_1=0/1	AR_2=0/2	AR_3=2/3	AR_4=3/3	AR=5/9
AllocTC Strategy	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Link**
Load by priority	10	10	10	7	37
Utilization (U)	U_1=9/40 =22.5%	U_2=6+7/40 =32.5%	U_3=8/40 =20%	U_4=7/40 =17.5%	U=37/40 =92.5%
Blocking probability (Bp)	Bp_1=0/1	Bp_2=0/2	Bp_3=2/3	Bp_4=2/3	Bp=4/9
Acceptance ratio (AR)	AR_1=1/1	AR_2=2/2	AR_3=1/3	AR_4=1/3	AR=5/9
RDM Strategy	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Link**
Load by priority	10	10	10	4	34
Utilization (U)	U_1=9/40 =22.5%	U_2=6/40 =15%	U_3=8+4/40 =30%	U_4=7/40 =17.5%	U=34/40 =85%
Blocking probability (Bp)	Bp_1=1/1	Bp_2=2/2	Bp_3=1/3	Bp_4=0/3	Bp=4/9
Acceptance ratio (AR)	AR_1=0/1	AR_2=0/2	AR_3=2/3	AR_4=3/3	AR=5/9

$$AR(T) = AD(T)/D(T) \tag{7}$$

$$AR_c(T) = AD_c(T)/D_c(T) \tag{8}$$

$$U(T) = \frac{\sum_{j=1}^{D} d_j(CT_c)\, I_{A(j)}\, T_j}{R * T} \tag{9}$$

$$U_c(T) = \frac{\sum_{j=1}^{D_c} d_j(CT_c)\, I_{A_c(j)}\, T_j}{RC_c * T} \tag{10}$$

Note that the definition of $I_{A(j)}$ and $I_{A_c(j)}$ for online scenario as in Eq. 5 and Eq. 6 respectively.

1) Online Simulation Scenarios: To evaluate our solution, the system used consists of a resource capacity equal to R = 160 units. Each class has RC_c=40 units. This resource capacity is divided into four classes considered in the system. The proposed strategy is used to check whether there are sufficient resources according to the class of the demand that needs to be admitted into the system, and then evaluate the metrics for comparison with other strategies for an online scenario. In the simulations, the demands are generated with a fixed lifetime equal 1-time slot, and the size is also fixed equal to 1 unit as the minimum granularity for allocation. Each demand has a single priority generated randomly from (1 to 4) with a generation rate of demands per each unit time equal to 200 demand. The total number of demands among classes generated until 100 unit time is 20,000 demands for each scenario. The traffic load consideration of the validation scenarios in each unit time is as follow: **Scenario 01:** Higher load in higher priority classes ($CT_1 = 20units > CT_2 = 40units > CT_3 = 60units > CT_4 = 80units$). **Scenario 02:** Higher load in all priority classes ($CT_4 = 50units > CT_3 = 50units > CT_2 = 50units > CT_1 = 50units$). Please also note that the used computer had Intel (R) Core (TM) 2 CPU 6400 @ 2.13GHz Memory 6GB and the used tool was Eclipse Java Oxygen.

V. OBTAINED RESULTS AND DISCUSSION

The performance of SKM is evaluated and compared with AllocTC and RDM in terms of the number of performance metrics as described below. The main objective of the above scenarios is to analyze the performance of SKM under different load distributions among the different priority classes.

TABLE VIII: Summary of scenario 1 results

Scenario1	**Simulations results (Values in %)**									
Metrics	U_1	U_2	U_3	U_4	U	AR_1	AR_2	AR_3	AR_4	AR
SKM	0	12.5	37.5	50	100	0	50	100	100	80
AllocTC	12.5	25	25	37.5	100	100	100	66.67	75	80
RDM	12.5	25	25	25	87.5	100	100	66.67	50	70

TABLE IX: Summary of scenario 2 results

Scenario2	**Simulations results (Values in %)**									
Metrics	U_1	U_2	U_3	U_4	U	AR_1	AR_2	AR_3	AR_4	AR
SKM	6.25	31.25	31.25	31.25	100	20	100	100	100	80
AllocTC	25	25	25	25	100	80	80	80	80	80
RDM	25	25	25	25	100	80	80	80	80	80

The obtained simulation results from scenario 1 are summarized in Table. VIII, in terms of AR_c, AR, U, U_c and shown in Fig. 3a for SKM, Fig. 3b for AllocTC and Fig. 3c for RDM. From the obtained results, the algorithms show a constant behavior in time since we assumed that 200 demands need to be allocated in each unit time along 100 unit times, on a single link with capacity equal to 160 resources (should cause link saturation). In light of that, the SKM outperforms RDM and AllocTC in the highest priority class by 50% and 25% in terms of AR_4, and by 25% and 12.5% in terms of U_4. AllocTC achieved higher acceptance ratio and utilization than RDM in class 4, since, in AllocTC performance, the higher priority classes can borrow unused resources from the lower ones to admit the demands (class 4 shared 20 resources from the lowest class). This is attributed to the fact that scenario one considered the higher priority classes to have more demand than the lower priority classes. Also, from the results, SKM outperforms RDM and AllocTC in class 3 by 33.33 % in terms of AR_3 and by 12.5% in terms of U_3 (as the expected from the behaviours). The SKM approach registers highest AR and U performance in the higher priority classes, due to the kicking operation as explained earlier. Moreover, even when the lower classes have fewer demands than the assigned resources for admitting demands, the unused resources can be shared by higher priority classes, which is not the case with RDM. If there are any unused resources in class 1 or 2 for the case of RDM, these resources will stay idle even if there is congestion in the higher priority classes.

In terms of total U and total AR, when we increase the load in higher priority classes, the RDM performance is the lowest one among the three strategies, achieving 70% as AR and 87.5% as U. Where the lower priority classes can only share resources from the higher ones. Therefore, in all unit times, the total acceptance ratio along the system will not exceed 160/200 = 80% as in SKM and AllocTC even if the number of demands was more than the capacity of the system. This is because each class cannot exceed its resources constraints (class 1 = 20 units, class 2 = 40 units, class 3 = 40 units, class 4 = 40 units).

Finally, from the results of scenario one, by increasing the number of demands in the higher priority classes we can realize a significant performance difference between SKM, AllocTC and RDM approach in terms of the strictness on priority. Thus, SKM provided better performance in terms of AR and U.

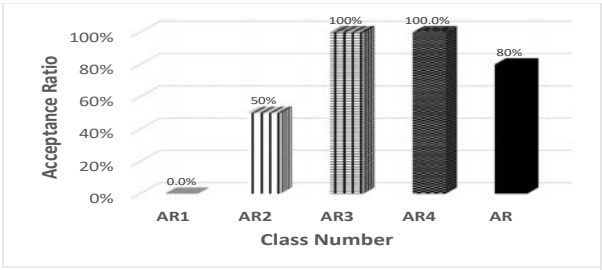
The obtained simulation results from scenario two are highlighted in Fig. 4 and summarized in Table IX. The results indicate that SKM, RDM and AllocTC, resulted in 100% U and 80% AR, where 160 demands are accepted from 200 demands per each unit time. From the obtained results in this scenario, the algorithms also show constant behavior in time. As expected, SKM registered the highest performance among the other two strategies (RDM, AllocTC) by 20% in terms of AR_4. Similarly, SKM outperforms RDM and AllocTC by 20% in terms of AR_3. Further, in terms of U_c, SKM, achieved 6.5% for class 4 and, 6.5% for class 3 more than both RDM and AllocTC. The above results show a superior performance of SKM for class 4 and 3 in terms of both AR_c and U_c. This can be justified by the nature of SKM, which permits higher priority classes to share unused resources from the lower ones and vice versa. The results also reveal that RDM has the same performance as AllocTC for the above classes under the considered scenario in terms of both AR_c and U_c. This can also be justified by the nature of AllocTC, which permits lower priority classes to share unused resources from the higher ones and vice versa similar to our proposal. However, in case of system saturation, unlike SKM, all borrowed resources should be returned in both senses for AllocTC case. Therefore, as illustrated in this scenario settings with the same traffic load in all classes, each class accepted 40 demands from 60 demands that needed to be admitted. In terms of RDM performance, the higher priority classes can not share unused resources from the lower ones, so it had the same equivalent performance to AllocTC.

SKM achieves the lowest performance in lower classes due to the kicking operation, which results in expelling the lower priority users to satisfy the demand requirements of the high priority classes. On the other hand, SKM intends to favour users belonging to high priority classes in terms of admission and resource allocation, hence the observed superior performance for high classes at the expense of low priority classes. Moreover, this behaviour makes SKM a right candidate for prioritized admission control.

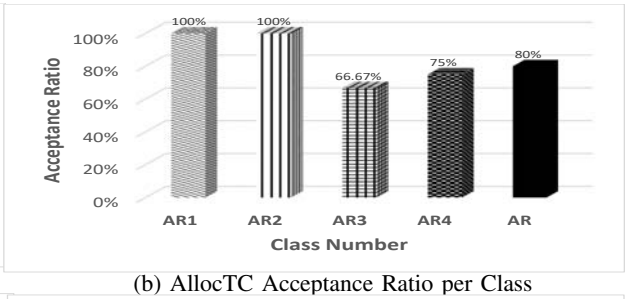
From the considered scenarios, SKM can guarantee to achieve 100% AR_c as long as the demanded resources from higher priority classes not exceed the capacity of the system. It also registers a better overall resource utilization compared to RDM in both traffic scenarios and the same performance as AllocTC. These results justify that SKM is a better admission control model for prioritized services than the existing schemes based in BAMs.

VI. CONCLUSIONS AND FUTURE WORK

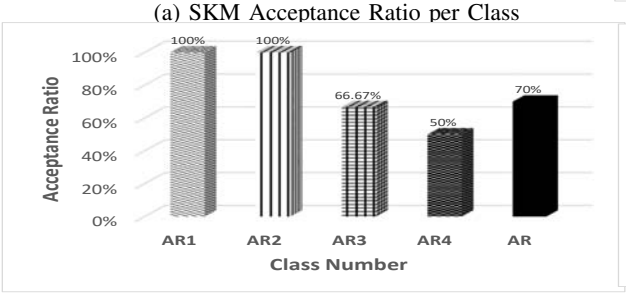
In this paper, a novel admission control model has been proposed, able to guarantee 100% utilization under different



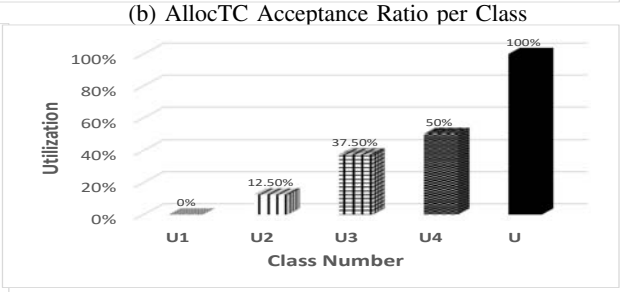
(a) SKM Acceptance Ratio per Class



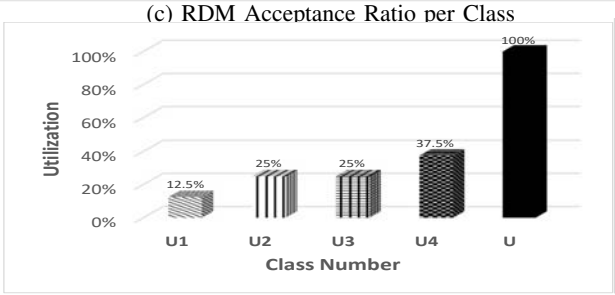
(b) AllocTC Acceptance Ratio per Class



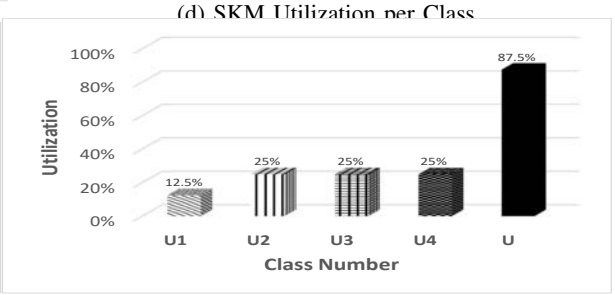
(c) RDM Acceptance Ratio per Class



(d) SKM Utilization per Class

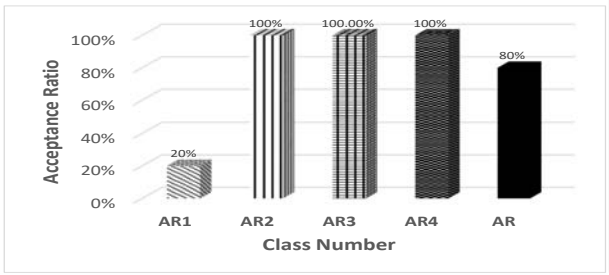


(e) AllocTC Utilization per Class

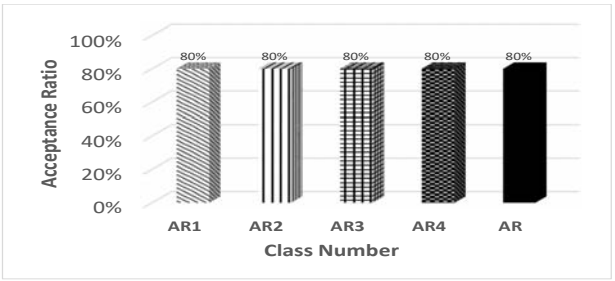


(f) RDM Utilization per Class

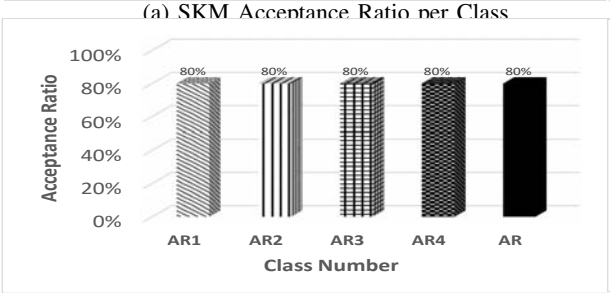
Fig. 3: SKM, AllocTC and RDM AR_c, U_c Comparison of Scenario 01



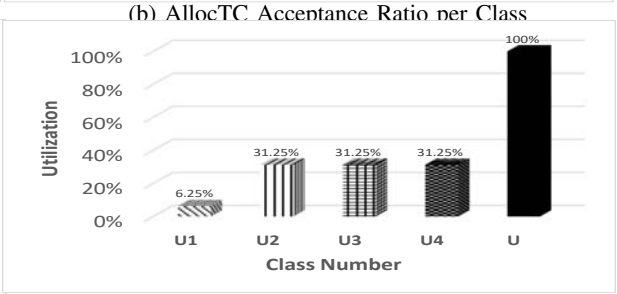
(a) SKM Acceptance Ratio per Class



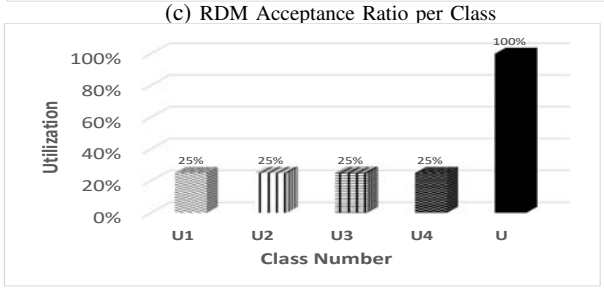
(b) AllocTC Acceptance Ratio per Class



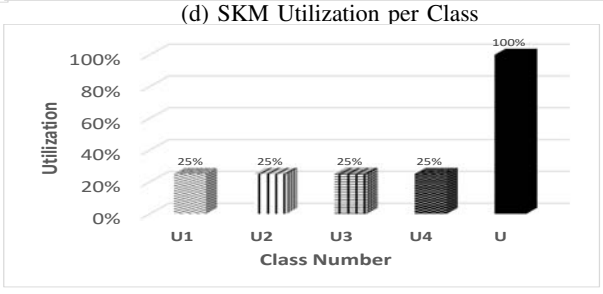
(c) RDM Acceptance Ratio per Class



(d) SKM Utilization per Class



(e) AllocTC Utilization per Class



(f) RDM Utilization per Class

Fig. 4: SKM, AllocTC and RDM AR_c, U_c Comparison of Scenario 02

priorities consideration, specially designed for highly congested scenarios with strict constraints for priority classes.

On the other hand, for the case of uncongested scenarios the SKM behaves similar to MAM, RDM and AllocTC. In RDM, the reservation of resources is made from bottom to top and not the reverse. So, in this way, resources utilization is more effective in comparison to MAM, which does not permit resource sharing across classes, but there is no guaranteed bandwidth for higher priority classes. Therefore, the benefit of using SKM is that the given class can be accepted regarding other classes (high or low) by means of initiating a squatting process, this is similar to the AllocTC per link behaviour of traffic distribution scenario. Beyond that, in SKM, the usage of resources for the higher priority classes is greater than originally reserved. SKM guarantees 100 percent of admission of high priority demands as long as there are resources in the lower priority classes, regardless of whether these resources are unused or occupied by the lower priority classes by means of initiating a kicking process. It is expected that groups of higher priority applications on multi-service networks could benefit from improved link utilization achieved by SKM. This corresponds to dynamically providing support to improve the quality of the application (SLA) for traffic distributions that occur in actual system operation, which means that the SKM is strict on priorities more than AllocTC and RDM.

Simulations validated the performance in the considered system in terms of utilization and acceptance ratio, including metrics per priority class, such as in scenario one SKM outperforms RDM and AllocTC in the highest priority class by 50% and 25%, in terms of AR_4, and by 25% and 12.5% respectively in terms of U_4. Also, SKM outperforms RDM and AllocTC in class 3 by 33.33 %, in terms of AR_3 and by 12.5% in terms of U_3. In terms of total U and total AR, when we increase the load in higher priority classes, the RDM performance is the lowest one among the three strategies, achieving 70% as AR and 87.5% as U compared to 80% AR and 100% U in both AllocTC and SKM.

As future work, the authors are planning to extend the SKM to consider other scenarios to study more the behaviour of SKM, as well as studying the complexity of the SKM implementation and propose a fast heuristic of SKM. As another future work, SKM will be improved by considering aforementioned thresholds to define and guarantee minimum resources for each class that will avoid resources beat down for lower priority classes.

ACKNOWLEDGMENT

This work has been partially supported by the Ministerio de Economy Competitividad of the Spanish Government under project TEC2016-76795-C6-1-R and AEI/FEDER, UE, and the SGR project, grant number 2017 SGR 397, from the Generalitat de Catalunya.

REFERENCES

[1] H. Zhang et al., "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," in IEEE Commun. Mag., vol. 55, no. 8, pp. 138-145, Aug. 2017.

[2] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez and A. Azcorra, "Network Slicing for Guaranteed Rate Services: Admission Control and Resource Allocation Games," in IEEE Transactions on Wireless Communications, vol. 17, no. 10, pp. 6419-6432, Oct. 2018.

[3] B. Han et al., "Admission and Congestion Control for 5G Network Slicing," 2018 IEEE Conference on Standards for Communications and Networking (CSCN), Paris, 2018, pp. 1-6.

[4] D. T. Hoang, D. Niyato, P. Wang, A. De Domenico and E. C. Strinati, "Optimal Cross Slice Orchestration for 5G Mobile Services," 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 2018, pp. 1-5.

[5] M. Jiang, M. Condoluci and T. Mahmoodi, "Network slicing management and prioritization in 5G mobile systems," European Wireless 2016; 22th European Wireless Conference, Oulu, Finland, 2016, pp. 1-6.

[6] R.F. Reale, R.Bezerra, J.Martins, "A preliminary evaluation of bandwidth allocation model dynamic switching," Int. J. Comput. Netw. Commun. (IJCNC), vol. 6, no. 3, pp. 131-143, May. 2014.

[7] G. M. Duraes et al., "Evaluating the applicability of bandwidth allocation models for EON slot allocation," in IEEE Intl. Conf. on Adv. Net. and Tel. Sys. (ANTS), Bhubaneswar, pp. 1-6, 2017.

[8] F. Le Faucheur, W. Lai, "Maximum Allocation Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering," RFC 4125, 2005.

[9] F. Le Faucheur, "Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering," RFC 4127, 2005.

[10] D. Adami, C. Callegari, S. Giordano, M. Pagano, M. Toninelli, "G-RDM: a new bandwidth constraints model for DS-TE networks," in: Proceedings of the IEEE Global Telecommunications Conference, 2007, pp. 2472-2476.

[11] R.F. Reale, W. d. C. P. Neto and J. S. B. Martins, "AllocTC-sharing: A new bandwidth allocation model for DS-TE networks," in: Proc. of the IEEE Net. Oper. and Mgmt. Symposium, pp. 1-4, Quito, 2011.

[12] R.F. Reale, Rafael Freitas, Romildo Martins da Silva Bezerra and Joberto S. B. Martins, "G-BAM: A Generalized Bandwidth Allocation Model for IP/MPLS/DS-TE Networks," CoRR abs/1806.07292 (2014): n. pag.

[13] S.K. Sadon, N.M. Din, M.H. Al-Mansoori, N.A. Radzi, I.S. Mustafa, M. Yaacob, M.S.A. Majid, "Dynamic hierarchical bandwidth allocation using Russian Doll Model in EPON," Comput. Electr. Eng. 38 (6) (2012) 1480-1489.

[14] R. Trivisonno, R. Guerzoni, I. Vaishnavi, A. Frimpong, "Network resource management and QoS in SDN-enabled 5G systems," in: Proceedings of the IEEE Global Communications Conference, 2015, pp. 1–7.

[15] N. Subhashini, A.B. Therese, "User prioritized constraint free dynamic bandwidth allocation algorithm for EPON networks," Indian J. Sci. Technol. 8 (33) (2015) 1–7.

[16] A. Ayoub Bahnasse et al., "Novel SDN architecture for smart MPLS Traffic Engineering-DiffServ Aware management," in Future Generation Computer Systems, Volume 87, pp. 115-126, 2018.

[17] X. Hesselbach et al., "Management of resources under priorities in EON using a modified RDM based on the squatting-kicking approach," Intl. Conf. on Transparent Optical Networks (ICTON), Trento, pp. 1-5, 2016.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

SERA: Sistema para la Evaluación y Retroalimentación Automática de Prácticas

Isaac Agudo, Rubén Ríos, Ana Nieto

Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga, Spain
Email: {isaac, ruben, nieto}@lcc.uma.es

***Resumen*—En este artículo presentamos una sistema modular y altamente configurable que permite no sólo la generación y evaluación automática de prácticas de laboratorio sino también proporcionar una retroalimentación instantánea al estudiante para orientar el proceso de aprendizaje y fomentar su autonomía. Este sistema ha sido integrado dentro de la plataforma Moodle en varias asignaturas del Área de Ingeniería Telemática en la Universidad de Málaga.**

***Palabras Clave*—Reutilización, automatización, prácticas, feedback instantaneo, LTI, Moodle, autonomía estudiante**

Tipo de contribución: *Formación e innovación educativa en desarrollo*

I. Introducción

Entre los diferentes objetivos promovidos por el Espacio Europeo de Educación Superior (EEES) se encuentra impulsar un cambio en las metodologías docentes con el fin de mejorar el proceso de aprendizaje del estudiante. Este cambio de metodologías supone, en la mayoría de casos, una mayor dinamización de la docencia, haciendo al estudiante más partícipe de su propio aprendizaje. En definitiva, el objetivo es orientar la enseñanza a la acción, esto es, a que los estudiantes aprendan haciendo en lugar de ser meros consumidores de información.

Este objetivo se ha materializado en diferentes tipos de metodologías como el aprendizaje basado en proyectos/problemas o las clases invertidas [1], reduciendo el número de horas dedicadas a clases magistrales en favor de un mayor contenido práctico. Esto es especialmente relevante en el contexto particular de las ingenierías, donde los autores de este artículo tienen la mayor parte de su docencia, pero también en otras ramas de conocimiento debido a la implantación de las TIC en todos los ámbitos.

Este cambio de perspectiva, que sin duda es beneficiosa para el estudiante, supone un gran esfuerzo adicional para el docente, que no sólo debe preparar los contenidos teóricos necesarios para la asignatura sino que además se ve inmerso en la constante creación y corrección de prácticas con el fin de que los estudiantes reciban retroalimentación. Del mismo modo, el docente debe guiar a aquellos estudiantes que encuentran alguna dificultad para finalizar la práctica, ya sea en el aula o fuera de ella. En estos casos, si la retroalimentación del profesor no es inmediata, el estudiante suele perder el interés.

Lamentablemente, esta forma de proceder es sólo abordable cuando el número de estudiantes es relativamente reducido ya que, en otro caso, el docente es incapaz de hacer frente a las necesidades del alumnado sin que esto suponga un exceso de carga de trabajo. Ante esta situación, los autores de este trabajo sondeamos diferentes alternativas que pudieran ser integradas fácilmente en la plataforma de enseñanza virtual de la Universidad de Málaga, actualmente basada en Moodle [2]. Entre ellas se consideró utilizar herramientas como Siette (Sistema de Evaluación Inteligente mediante Tests) [3] y VPL (Virtual Programming lab for Moodle) [4] pero no se adaptaban completamente a nuestras necesidades. Siette por estar principalmente orientado a pruebas adaptativas en función de las respuestas de los estudiantes, algo que queda un fuera de nuestras aspiraciones, y si bien es altamente personalizable, el esfuerzo para la creación de nuevas actividades no era abarcable. VPL al estar más enfocado a actividades de programación y, a pesar de ser también altamente personalizable, su interfaz de usuario lo hace poco amigable para otro tipo de actividades.

Por ello, comenzamos realizando algunas experiencias ad-hoc pero (i) por un lado no era sencillo integrar nuestros desarrollos con el Campus Virtual, al que los estudiantes están acostumbrados y (ii) por otro lado, las prácticas planteadas eran difícilmente reutilizables ya que estaban diseñadas para cumplir un objetivo muy concreto.

A partir de estas primeras experiencias nos planteamos la necesidad de desarrollar un sistema que permitiera el diseño de prácticas de laboratorio reutilizables, autoevaluables y capaces de ofrecer al estudiante asesoramiento en el caso de que el resultado no fuera el esperado. Además, entre nuestros objetivos irrenunciables se encontraba

que se pudiera integrar con la plataforma Moodle para unificar todos los contenidos y resultados de la asignatura en un único lugar. El resultado de este trabajo es SERA.

A continuación se presenta el contexto en el que se desarrolla nuestra propuesta, respaldada por dos proyectos de innovación educativa financiados por la Universidad de Málaga. Seguidamente, en la sección III presentamos detalles sobre el diseño e implementación de la solución. La sección IV describe el lenguaje creado para la especificación de actividades además de mostrar un ejemplo. La sección V presenta nuestras conclusiones y trabajo futuro.

II. Contexto

El sistema SERA comienza a gestarse en un Proyecto de Innovación Educativa (PIE) [5] centrado en la evaluación basada en hitos, donde se observó la necesidad de automatizar la evaluación de competencias prácticas del alumnado.

La experiencia se ha llevado a cabo en asignaturas de varias titulaciones donde los autores de este artículo imparten asignaturas relacionadas con la Seguridad en Redes. Las asignaturas en cuestión son las siguientes:

- *Seguridad en Redes* del Grado en Ingeniería Telemática (Obligatoria)
- *Seguridad Informática e Informática Forense* del Grado en Criminología (Optativa)
- *Seguridad en Redes y Transacciones Online* del Máster en Dirección y Marketing Digital (Obligatoria)

Aunque estas asignaturas están dirigidas a grupos muy heterogéneos, tanto en el número de estudiantes como en sus conocimientos previos sobre redes e informática, todas están orientadas a iniciar al estudiante en el ámbito la Seguridad de la Información. Por ello, a pesar de que cada asignatura tiene una serie de contenidos específicos a la titulación, comparten una serie de contenidos básicos sobre amenazas, servicios de seguridad, criptografía, protocolos y herramientas de comunicación segura. Entre las prácticas más recurrentes en estas asignaturas cabe destacar:

- Implementación de algoritmos criptográficos
- Gestión de certificados digitales
- Cifrado y firma digital de documentos
- Correo electrónico seguro (S/MIME y PGP)
- Servidor Web seguro (TLS)
- Configuración de Cortafuegos
- Redes Privadas Virtuales (IPSEC)

A pesar de que la orientación y el nivel de profundidad alcanzado en cada una de las prácticas depende del perfil de cada titulación, es interesante observar que las practicas más complejas se pueden descomponer en actividades más sencillas. Estas prácticas sencillas suelen ser comunes a todas las asignaturas, lo que facilita la reutilización mediante la concatenación de actividades.

En lo que a medios tecnológicos se refiere, para la realización de las prácticas nos encontramos con la dificultad de que es necesario utilizar multitud de entornos y herramientas, lo cual dificulta enormemente el seguimiento y evaluación automática de las mismas. Más aún si consideramos que entre nuestros objetivos está tener un sistema completamente integrado en Moodle.

III. El sistema SERA

En esta sección presentamos nuestra solución de manera detallada. En primer lugar se ofrece una visión general del funcionamiento del sistema y a continuación se describe la arquitectura e implementación del mismo.

A. Visión Global

SERA es un sistema para el diseño de prácticas de laboratorio que permite la evaluación de competencias relacionadas con la seguridad en redes, la programación y el diseño de protocolos de comunicación.

El flujo de trabajo con nuestro sistema comienza con la especificación de una actividad por parte del docente. Para ello se crea una plantilla con parámetros que permiten instanciar tareas personalizadas para cada estudiante. El valor de estos parámetros es generado por el sistema SERA utilizando unos módulos específicos para cada tipo de actividad. Los parámetros generados servirán, además, a nuestro sistema para comprobar si los envíos realizados por los estudiantes son correctos.

B. Arquitectura

El sistema SERA consta de tres componentes principales (ver Figura 1):

- *Generador* encargado de la creación de las actividades en función de los parámetros proporcionados por el docente
- *Validador* encargado de comprobar si la actividades son correctas en función de elementos enviados por el estudiantes a través de un formulario
- *Monitor* encargado de validar actividades pero mediante la monitorización de acciones realizadas por el estudiante en el sistema

Estos tres componentes dependen de un conjunto de módulos que proporcionan métodos específicos para cada actividad encargados de la generación de las actividades y su evaluación. Los módulos deben ser programados por el docente y podrán ser utilizados en diferentes actividades. Esto permite la modularidad y extensibilidad de SERA.

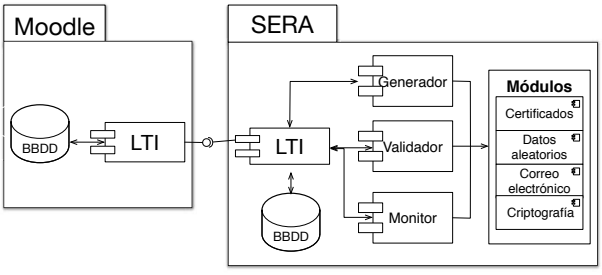


Figura 1: Componentes principales de SERA

Tal como se aprecia en la figura, el sistema SERA está conectado con *Moodle* a través del estándar LTI (*Learning Tool Interoperability*)[1], ampliamente utilizado

[1] http://www.imsglobal.org/activity/learning-tools-interoperability

por la comunidad educativa y que ya se ha utilizado con éxito para integrar actividades de programación en Moodle [6].

El estándar LTI permite conectar de forma fácil y segura aplicaciones externas con los sistemas de gestión de aprendizaje (LMS - *Learning Management Systems*) más comunes hoy en día. Mediante LTI se establece una conexión segura entre ambas partes que permite (a) recibir en la aplicación externa, en nuestro caso el sistema SERA, la sesión del estudiante y (b) enviar a la plataforma de aprendizaje, en nuestro caso Moodle, las calificaciones correspondientes.

C. Implementación

Para la implementación de SERA se ha usado *Node.js* y en particular el marco de trabajo para aplicaciones web *express*, que nos permite crear rutas de acceso al sistema. En nuestro caso definimos:

- *lti*. Se encarga de recibir las peticiones de Moodle, verificar los datos del estudiante y llamar al componente Generador para crear una sesión local para el estudiante dentro del sistema SERA. En esencia, transforma las plantillas definidas por el docente (ver sección IV) en actividades.
- *solve*. Se encarga de validar el formulario de entrada del estudiante. Comprobará que el estudiante ha enviado todos los datos solicitados en la actividad y verificará si se cumplen las condiciones especificadas en la actividad.
- *monitor*. Se encarga de proporcionar una conexión entre los monitores de actividad de cada módulo y la interfaz web mediante AJAX. De esta forma, cuando el Monitor detecte que se ha completado la actividad, la interfaz web se refresca y muestra la información necesaria al estudiante.

La funcionalidad del sistema recae en última instancia en los módulos específicos para cada tipo de actividad.

Todos los módulos deben implementar una función de generación de parámetros de forma obligatoria, así como una serie de funciones de comprobación. En el código 1 podemos ver la estructura de esta función para el caso particular del módulo x509.

Código 1 : Función de generación del módulo x509

```
1  async function generateParam(template){
2    let param = {};
3    if (template.hasOwnProperty('type'))
4    {
5      let subject = template.input.subject;
6      switch(template.type.toLowerCase()) {
7      case 'csr':  // Creamos un CSR
8        param = {pem: await pki.createPKCS10(subject),
               subject: subject};
9        break;
10     case 'p12':  // Creamos un Cert
11       param = {pem: await pki.createPKC12(subject),
              subject: subject};
12       break;
13     ...
14     }
15   }
16   return param;
17 }
```

Esta función es la que invocará el componente Generador cuando tenga que instanciar alguna actividad que utilice dicho módulo. En este caso, el módulo permite generar un CSR y un certificado (y clave privada) en formado PKCS12 para ser utilizados en la actividad que se quiera definir.

En cuanto a las funciones de comprobación, estas permiten comprobar propiedades de los envíos de los estudiantes, así como comprar los parámetros generados anteriormente frente a los envíos de los estudiantes.

IV. Especificación de actividades

La especificación de las actividades se descompone en 4 bloques:

- **Descripción**. Además de un título, cada actividad debe incluir una descripción que se personalizará para cada estudiante en función de los parámetros generados por SERA.
- **Envíos**. Las actividades suelen solicitar al estudiante el envío de algún tipo de contenido usando un formulario. De no ser así, se debe especificar la función que monitorizará las actividades realizadas por el estudiante.
- **Parámetros**. De cara a personalizar la actividad, el docente puede especificar un conjunto de parámetros que se instanciarán usando los módulos específicos. Hay tres tipos de parámetros:
  - *Literales*. Cadenas de texto personalizadas en función de la sesión del estudiante y otros parámetros generados en la misma actividad.
  - *Referencias*. Referencias a envíos o parámetros de otras actividades ya superadas por el estudiante.
  - *Generados*. Parámetros creados usando módulos específicos.
- **Comprobaciones**. Por último, la actividad debe incluir una serie de comprobaciones que relacionen los envíos del estudiante con los parámetros generados previamente.

Esta forma de especificar actividades permite (a) que cada estudiante reciba actividades personalizadas con distintos valores y (b) establecer dependencias para construir actividades complejas a partir de otras más simples.

Por limitaciones de espacio, en lugar de describir el lenguaje de especificación de actividades en detalle, se muestra un ejemplo de uso.

A. Ejemplo de uso

En el código 2 se puede ver un ejemplo de especificación de actividad. En ella se aprecia como la descripción (línea 3) utiliza código HTML con unas etiquetas especiales que permiten incluir los parámetros de la actividad (líneas 6-12). Entre los parámetros, se incluyen tanto parámetros literales (línea 9) como parámetros generados (línea 11). También se puede ver los campos que debe enviar el estudiante para resolver la actividad (líneas 4-5).

Por último se especifican las comprobaciones (líneas 13-19) de la actividad, que devuelven en caso de fallo

Código 2 : Ejemplo de especificación de actividad

```
1 {id: 20,
2  title: 'Crear CSR para S/MIME',
3  desc:`Crea un Certificate Signing Request (CSR)
       para pedir un certificado de correo electrónico
        poniendo en el CN tu nombre, es decir: <span
       class= "param"><%= cn %></span>. No olvides
       especificar también en el Subject Alternative
       Name tu correo electrónico <span class= "param"
       ><%= email %></span>. Es importante que la
       longitud de las claves que utilices sea de <
       span class= "param"> <%= length %></span> bits.
       `,
4  submissions: [
5    {name: 'csr', label: 'CSR', type:'file'}],
6  params: [
7    {name: 'email', value: '<%= eMail %>'},
8    {name: 'san', value: 'email:<%= eMail %>'},
9    {name: 'length', value: '2048'},
10   {name: 'type', value: 'smime'},
11   {name: 'random', module: 'data', type:'dec',
       input:{low:0,high:99}},
12   {name: 'cn', value: '<%= fullName %> <%= random
       %>'}],
13 checks: [
14   {module: 'x509', check:'checkSecureHash', input:{
       csr: 'csr'}},
15   {module: 'x509', check:'isCsr', input:{csr: 'csr'
       }},
16   {module: 'x509', check:'checkType', input:{csr: '
       csr', type:'type'}},
17   {module: 'x509', check:'checkCN', input:{csr: '
       csr', CN:'cn'}},
18   {module: 'x509', check:'checkSan', input:{csr: '
       csr', san:'san'}},
19   {module: 'x509', check:'checkKeyLength', input:{
       csr: 'csr', length:',length'}},]}
```

una retroalimentación que puede orientar al estudiante en la consecución de su objetivo. El módulo utilizado para las comprobaciones es el encargado de proporcionar la retroalimentación especifica.

En la figura 2 se muestra la vista del estudiante a partir de la plantilla descrita anteriormente. En el menú de la izquierda se muestran las actividades disponibles y se indica, con un código de colores, aquellas que aún no se han empezado (blanco), aquellas se han finalizado con éxito (verde) y aquellas que se han intentado de forma fallida (rojo). En la descripción de la actividad, se puede ver como cada parámetro se ha instanciado de forma particular (en azul) de acuerdo a la especificación y como aparece un formulario de entrega con un único campo, CSR, tal como se indicaba en la plantilla (línea 5).

El estudiante tiene libertad para elegir qué actividad realizar. Las actividades con algún parámetro de tipo referencia, como la actividad A6, requieren que se realicen previamente las actividades que la bloquean.

V. Conclusiones y Trabajo Futuro

En este articulo hemos presentado SERA, un sistema integrado con la plataforma Moodle mediante LTI, que ofrece al estudiante retroalimentación instantánea sobre las actividades que está desarrollando, guiándole en todo momento hacia la solución de manera autónoma.

El sistema está en estado de pruebas en asignaturas de diferentes titulaciones y esperamos seguir ampliando la experiencia a otras asignaturas del área, así como el número de módulos disponibles para permitir que cualquier

Figura 2: Interfaz Web de SERA

docente pueda definir sus propias actividades en SERA sin necesidad de programar un módulo especifico.

Estamos en proceso de obtener impresiones de los estudiantes que lo han utilizado y comparar los resultados de las prácticas obtenidos por estudiantes de años anteriores. Aunque aún no disponemos de esos datos, las primeras impresiones por parte del alumnado son muy positivas porque les permite tener una visibilidad clara de su progreso y trabajar de manera más autónoma. Nuestro objetivo es que SERA permita reducir la carga docente asociada a la evaluación y seguimiento continuo.

Agradecimientos

Este trabajo ha estado parcialmente financiado por el proyecto de innovación educativa PIE17-137 de la Universidad de Málaga.

Referencias

[1] R. Clark, A. Kaw, Y. Lou, A. Scott, and M. Besterfield-Sacre, "Evaluating blended and flipped instruction in numerical methods at multiple engineering schools," *International Journal for the Scholarship of Teaching and Learning*, vol. 12, no. 1, 2018.

[2] S. Kumar, A. K. Gankotiya, and K. Dutta, "A comparative study of moodle with other e-learning systems," in *2011 3rd International Conference on Electronics Computer Technology*, vol. 5, April 2011, pp. 414–418.

[3] R. Conejo, E. Guzmán, E. Millán, M. Trella, J. L. Pérez-De-La-Cruz, and A. Ríos, "Siette: A web-based tool for adaptive testing," *Int. J. Artif. Intell. Ed.*, vol. 14, no. 1, pp. 29–61, Jan. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1434852.1434855

[4] J. C. Rodríguez-del Pino, E. Rubio Royo, and Z. Hernández Figueroa, "A virtual programming lab for moodle with automatic assessment and anti-plagiarism features," in *Proceedings of The 2012 Internacional Conference on e-Learning, e-Business, Entreprise Information Systems, & e-Government*, 2012.

[5] D. Nunez, F. Moyano, A. Nieto, J. J. Ortega, I. Agudo-Ruiz, and J. López, "A milestone-driven approach for lab assignments evaluation in information security," in *International Conference on e-Learning 2014*, 2014.

[6] A. J. Sierra, Á. Martín-Rodríguez, T. Ariza, J. Muñoz-Calle, and F. J. Fernández-Jiménez, "LTI for interoperating e-assessment tools with LMS," in *Methodologies and Intelligent Systems for Technology Enhanced Learning*. Springer International Publishing, 2016, pp. 173–181.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Optimización conjunta del nivel *split* y *scheduling* en redes 5G

Luis Diez, Víctor González y Ramón Agüero
Departamento de Ingeniería de Comunicaciones,
Universidad de Cantabria
Plaza de la Ciencia s.n. 39005, Santander, Cantabria. España
ldiez@tlmat.unican.es, victor.gonzalezcar@alumnos.unican.es, ramon@tlmat.unican.es

***Resumen*—Las técnicas de virtualización serán una pieza fundamental en las futuras tecnologías 5G. Sin embargo, las soluciones totalmente centralizadas, tales como *Cloud Radio Access Network* (C-RAN), podrían no ser factibles, debido a las requisitos adicionales impuestos al *fronthaul* de la red. En este sentido, las técnicas *flexible functional split* permiten definir niveles de centralización de manera flexible, proporcionando, así, un compromiso entre el rendimiento y la aplicabilidad práctica. A pesar del creciente interés en estas técnicas, no se ha prestado mucha atención a su interacción con la planificación a la hora de acometer la transmisión de tramas (*scheduling*). Es por ello que en este trabajo se analiza la gestión combinada del nivel de centralización y el envío de tramas, a fin de minimizar el retardo en la red de acceso. En concreto, se compara, en diferentes escenarios, la solución óptima global con aquellas aportadas por optimizaciones parciales, cuya implementación puede resultar más asequible desde el punto de vista de complejidad computacional. A la vista de los resultados obtenidos, se puede concluir que las políticas de planificación fija para el envío de tramas presenta un comportamiento similar al óptimo global en escenarios con tecnologías de acceso heterogéneas y tráfico homogéneo. Además, se ha comprobado que, para los escenarios analizados, es preferible mantener un política de envío fija y optimizar únicamente el nivel de *split*.**

***Palabras Clave*—functional split, scheduling, 5G, cloud RAN, retardo, NFV**

I. INTRODUCCIÓN

Entre los cambios en la arquitectura de red que van a caracterizar los despliegues 5G, uno de los que se consideran más determinantes es, sin duda, la virtualización de funciones de red *Network Function Virtualization* (NFV), utilizando técnicas *Software Defined Networking* (SDN). Mientras que en el pasado las arquitecturas de red celular, tales como 4G, han evolucionado hacia topologías descentralizadas, los cada vez más exigentes requisitos de las redes 5G requieren una mayor coordinación de los elementos de acceso, lo que, a su vez, solo puede lograrse mediante arquitecturas centralizadas.

Esto se consigue mediante la separación de las funcionalidades pertenecientes a la red de acceso, de modo que un conjunto de ellas se virtualizan, centralizándose en centros de procesado de datos, mientras que el resto permanece en los elementos de acceso. Con la aparición de la virtualización de funciones de red, se propusieron inicialmente soluciones totalmente centralizadas (C-RAN). En estos casos, se virtualizan todas las funciones de red, quedando en los elementos de acceso o *Remote Radio Head* (RRH), las funciones básicas de la capa física. Sin embargo, este tipo de soluciones requieren capacidades de comunicación muy altas entre la RRH y la unidad de banda base o *Base-Band Unit* (BBU) virtualizada, lo que puede ser difícil de satisfacer en ciertos escenarios. Por esta razón, en los últimos años han surgido propuestas que permitan niveles de centralización flexibles (*flexible functional split*) [1], [2].

Este cambio de paradigma en la arquitectura de red, permitirá implementar soluciones que den respuesta a los requisitos de los despliegues 5G, a la vez que habilitará una reducción de costes, en comparación con las soluciones totalmente centralizadas. Por otro lado, la posibilidad de definir diferentes niveles de centralización conlleva responder a varias cuestiones. Por un lado, se ha de decidir el nivel de centralización (*functional split*) de los diferentes elementos de acceso, de acuerdo al tipo de coordinación necesaria. Por otro lado, es necesario planificar la transmisión de tramas en las BBU hacia las RRH que gestiona. Esta planificación, o *scheduling*, deberá hacerse de forma que se minimice el retardo, ya que se trata de uno de los principales parámetros a optimizar en la tecnología 5G.

En este trabajo se analiza, sobre una arquitectura de *flexible split*, la configuración conjunta del nivel de centralización y planificación de envío de tramas. Para ello, se toma como base el trabajo realizado por Koutsopoulos [3], en el que se describe, de forma teórica, el problema de optimización subyacente. El autor caracteriza

la complejidad del problema conjunto, proporcionando pautas para solucionar los casos en los que uno de los dos parámetros (nivel de centralización o planificación del envío) se conoce. Sin embargo, no se proporcionan resultados prácticos de ninguno de los problemas.

En concreto, este trabajo persigue los siguientes objetivos:

- Implementación de las soluciones descritas en [3] para minimizar el retardo.
- Evaluación de dichas soluciones sobre diferentes escenarios, usando configuraciones realistas.
- Análisis de la pérdida de rendimiento cuando se usan optimizaciones parciales (uno de los parámetros conocidos), con respecto a la solución óptima global.

El resto del documento sigue la estructura que se indica a continuación. En la Sección II se analizan los trabajos relacionados existentes en la literatura, indicando las principales diferencias con el estudio que aquí se presenta. Posteriormente, en la Sección III se describe el modelo del sistema, así como las variantes del problema de optimización para obtener la solución óptima de selección de *split* y *scheduling*. A continuación, en la Sección IV se evalúa el rendimiento de estos problemas sobre diferentes escenarios. Finalmente, el artículo concluye en la sección V, donde se resumen las contribuciones y se enumeran líneas futuras de investigación.

II. TRABAJOS PREVIOS

Como se ha mencionado anteriormente, las arquitecturas C-RAN [4], [5] se consideran una de las soluciones clave para satisfacer los requisitos impuestos por la tecnología 5G. La principal idea que subyace en este tipo de soluciones es la de trasladar funciones de red, típicamente localizadas en las estaciones base, a un controlador central. Sin embargo, las soluciones en las que se centralizan todas las funciones de red pueden no resultar prácticas, debido a las altas capacidades (tanto de transmisión como de retardo) que se impondrían sobre los enlaces del *fronthaul*. En este sentido, las soluciones totalmente centralizadas requerirían la implementación de un *fronthaul* basado únicamente en fibra óptica [6], [7], lo que a su vez implica costes de despliegue muy altos. Por ello, han aparecido varias iniciativas que proponen un cambio de diseño del *fronthaul* [8], de modo que se puedan definir diferentes niveles de centralización. El lector puede encontrar una revisión detallada de los niveles de *split* que se están definiendo en [9].

Existen además trabajos que proponen no solo la selección de niveles de *split*, sino que el proceso de selección se lleve a cabo de forma dinámica, dando lugar al concepto de *flexible functional split*. En este caso el nivel de centralización se puede adaptar en función de los requisitos de retardo, así como del estado de los enlaces del *fronthaul*. Este tipo de arquitectura basada en centralización flexible para 5G se ha descrito en [10] y validado, en entorno de laboratorio, en [11]. Por otro lado, en [12], [13] se describen las principales características de este tipo de soluciones.

A partir del concepto de *flexible functional split* algunos trabajos han propuesto técnicas de mejora de la eficiencia energética [14], [15]. Otros se han centrado en su combinación con redes de transporte ópticas [16], [17]. Por otro lado, algunos trabajos han estudiado la interacción de la selección de *split* con la gestión de recursos radioeléctricos. Sin embargo, más allá del estudio realizado por Koutsopoulos [3], citado anteriormente, no se ha prestado atención a la optimización conjunta de la selección de *split* y la planificación de envío de tramas desde la BBU a los puntos de acceso gestionados por este, que es precisamente donde se sitúa la principal contribución de este trabajo.

III. MODELADO DEL SISTEMA

Como se ha mencionado en secciones anteriores, en este trabajo se adopta el modelo propuesto en [3], que se reproduce a continuación de forma resumida.

Se considera una arquitectura de red celular, en la que se realiza *functional split*, de manera que un conjunto de RRH, $\mathcal{R}$, se conectan a una unidad central en la nube, capaz de virtualizar varios BBU mediante un conjunto de enlaces $\mathcal{L}$. Se asume que la unidad central dispone de un solo procesador con capacidad computacional C^B, de forma que solo se puede procesar de forma concurrente una trama, perteneciente a una RRH. De forma similar, se usa C_i para denotar la capacidad computacional de la RRH i, mientras que L_i indica la capacidad de transferencia de datos del enlace entre la RRH i y la unidad en la nube. Cabe mencionar que en este trabajo únicamente se considerará el enlace descendente, aunque el modelo puede ser igualmente aplicado al ascendente.

Se asume un escenario en el que el tiempo está ranurado, de modo que se transmite una nueva trama para cada RRH al inicio de cada ranura. A fin de gestionar el envío de tramas, existe un controlador en la unidad central que, de manera global, decide el nivel de *split* aplicado a cada trama, así como el orden en que estas se envían.

Por simplicidad, se asume que todas las RRH pueden acomodar el mismo conjunto de niveles de *split*, $\mathcal{F}$. De esta manera, en cada ranura el controlador selecciona un vector de niveles de *split* $\boldsymbol{s} := \{s_1, ...s_{|R|}\} \in \mathcal{S}$, donde $s_i \in \mathcal{F}$ se corresponde con la decisión tomada para la RRH $i \in \mathcal{R}$ y $\mathcal{S}$ representa el conjunto de todas las decisiones posibles. Cabe indicar que el número de posibles configuraciones crece exponencialmente con el número de opciones de centralización, de modo que $|\mathcal{S}| = |\mathcal{F}|^{|\mathcal{R}|}$.

A partir de la decisión de *split* de cada RRH i, s_i, se definen las variables ω_{i,s_i} y $\hat{\omega}_{i,s_i}$ para indicar la carga computacional necesaria, respectivamente, en la BBU y RRH. De forma similar se define la función d_{i,s_i} para representar la cantidad de datos que tienen que transmitirse a través del enlace L_i, dependiendo de la decisión concreta de *split*.

De manera similar al modelado de la centralización, Π representa el conjunto de posibles políticas de transmisión (*scheduling*). Una política concreta se representa mediante un vector, $\boldsymbol{\pi} = \{\pi_1, ..., \pi_{|R|}\}$, de números enteros positivos, donde π_i indica el orden en que se envía la trama

perteneciente a la RRH i. Por ejemplo, si se asume que se tiene un escenario con 4 RRH, la política de envío $\boldsymbol{\pi} = \{3, 4, 2, 1\}$ indicaría que en primer lugar se envía la trama de la RRH 4 y a continuación las de las estaciones 3, 1 y 2. Cabe destacar que el espacio de las políticas de envío es $|\Pi| = |\mathcal{R}|\,!$.

Teniendo en cuenta la notación presentada, se puede calcular el retardo de procesado en la BBU de una trama i como:

$$\delta^B_{i,s_i} = \frac{\omega_{i,s_i}}{C^B} \tag{1}$$

De forma similar, considerando la cantidad de datos que se deben transmitir como consecuencia de la decisión de *split* de cada trama i, el retardo de transmisión se define como:

$$\delta^L_{i,s_i} = \frac{d_{i,s_i}}{L_i} \tag{2}$$

Finalmente, el retardo asociado al procesado en la estación viene dado por la siguiente expresión:

$$\delta^R_{i,s_i} = \frac{\hat{\omega}_{i,s_i}}{C_i} \tag{3}$$

De este modo, se puede calcular el retardo total de una trama i como:

$$d_i(\boldsymbol{\pi}, \boldsymbol{s}) = \delta^B_{i,s_i} + \delta^L_{i,s_i} + \delta^R_{i,s_i} + \sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j} \tag{4}$$

donde $\sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j}$ indica el tiempo en que la trama espera en la BBU para ser enviada. Igualmente, se puede definir el retardo total en el sistema, D, como la suma de retardos de cada trama:

$$\begin{aligned} D(\boldsymbol{s}, \boldsymbol{\pi}) &= \sum_{i\in\mathcal{R}} d_i(\boldsymbol{s}, \boldsymbol{\pi}) = \\ &= \sum_{i\in\mathcal{R}} \left(\delta^B_{i,s_i} + \delta^L_{i,s_i} + \delta^R_{i,s_i} + \sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j} \right) \end{aligned} \tag{5}$$

El retardo global definido en la Ecuación 5 se puede reformular para expresarlo en función de la contribución realizada por cada trama. Para ello, se agrupan los retardos asociados al procesado en la BBU de la siguiente manera:

$$\begin{aligned} &\sum_{i\in\mathcal{R}} \left(\delta^B_{i,s_i} + \sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j} \right) = \sum_{i\in\mathcal{R}} \sum_{j:\pi_j\le\pi_i} \delta^B_{j,s_j} = \\ &= \delta^B_{1,s_1} + (\delta^B_{1,s_1} + \delta^B_{2,s_2}) + ... + (\delta^B_{1,s_1} + ... + \delta^B_{|\mathcal{R}|,s_{|\mathcal{R}|}}) = \\ &= \sum_{i\in\mathcal{R}} \delta^B_{i,s_i} \cdot (|\mathcal{R}| - \pi_i + 1) \end{aligned} \tag{6}$$

De este modo, se puede definir el retardo global como la suma de los generados por cada trama i, $g^i_{s_i,\pi_i}$, en lugar del experimentado por ellas:

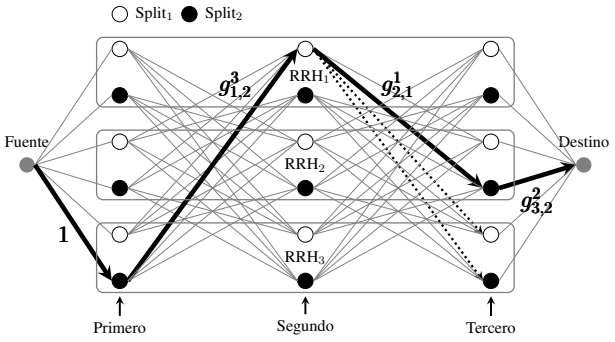


Fig. 1: Instancia del problema con 3 RRH y 2 posibles niveles de *split*. La línea continua resalta indica la solución seleccionada, mientras que las líneas discontinuas denotan soluciones no admisibles.

$$\begin{aligned} D(\boldsymbol{s}, \boldsymbol{\pi}) &= \sum_{i\in\mathcal{R}} g^i_{s_i,\pi_i} = \\ &= \sum_{i\in\mathcal{R}} \left(\delta^L_{i,s_i} + \delta^R_{i,s_i} + \delta^B_{i,s_i}(|\mathcal{R}| - \pi_i + 1) \right) \end{aligned} \tag{7}$$

A. Formulación del problema

Aunque pueden existir varios parámetros a optimizar, tales como la minimización del retardo máximo o de las diferencias de retardo entre tramas, este trabajo se centra en la reducción del retardo total del sistema.

De acuerdo con el modelo presentado en [3], el problema de optimización global del retardo se puede plantear definiendo el sistema como un grafo dirigido. Los nodos en el grafo forman una trama regular con $|\mathcal{R}|$ columnas y $|\mathcal{R}| \times |\mathcal{F}|$ filas, donde cada nodo se corresponde con una decisión conjunta de *split* y política de envío para una trama determinada, mientras que los arcos que conectan los nodos tienen pesos iguales al retardo asociado a dicha decisión. De este modo, las columnas en el grafo representan el orden de envío, mientras que las filas indican la decisión de *split* y la trama seleccionada. Finalmente, añadiendo dos nodos virtuales (*Fuente* y *Destino*), el problema se reduce a la búsqueda del camino más corto entre dichos nodos.

A fin de ilustrar el planteamiento del problema, la Figura 1 muestra el grafo resultante de un sistema compuesto por 3 RRH y 2 niveles de centralización. Como se puede ver, todas las posibles soluciones de *split* y orden de envío de cada RRH se agrupan en dos filas. En la figura se resalta una posible solución completa con líneas continuas, en las que el coste de cada enlace se corresponde con la decisión asociada el nodo fuente de dicho enlace. En el ejemplo la trama 3 se envía en primer lugar con nivel de *split* 2, a continuación la trama 1 con nivel 1 y, finalmente, la trama 2 usando nuevamente el nivel de *split* 2. Como se puede observar, tras seleccionar la trama 3 es necesario añadir restricciones adicionales, que garanticen que dicha trama no se selecciona de nuevo, lo que se indica en la figura con líneas discontinuas.

Se denota el grafo del sistema como $G(\mathcal{V}, \mathcal{A})$, donde $\mathcal{A}$ es el conjunto de enlaces y $\mathcal{V}$ el conjunto de vértices, de modo que v_0 y $v_{|\mathcal{R}|+1}$ se corresponden, respectivamente, con los nodos virtuales *Fuente* y *Destino* ($|\mathcal{V}| = |\mathcal{R}| + 2$). A continuación se define el subconjunto de nodos correspondientes a una RRH i como $\mathcal{V}_i \subseteq \mathcal{V}$.

La selección de los enlaces se realiza usando una variable binaria $x_{i,j}$, que toma valor 1 si se selecciona el enlace entre los nodos i y j, y 0 en caso contrario. Por simplicidad, se usará la variable w_{ij} para indicar el coste del enlace que conecta cada par de nodos (i, j). Finalmente, el problema global de reducción del retardo se puede expresar como:

Problem 1 (Selección conjunta de *split* y política de envío)**.**

$$\textbf{\textit{min.}} \quad \sum_{i,j} x_{ij} \cdot w_{ij} \tag{8}$$

$$\textbf{\textit{s.t.}} \quad \sum_{i \in \mathcal{V}/k} x_{ik} + \sum_{i \in \mathcal{B}/k} x_{ki} = T_k \qquad \forall k \in \mathcal{V} \tag{9}$$

$$\sum_{i \in \mathcal{V}/\mathcal{V}_i} x_{ik} = 1 \qquad \forall k \in \mathcal{V}_i \tag{10}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in \mathcal{V} \tag{11}$$

donde la Ecuación 9 se corresponde con la restricción de conservación de flujo; la constante T_k indica el flujo entrante y saliente de cada nodo. Esta constante toma valor 1 y -1 respectivamente para los nodos *Fuente* y *Destino* ($T_0 = 1; T_{|\mathcal{R}|+1} = -1$) y 0 para el resto. Mediante la Ecuación 10 se asegura que únicamente se toma una decisión por RRH, tal y como se mostró en la Figura 1 mediante las líneas discontinuas.

Como se puede observar, el problema resultante es binario y lineal, *Binary Linear Program* (BLP), que es conocido por ser *np-hard* y, por tanto, difícil de resolver. Además, el tamaño crece exponencialmente con el número de RRH y posibles niveles de *split*, siendo el espacio de posibles soluciones $|\mathcal{F}|^{|\mathcal{R}|} \times |\mathcal{R}|!$ y el número de variables $2|\mathcal{R}||\mathcal{F}| + (|\mathcal{R}| - 1) \times |\mathcal{F}||\mathcal{R}| \times (|\mathcal{F}||\mathcal{R}| - |\mathcal{F}|)$[1].

En las siguientes secciones se describen brevemente las modificaciones del problema original cuando bien el nivel de *split* o la política de envío son conocidas.

B. Selección de split *conocida*

Si se fija el nivel de *split* para cada RRH, se pueden conocer los valores de retardo, de modo que el problema se reduce a minimizar el producto del retardo en la BBU por el orden de transmisión. Se puede observar que la optimización global en este caso se obtiene usando una política que envía en primer lugar las tramas con menor retardo en la BBU (*shortest-job-first*).

[1]El primer término, $2|\mathcal{R}||\mathcal{F}|$, se corresponde con los enlaces salientes y entrantes a los nodos virtuales. En el segundo término se multiplica el número de columnas ($|\mathcal{R}| - 1$) por el número de filas $|\mathcal{R}||\mathcal{F}|$ y enlaces salientes de cada nodo ($|\mathcal{F}||\mathcal{R}| - |\mathcal{F}|$).

C. Política de envío conocida

En el caso en que se fije la política de envío, la complejidad del problema se reduce. En concreto se trataría de seleccionar el nivel de *split* que minimiza la expresión $\delta^L_{i,s_i} + \delta^R_{i,s_i} + \delta^B_{i,s_i}(|\mathcal{R}| - \pi_i + 1)$, donde π_i es conocido para cada RRH. Si se tiene en cuenta que, en la práctica, el posible número de *splits* es bajo, esta tarea se puede realizar de manera eficiente mediante simples algoritmos de búsqueda.

IV. EVALUACIÓN DE RENDIMIENTO

Como se ha visto anteriormente, la complejidad asociada al planteamiento general del problema podría no ser de utilidad práctica en casos reales, especialmente en escenarios con un número elevado de elementos de acceso. Por esta razón, parece razonable explorar alternativas subóptimas y analizar su rendimiento en diferentes escenarios.

En esta sección se analiza el comportamiento del problema general definido en Problema 1 y, a continuación, se compara con el resultado obtenido por las soluciones subóptimas que se han presentado anteriormente. Para realizar esta comparación se usarán 3 escenarios diferentes, en los que se varían los dos parámetros con mayor impacto en el problema: la relación de las capacidades computacionales entre la RRH y BBU y la longitud de las tramas, que es directamente proporcional a los retardos. A fin de obtener unos resultados más claros y fáciles de comparar, se usará la capacidad computacional de la BBU como referencia. De esta manera, los escenarios se definen en función de la relación de capacidades de procesado entre la RRH y BBU , $r^R_i = C_i/C^B$, y del retardo de procesado de esta última δ^B_{i,s_i} para una longitud de trama fija. Respecto al retardo de transmisión, se asume que se dispone de enlaces con alta capacidad [18], por lo que se considera despreciable en comparación con los otros.

En general, los escenarios están compuestos por una BBU y 10 RRH, y se analiza el comportamiento estadístico de cada algoritmo, realizando 1000 experimentos independientes para cada configuración. Las capacidades de cómputo de los elementos de la red están obtenidos de los modelos reales descritos en [19], [20] y para la longitud de las tramas se usan como referencia los valores mencionados en [21], [22]. Finalmente, cabe indicar que el nivel de *split* se define como la variación de carga de procesado de las tramas entre la BBU y RRH. Por lo tanto, el nivel de centralización de una RRH i para un *split* s viene dado por la expresión $w_{i,s_i}/(w_{i,s_i} + \hat{w}_{i,s_i})$, como se detalla en [23], [24]. El problema conjunto se soluciona usando la herramienta de optimización GLPK [25] y para cada uno de los escenarios se compara el comportamiento obtenido con:

- La solución usando una política de envío conocida, según la cual las tramas se envían en orden ascendente en relación a su tamaño.
- La solución con la selección de *split* conocida, aplicando niveles de centralización del 0 (red tradicional), 50, y 100% (C-RAN).

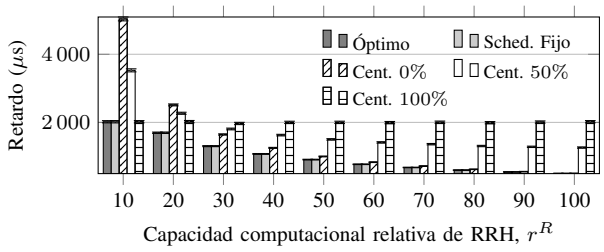


Fig. 2: Retardo medio por trama con longitud de trama heterogénea y diferentes valores de capacidad computacional de las RRHs

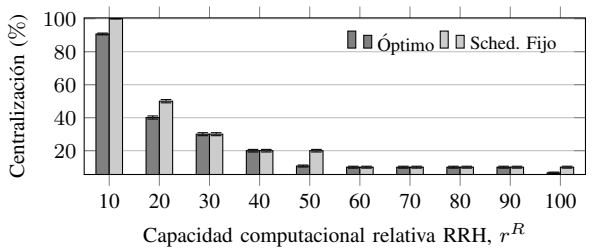


Fig. 3: Nivel medio de centralización con longitud de trama heterogénea y diferentes valores de capacidad computacional de las RRHs

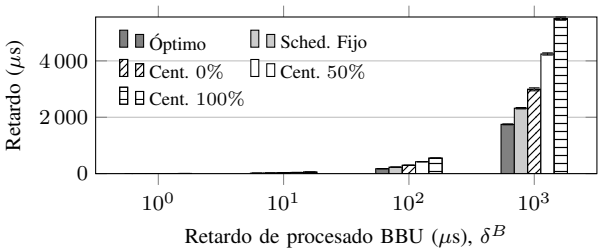


Fig. 4: Retardo medio por trama con capacidad de cómputo de las RRHs heterogénea y diferentes valores de longitud de trama

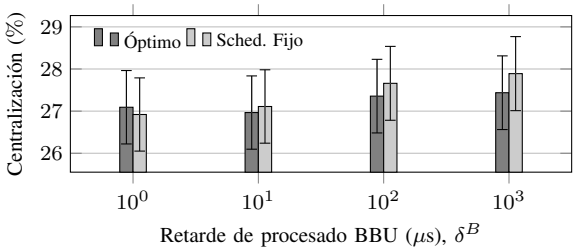


Fig. 5: Nivel medio de centralización con capacidad de cómputo de las RRHs heterogénea y diferentes valores de longitud de trama

A. RRHs homogéneos y tráfico heterogéneo

En el primer escenario se considera un despliegue de red homogéneo, con elementos de acceso iguales, de forma que todos tienen la misma capacidad de cómputo ($r_i^R = r^R \quad \forall i \in \mathcal{R}$). Por otro lado, el retardo de procesado en la BBU, δ^B, se distribuye uniformemente en el intervalo $[1, 1000]$ μs, en cada experimento independiente.

La Figura 2 muestra, para diferentes valores de la carga computacional en los elementos de acceso, el retardo medio por trama, así como su intervalo de confianza del 95%. Como se puede observar, la solución con política de envío fija (primero las tramas más cortas) muestra un comportamiento similar al óptimo global para todas las configuraciones. Por otro lado, el esquema con el nivel de *split* fijo manifiesta un comportamiento menos previsible. En este sentido, para configuraciones en las que los elementos de acceso tienen menor capacidad el mejor resultado es aquel en que se tiene un mayor nivel de centralización (C-RAN), como cabría esperar. En el otro extremo, cuando la capacidad de las RRH es similar a la de la BBU la mejor solución sería un esquema distribuido. En general, cuando se usa el esquema que fija un nivel de *split*, es necesario adaptar esa configuración al escenario concreto sobre el que se aplica.

En lo que se refiere a la selección del *split*, en la Figura 3 se muestra el nivel de centralización para las soluciones óptima y de política de envío fija. Como se puede ver, de nuevo, ambos esquemas tienen un comportamiento muy similar, de forma que son capaces de seleccionar el *split* de acuerdo a las características del escenario.

B. RRHs heterogéneos y tráfico homogéneo

En el segundo escenario se fija la longitud de las tramas, lo que da lugar a una configuración de tráfico homogéneo, que se tiene en cuenta fijando el valor del retardo de cómputo en la BBU, δ^B. Por otro lado, la capacidad relativa de las RRH, r_i^R, se selecciona de manera aleatoria en el intervalo $[0.1, 1]$.

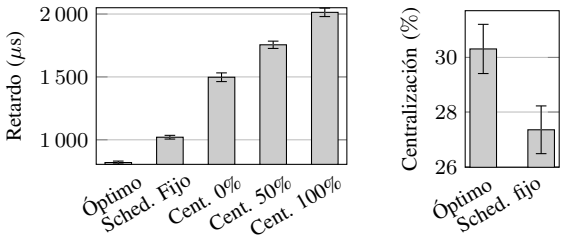
De forma similar al escenario anterior, en primer lugar se muestra en la Figura 4 el retardo medio de las tramas, representando también el intervalo de confianza. Como cabía esperar, con independencia del algoritmo seleccionado, el retardo aumenta con el tamaño de las tramas. Además, se puede ver que el esquema de política de transmisión fija siempre tiene un comportamiento mejor al mostrado por el de nivel de centralización fijo.

Seguidamente, la Figura 5 muestra el nivel de centralización seleccionado. En este caso, se puede observar que para valores de longitud de trama pequeños (menor valor de δ^B) la solución óptima selecciona niveles de centralización superiores, mientras que el comportamiento es precisamente el contrario cuando se incrementa la longitud de trama.

C. RRHs heterogéneos y tráfico heterogéneo

En el último escenario se selecciona de forma aleatoria tanto la longitud de trama como la capacidad computacional de las RRH usando los intervalos indicados anteriormente.

La Figura 6a muestra el retardo experimentado por las tramas cuando se usan los diferentes esquemas de selección. Se puede ver que, de forma similar a los casos anteriores, la política de transmisión fija siempre se comporta mejor que aquella en la que se fija el nivel



(a) Retardo medio

(b) Nivel de centralización

Fig. 6: Rendimiento de los algoritmos con longitud de trama y capacidad de cómputo de las RRHs heterogéneos

de centralización. Sin embargo, en este escenario ambas soluciones sub-óptimas están alejadas del óptimo global.

Finalmente, en la Figura 6b se puede apreciar que la solución óptima tiende a niveles de centralización superiores.

V. CONCLUSIONES

En este trabajo se han analizado diferentes algoritmos para minimizar el retardo en escenarios con *flexible functional split* en los que la selección del nivel de centralización y la gestión de envío de tramas afectan notablemente al retardo. Con esta premisa, resulta preciso tener en cuenta ambos aspectos de manera conjunta para obtener el mejor rendimiento del sistema en términos de retardo. Sin embargo, el problema de optimización subyacente resulta *np-hard*, por lo que su uso práctico podría verse comprometido. Por ello, se han analizado alternativas que, bajo ciertos supuestos, reducen de forma considerable la complejidad del problema original.

A partir del trabajo realizado por Koutsopoulos [3], se ha implementado un algoritmo que soluciona tanto el problema original como las alternativas que llevan a cabo optimizaciones parciales. Se han analizado y evaluado las diferentes soluciones en diversos escenarios, comparando los retardos y niveles de centralización de las optimizaciones parciales con aquellas proporcionadas por el problema original. A la luz de los resultados, en escenarios donde los elementos de acceso son homogéneos en términos de capacidad de procesado, el rendimiento de soluciones que consideren esquemas fijos de envío de tramas es similar al óptimo global. Además, también se ha observado que la selección del *split*, realizado por este tipo de optimización parcial, también es semejante al óptimo. Por otro lado, las soluciones con nivel de centralización fijo presentan un rendimiento notablemente peor que, a su vez, se ve afectado por las características concretas del escenario.

También se ha estudiado el comportamiento de las diferentes alternativas en escenarios donde la red de acceso es heterogénea. En este caso, el rendimiento mostrado por las optimizaciones parciales se encuentra lejos del óptimo global. Sin embargo, de acuerdo a los resultados presentados, nuevamente las soluciones con políticas de envío constante muestran un mejor comportamiento que aquellas en las que se fija en nivel de *split*.

Partiendo de este trabajo, en el futuro se pretende analizar diferentes alternativas a la optimización global, especialmente en escenarios con heterogeneidad en la red de acceso. En concreto, se analizarán técnicas de agrupamiento (*clustering*) que agruparían tramas o elementos de acceso con características similares. Así, se reduciría la complejidad del problema. Por otro lado, se analizarán escenarios más complejos en los que se considerarán procesos de llegada de tramas a la BBU menos predecibles, lo que precisará otro tipo de soluciones, tales como las basadas en teoría de colas.

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 de los proyectos ADVICE: *Dynamic provisioning of connectivity in high density 5G wireless scenarios* (TEC2015-71329-C2-1-R) y FIERCE: *Future Internet Enabled Resilient Cities* (RTI2018-093475-A-100).

REFERENCIAS

[1] I. . W. Group, "Next generation fronthaul interface." [Online]. Available: http://sites.ieee.org/sagroups-1914/

[2] "Study on new radio access technology: Radio access architecture and interfaces," 3rd Generation Partnership Project (3GPP), TR 38.801, 2017.

[3] I. Koutsopoulos, "Optimal functional split selection and scheduling policies in 5g radio access networks," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017, pp. 993–998.

[4] 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.

[5] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks—a technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[6] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5g," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 6, pp. 573–581, June 2018.

[7] A. Garcia-Saavedra, J. X. Salvat, X. Li, and X. Costa-Perez, "Wizhaul: On the centralization degree of cloud ran next generation fronthaul," *IEEE Transactions on Mobile Computing*, vol. 17, no. 10, pp. 2452–2466, Oct 2018.

[8] C. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink fronthaul for soft ran," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 82–88, Sep. 2015.

[9] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5g mobile crosshaul networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, Firstquarter 2019.

[10] P. Arnold, N. Bayer, J. Belschner, and G. Zimmermann, "5g radio access network architecture based on flexible functional control / user plane splits," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[11] Y. Alfadhli, M. Xu, S. Liu, F. Lu, P. Peng, and G. Chang, "Real-time demonstration of adaptive functional split in 5g flexible mobile fronthaul networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[12] D. Harutyunyan and R. Riggio, "Flex5g: Flexible functional split in 5g networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 961–975, Sep. 2018.

[13] ——, "Flexible functional split in 5g networks," in *2017 13th International Conference on Network and Service Management (CNSM)*, Nov 2017, pp. 1–9.

[14] D. A. Temesgene, M. Miozzo, and P. Dini, "Dynamic functional split selection in energy harvesting virtual small cells using temporal difference learning," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2018, pp. 1813–1819.

[15] L. Wang and S. Zhou, "Flexible functional split in c-ran with renewable energy powered remote radio units," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[16] A. Marotta, D. Cassioli, K. Kondepu, C. Antonelli, and L. Valcarenghi, "Efficient management of flexible functional split through software defined 5g converged access," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[17] Y. Li, J. Mårtensson, M. Fiorani, B. Skubic, Z. Ghebretensaé, Y. Zhao, J. Zhang, L. Wosinska, and P. Monti, "Flexible ran: A radio access network concept with flexible functional splits and a programmable optical transport," in *2017 European Conference on Optical Communication (ECOC)*, Sep. 2017, pp. 1–3.

[18] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5g network capacity: Key elements and technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, Mar. 2014.

[19] P. Rost, I. Berberana, A. Maeder, H. Paul, V. Suryaprakash, M. Valenti, D. Wübben, A. Dekorsy, and G. Fettweis, "Benefits and challenges of virtualization in 5g radio access networks," *IEEE Communications Magazine*, vol. 53, no. 12, pp. 75–82, Dec. 2015.

[20] K. Wang, K. Yang, H. Chen, and L. Zhang, "Computation diversity in emerging networking paradigms," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 88–94, Feb. 2017.

[21] X.-L. Wu, W.-M. Li, F. Liu, and H. Yuand, "Packet size distribution of typical internet applications," in *2012 International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP)*, Dec. 2012, pp. 276–281.

[22] Z. Sun, D. He, L. Liang, and H. Cruickshank, "Internet qos and traffic modelling," *IEE Proceedings - Software*, vol. 151, no. 5, pp. 248–255, Oct. 2004.

[23] N. Makris, P. Basaras, T. Korakis, N. Nikaein, and L. Tassiulas, "Experimental evaluation of functional splits for 5g cloud-rans," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[24] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5g signal processing: Flexible centralization through cloud-ran," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[25] G. Project, "Gnu linear programming kit." [Online]. Available: https://www.gnu.org/software/glpk/


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Modelado del control de potencia del enlace ascendente para redes heterogéneas y densas basadas en OFDMA

Luis Diez y Ramón Agüero
Departamento de Ingeniería de Comunicaciones, Universidad de Cantabria
Plaza de la Ciencia s.n. 39005, Santander, Cantabria. España
{ldiez, ramon}@tlmat.unican.es

***Resumen*—En este trabajo se presenta un modelado novedoso del enlace ascendente en redes heterogéneas, que se ha planteado para su uso en simulación a nivel de sistema. A diferencia de otros trabajos, el modelo propuesto tiene en cuenta la interferencia mutua causada por las conexiones de otros usuarios, y establece los niveles de potencia mínimos necesarios para satisfacer una Relación Señal a Ruido e Interferencia (SINR) arbitraria. El planteamiento da lugar a un problema de optimización lineal que se puede resolver de manera sencilla con diferentes herramientas. A fin de validar el modelo, se ha comparado el rendimiento obtenido al aplicarlo con el observado al utilizar un esquema tradicional de lazo cerrado, implementado de forma iterativa. El análisis demuestra que el modelo propuesto proporciona, en una sola iteración, los mismos resultados que los métodos de simulación iterativos, permitiendo, por lo tanto, reducir la complejidad de las simulaciones. Finalmente, se ha estudiado el comportamiento del modelo en escenarios heterogéneos, haciendo uso de diferentes técnicas de selección de acceso.**

***Palabras Clave*—Control de potencia, LTE, HetNets, optimización, modelado.**

I. INTRODUCCIÓN

En los últimos años se ha podido ver la aparición de servicios bi-direccionales, tales como los basados en *Mobile Edge Computing* (MEC) [1], [2], lo que llevará posiblemente a un aumento significativo del volumen de tráfico en el enlace ascendente de las comunicaciones celulares. Por otro lado, la creciente densidad de la red de acceso, así como las diferencias entre los elementos que la componen, plantea nuevos retos para la definición de mecanismos de gestión, cobrando especial relevancia la interferencia entre celdas. Este aspecto es, incluso, más relevante cuando se hace uso de nuevos esquemas de selección de acceso, tales como *cell-range extension* (CRE) o *Downlink Uplink Decoupling* (DUDe).

En este contexto, el control de potencia del enlace ascendente, que minimice el nivel de interferencia, puede mejorar de manera significativa el rendimiento de la red. Al mismo tiempo, incrementar eficiencia la energética permitiría aumentar la autonomía de los dispositivos de usuario, *User Equipment* (UE).

En análisis de los diferentes mecanismos de gestión en general, y de la gestión de potencia en particular, se suele abordar mediante técnicas de simulación. Dependiendo del nivel de precisión exigido, y la forma en que la red se modela, existen tres tipos principales de simulación [3]: simulación de red, simulación a nivel de enlace y a nivel de sistema.

En el primer tipo se modelan los elementos de red de forma realista, lo que suele requerir la implementación de la pila de protocolos. En el otro extremo, la simulación a nivel de enlace se centra en el comportamiento del enlace radio, típicamente entre un único par transmisor-receptor, y considera aspectos concretos de las capas física y de acceso al medio, así como propagación a pequeña escala temporal; por ejemplo, a nivel de símbolo. Finalmente, la simulación a nivel de sistema busca la caracterización del comportamiento global de la red. En este sentido, este último tipo de simulación se centra en fenómenos que ocurren en escalas temporales mayores que aquellos estudiados a nivel de enlace, tales como selección de acceso o traspasos. Por ello, el modelado a nivel de sistema hace uso de valores estadísticos obtenidos a partir de estudios a nivel de enlace, o modelos analíticos, para tener en cuenta el comportamiento de los enlaces.

A su vez, la simulación a nivel de sistema puede tener dos enfoques distintos, estático o dinámico [3], de acuerdo a la posibilidad de considerar la evolución temporal de la red (por ejemplo, patrones de tráfico o movimiento). La simulación estática se basa en análisis *Monte Carlo*, en los que se usan múltiples fotografías independientes del escenario para obtener figuras de rendimiento, tales como nivel de cobertura. Por otro lado, el enfoque dinámico busca estudiar la evolución temporal del sistema, para lo que se considera que la red cambia a lo largo del tiempo, y la simulación mantiene el estado del sistema

entre fotografías consecutivas. Esta metodología permite analizar parámetros como los traspasos, o incluir patrones de tráfico de servicios.

De las diferentes opciones de simulación, para el estudio de técnicas de control de potencia normalmente se adopta la alternativa a nivel de sistema, usando en enfoque estático o dinámico, en función de los objetivos concretos. En el caso particular de la gestión de potencia en el enlace ascendente para sistemas basados en OFDMA, se han propuestos diferentes modelos, tal como se elaborará en la Sección II. En la mayoría de los estudios se asume únicamente control de potencia en lazo cerrado, lo que puede dar lugar a resultados poco precisos. En los escasos casos en los que se modela el control de potencia en lazo cerrado, el análisis requiere una mayor complejidad de simulación, a fin de modelar la convergencia de este método, por tanto limitando su aplicabilidad para estudios que incluyan fenómenos que tienen lugar en escalas temporales mayores, tales como los traspasos.

En este trabajo se propone un modelo analítico que permite modelar el control de potencia en lazo cerrado del enlace ascendente en sistemas basados en OFDMA, para su posterior aplicación en simulación, tanto estática como dinámica, a nivel de sistema. El modelo propuesto da lugar a um problema de optimización que tiene como objetivo minimizar la potencia total de transmisión del sistema, asegurando, a la vez, que las conexiones alcanzan un nivel determinado de *Signal to Interference plus Noise Ratio* (SINR). En este sentido, a diferencia de los enfoques existentes, la solución que se presenta proporciona el nivel de potencia que se obtendría aplicando el control de potencia en lazo cerrado, sin necesidad de realizar varias iteraciones para que el sistema converja. En concreto, la formulación propuesta da lugar a un problema de optimización lineal, que se puede resolver con multitud de herramientas. La validez del modelo se ha evaluado comparando los resultados que proporciona con aquellos que se obtendrían mediante una implementación iterativa. Posteriormente se ha usado el modelo para analizar diferentes estrategias de selección de acceso en redes heterogéneas, así como su impacto sobre los niveles de potencia necesarios.

El resto del documento se estructura como sigue. En la Sección II se presenta un análisis de la literatura relaciona existente. A continuación, el modelado del sistema se describe en la Sección III, en la que se presenta, además, el problema de optimización resultante. A continuación, en la Sección IV se evalúa el modelo propuesto, y se analiza el rendimiento, en términos de potencia, de diferentes técnicas de selección de acceso. Finalmente, el artículo concluye en la Sección V, donde se resumen las contribuciones y se enumeran líneas futuras de investigación.

II. TRABAJOS PREVIOS

A pesar de la creciente importancia de la gestión de la potencia del enlace ascendente en redes celulares, no hay mucha literatura reciente relacionada con este aspecto. Algunos trabajos previos se centraron en analizar las diferencias entre las soluciones de control de potencia en lazo abierto y cerrado en redes LTE [4], [5]. Más recientemente, los autores de [6] han comparado ambos esquemas de control de potencia, concluyendo que la técnica de lazo cerrado tiene un mejor rendimiento, y que los niveles de potencia obtenidos en lazo abierto no son siempre apropiados. Sin embargo, muchos de los estudios existentes se centran en analizar las técnicas en lazo abierto [7], buscando optimizar sus parámetros de configuración [8], [9], [10].

En lo que se refiere al modelado del control de potencia del enlace ascendente, algunos estudios previos has abordado el análisis estático a nivel de sistema, o planificación de red. Por ejemplo, en [11] se presenta un modelo general, mientras que los autores de [12] proponen uno específico para sistemas basados en WCDMA. También se han presentado modelos para redes heterogéneas basadas en OFDMA, tales como el descrito en [13], donde se considera el impacto de la interferencia entre celdas para obtener el rendimiento de la red. Por otro lado, los autores de [14] hacen uso de técnicas de geometría estocástica para definir un marco matemático para el análisis a nivel de sistema que ayude en el diseño de redes. De forma similar, en [15] se analiza el impacto de las topologías heterogéneas sobre el enlace ascendente, prestando especial atención a la interferencia entre elementos de acceso de diferentes capas. En relación con la selección de acceso, en [7] se analizan diferentes configuraciones del control de potencia del enlace ascendente, usando técnicas basadas en *Reference Signal Received Power* (RSRP) y CRE. Sin embargo, la mayoría de los trabajos existentes asumen políticas de acceso sencillas, como la distancia entre estación base y usuario [16].

Como se puede observar, la mayoría de los modelos propuestos asumen notables simplificaciones del sistema [16], lo que puede llevar a resultados poco precisos. Por otro lado, el uso de modelos más realistas requiere un tiempo simulación que podría resultar excesivo [4].

Como respuesta a esta problemática, el modelo propuesto en este trabajo tiene como objetivo proporcionar el mismo nivel de precisión que la simulación intensiva, sin necesitar el proceso iterativo para que la técnica de control de potencia converja. A continuación se resumen las principales contribuciones:

- Definición de un modelo analítico del control de potencia en lazo cerrado para sistemas basados en OFDMA. Cabe mencionar que no se trata de una propuesta de control de potencia, sino que tiene como objetivo emular el comportamiento del sistema real.
- El modelo se ha diseñado para ser usado en simulación a nivel de sistema, tanto estática como dinámica. El objetivo es proporcionar resultados similares a los obtenidos mediante la simulación del control de potencia en lazo cerrado, evitando la sobrecarga de las iteraciones necesarias para que el problema converja.
- En el modelo se considera el impacto de la interferencia mutua entre usuarios, aspecto que normalmente es ignorado. Dado lo genérico de la solución propuesta,

el nivel de interferencia se puede modular para tener en cuenta, de forma estadística, técnicas que mitiguen dicha interferencia.
- El planteamiento del modelo da lugar a un problema de optimización lineal, que puede ser resuelto con un gran número de herramientas y puede, por lo tanto, ser fácilmente integrado en entornos de simulación.

III. DESCRIPCIÓN DEL MODELO

En esta sección se describen los esquemas de control de potencia tradicionales de LTE. Tomando como base estas definiciones, posteriormente se presenta el modelo analítico para la configuración en lazo cerrado.

A. Esquemas de control de potencia del enlace ascendente

Se puede definir la potencia a transmitir, en lazo cerrado, por un UE como sigue [17]:

$$P = \min\{P_{\max}, P_0 \cdot N_{RB} \cdot \gamma^{\alpha} \cdot \delta_{\text{mcs}} \cdot f(\Delta)\} \qquad (1)$$

donde $P_{\max}$ es la potencia máxima que puede transmitir el UE, P_0 es un parámetro que refleja el nivel de interferencia esperado, N_{RB} es el número de recursos asignados al usuario, y los parámetros γ y α se corresponden, respectivamente, con las pérdidas de propagación y el factor de corrección de dichas pérdidas. El parámetro δ_{mcs} indica un nivel de potencia adicional que depende del *Modulation and Coding Scheme* (MCS) y que es específico para cada UE. Finalmente, $f(\Delta)$ es la función de corrección de lazo cerrado, que permite realizar ajustes de potencia de acuerdo a las variaciones del canal.

Tanto el parámetro P_0 como α se envían desde la red a los usuarios, mientras que la estimación de las pérdidas por propagación las obtiene el UE usando el nivel de RSRP. A partir de estos tres parámetros, que configuran el control de potencia en lazo abierto, el usuario puede establecer el nivel de potencia sin indicaciones adicionales desde la red, siendo incluso capaz de compensar variaciones lentas del canal.

Posteriormente, la estación base indica, mediante el parámetro $f(\Delta)$, sucesivos ajustes del nivel de potencia (lazo cerrado), de modo que el valor usado se desvía del obtenido mediante los parámetros en lazo abierto para compensar las variaciones del canal. Se puede encontrar una explicación más detallada de los procedimientos de control de potencia en [18].

B. Modelo del sistema

Se considera un escenario compuesto por un conjunto de usuario y celdas LTE, respectivamente $\mathcal{U}$ y $\mathcal{B}$, y se asume que todas las celdas comparten los mismos recursos espectrales, que se modelan como bloques asignables o *physical resource block* (PRB); los cuales son la unidad básica de asignación de recursos LTE. Por lo tanto, las conexiones de todos los usuarios se ven interferidas por las del resto, excepto por aquellos conectados a la misma celda. Finalmente, se asume que la selección de acceso ya ha tenido lugar, por lo que se conocen las conexiones de los usuarios. Cabe indicar que no se hace ninguna suposición sobre la técnica concreta de selección acceso.

La celda que da servicio (conectividad) a un usuario i se define como $\beta(i) \in \mathcal{B}$, mientras $\mathcal{U}(k) \subseteq \mathcal{U}$ es el subconjunto de usuarios conectados a la celda k, y C_k la capacidad de dicha celda, en número de PRBs. Por otro lado, γ_{ik} indica las pérdidas de propagación entre el usuario i y la celda k, mientras que, de forma particular, Γ_i se corresponde con las pérdidas con respecto a la celda a la que se conecta el usuario, $\Gamma_i = \gamma_{i,\beta(i)}$. El resto de los parámetros tienen el mismo significado que el indicado anteriormente.

A partir de estos parámetros se define la interferencia por PRB experimentada por la conexión del usuario i, en su celda, I_i, como sigue:

$$I_i = \sum_{k \in \mathcal{B}/\beta(i)} \sum_{j \in \mathcal{U}(k)} N_{RB_j} P_0 \ \delta_{\text{mcs}_i} \frac{\gamma_{jk}^{\alpha}}{\gamma_{j\beta(i)}} f(\Delta)_i \ S(i,j,C_k,C_{\beta(i)}) \qquad (2)$$

donde N_{RB_j} es la cantidad de recursos utilizados por el usuario interferente j. Además, se ha introducido el parámetro $S(i,j,C_k,C_{\beta(i)})$ para modelar, deforma estadística o determinista, la interferencia debida a otros usuarios como una función que dependería del *scheduling* y técnicas de reducción de interferencia. Por ejemplo, si se usara una política de *scheduling* concreta y determinista únicamente aquellos usuarios que transmiten en los mismos PRBs generarían interferencia. Por otro lado, dado que el objetivo de este trabajo es introducir el modelo, este parámetro se modela de forma simplificada asumiendo que no hay coordinación entre celdas y que el *scheduler* es aleatorio, de modo que $S(i,j,C_k,C_{\beta(i)}) = 1/C_k$.

Como se puede observar, la interferencia se define como la suma normalizada de los recursos asignados a otros usuarios, $j \in \mathcal{U}(k)$, $k \in \mathcal{B}/\beta(i)$, ponderada por el nivel de potencia interferente de dichos usuarios en la celda con la que se conecta el usuario i, $\beta(i)$

Entonces, la potencia transmitida por el usuario i, P_i, se puede definir en términos del nivel de SINR objetivo, St_i, y la interferencia correspondiente, de modo que $St_i \leq \frac{P_i/\Gamma_i}{\sigma^2+I_i}$, donde $P_i = P_0\Gamma_i^{\alpha}\delta_{\text{mcs}_i} f(\Delta)_i$. Si se agrupan todos los términos dependientes, se puede expresar la SINR de un usuario como una combinación lineal de la potencia transmitida por otros usuarios;

$$P_i - St_i \sum_{k \in \mathcal{B}/\beta(i)} \sum_{j \in \mathcal{U}(k)} \frac{N_{RB_j} P_j \Gamma_i}{C_k \gamma_{j\beta(i)}} \geq St_i \Gamma_i \sigma^2 \qquad (3)$$

Si se asume que cada UE transmitirá con la mínima potencia requerida para satisfacer su SINR objetivo, la configuración de potencia del enlace ascendente, en lazo cerrado, de todos los usuarios se puede modelar como un problema de optimización lineal, tal como se muestra a continuación:

Problem 1 (Asignación de potencia)**.**

$$\textbf{\textit{min}} \quad \sum_{i \in \mathcal{U}} P_i \tag{4}$$

$$\textbf{\textit{s.t.}} \quad P_i - St_i \sum_{k \in \mathcal{B}/\beta(i)} \sum_{j \in \mathcal{U}(k)} \frac{N_{RB_j} P_j \Gamma_i}{C_k \gamma_{j\beta(i)}} \geq St_i \Gamma_i \sigma^2 \quad \forall i \in \mathcal{U} \tag{5}$$

$$P_i \geq 0 \quad \forall i \in \mathcal{U} \tag{6}$$

Como se puede observar, la interferencia mutua que se considera en el Problema 1 proporciona, en una única evaluación, el nivel de potencia que se obtendría una vez que el proceso de lazo cerrado converge. Además, cabe destacar que el resultado del problema, P_i, incluye tanto los parámetros de lazo abierto como cerrado, por lo que los valores de estos dependerán de la configuración concreta. Por ejemplo, si se asumen valores fijos de P_0 y α, los parámetros en lazo cerrado (δ_{mcs_i} y $f(\Delta)_i$) se despejarían, usando la expresión $P_i = P_0 \Gamma_i^{\alpha} \delta_{\text{mcs}_i} f(\Delta)_i$.

Aunque no se incluyen en el modelo los desvanecimientos rápidos (*fast fading*), parece razonable asumir que su contribución se anularía (media cero) en el tiempo de convergencia del mecanismo de lazo cerrado. Por otro lado, aunque en el planteamiento del modelo se ha asumido que todas las celdas comparten los mismos recursos frecuenciales (escenario mono-portadora), el modelo se puede extender a escenarios multi-portadora, en la que los recursos asignados en diferentes bandas frecuenciales no interferirían. Por ejemplo, en un escenario con diferentes portadoras para las capas de estaciones macro y *small* no se consideraría la interferencia entre capas.

IV. VALIDACIÓN DEL MODELO Y ANÁLISIS DE ESCENARIOS

El modelo propuesto se ha validado en un escenario LTE con dos capas de celdas, macro y *small*. El primer nivel está compuesta por 7 estaciones base, con 3 sectores cada una, desplegadas según un patrón hexagonal. Por su parte, la segunda capa la conforman un número variable de *small-cells*, desplegadas de manera aleatoria dentro del área de cobertura de las estaciones base macro. Se considera que todos los elementos de acceso comparten una banda de 20 MHz (100 PRBs) y que no existe cooperación entre ellos. Sobre este escenario se despliegan 100 usuarios, que se consideran siempre activos y, por lo tanto, continuamente transmitiendo. En la Tabla I se resumen los parámetros principales del escenario.

Dado que el objetivo principal de este trabajo es validar el modelo propuesto, los resultados presentados en esta sección se han obtenido mediante simulación estática a nivel de sistema; sin embargo, cabe destacar que el modelo puede ser igualmente aplicable a simulación dinámica, en la que existan patrones de movilidad y tráfico. En cada iteración los usuarios y *small-cells* se despliegan aleatoriamente y el nivel de potencia recibido en las estaciones base se calcula para cada usuario. En función de dichos niveles, así como de otros parámetros, se lleva a cabo la selección de acceso y, a continuación, se aplica el mecanismo de control de potencia para establecer los niveles de potencia de transmisión de cada usuario. En los casos en que se utiliza el modelo propuesto, el problema de optimización se resuelve con la librería GLK[1]. En cuanto a la selección de acceso, se han estudiado las siguientes:

- Selección de acceso clásica basada en RSRP: cada usuario se conecta a la celda de la que recibe el mayor nivel de potencia. Mientras que esta técnica presenta un buen comportamiento en redes homogéneas, no es capaz de explotar la capacidad adicional de las *small-cells* debido a que no tiene en cuenta las diferentes potencias de transmisión de los diferentes tipos de estaciones base.
- *cell-range extension* (CRE): a fin de subsanar las limitaciones del la estrategia anterior, en este caso se añade un factor de *bias* a la potencia recibida desde las estaciones base en función de su tipo. De esta forma se incrementa de forma virtual el área de cobertura de las celdas con menor potencia de transmisión, favoreciendo que los usuarios se conecten a ellas.
- *Downlink Uplink Decoupling*: mientras que las técnicas anteriores se centran en las comunicaciones del enlace descendente, esta última trata de optimizar, en términos de potencia, ambos sentidos de la comunicación de manera individual. De este modo, los usuarios se conectan para el enlace descendente de acuerdo al nivel de RSRP y la conexión del enlace ascendente se realiza con la celda con la que haya menores pérdidas de propagación.

Tabla I: Características del escenario

Escenario LTE	**20MHz @2.1GHz**
Capa macro	*Inter Site Distance* (ISD) 500 m, 7 celdas con 3 sectores Max. potencia transmisión 46 dBm Ganancia de antena 14 dBi, 15^o down-tilt NF 3 dB
Capa *small*	Despliegue aleatorio Max. TX. power 30 dBm Antena-omni, 5.0 dBi NF 3 dB
Path Loss (Tables B.1.2.1-1 and B.1.2.1-2 from [19])	Macro-BS ↔ UE: *Urban macro* (UMA) Small-BS ↔ UE: *Urban micro* (UMi)
UE	Max. potencia transmisión 24 dBm (250 mW) Ganancia de antena 0 dB
PUSCH *Maximum Allowable Path Loss* (MAPL)	140 dB
δ_{mcs}	0 dB, i.e. independiente del MCS
α	1, i.e. compensación total

A. Validación del modelo

El primer objetivo de la validación es comprobar que el comportamiento del modelo propuesto es semejante el que se obtendría implementado el control en lazo cerrado

[1]https://www.gnu.org/software/glpk/

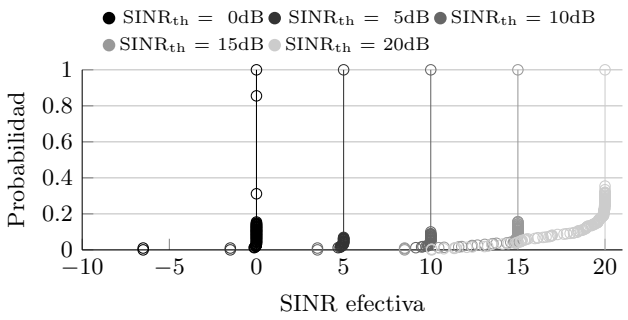


Fig. 1: *cdf* de la SINR efectiva usando el modelo propuesto (líneas continuas) y lazo cerrado iterativo (marcadores)

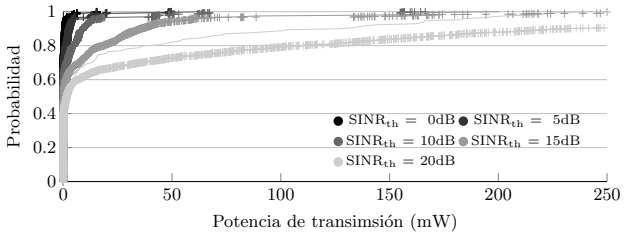


Fig. 2: *cdf* de la potencia de transmisión usando el modelo propuesto (líneas continuas) y lazo cerrado iterativo (marcadores)

de manera iterativa. Por simplicidad, este primer análisis se ha llevado a cabo sobre un escenario homogéneo, compuesto únicamente por estaciones base macro. Para la implementación iterativa se ha seguido el enfoque descrito en [4], donde en primer lugar se obtiene el componente en lazo abierto para establecer el nivel de potencia inicial. A continuación se calcula la SINR efectiva de las conexiones, y se ajusta la potencia en cada usuario hasta que se alcance el nivel de SINR requerido; este proceso se repite hasta que la potencia de todos los usuarios se estabiliza (el sistema converge). Por otro lado, el modelo propuesto se aplica en una escala de tiempo superior, de forma que se usan realizaciones independientes del escenario para cada instancia del problema. A fin de realizar una comparativa justa entre ambos procesos no se han incluido desvanecimientos, por lo que las diferencias se deben sólo a la interferencia mutua entre usuarios.

La Figura 1 muestra la *cumulative distribution function* (CDF) de la SINR efectiva obtenida para diferentes valores de SINR objetivo de los usuarios. Como se puede observar, la solución propuesta siempre proporciona el valor deseado, como consecuencia de incluir la interferencia mutua en su definición. Por otro lado, la implementación iterativa presenta una distribución ligeramente más dispersa, debido a las iteraciones necesarias para alcanzar la convergencia del sistema. En concreto se puede apreciar que a medida que el requisito de SINR es más exigente, la distribución presenta más valores por debajo del objetivo, lo cual indica un mayor tiempo de convergencia.

De manera similar, en la Figura 2 se muestra la distribución de la potencia de transmisión cuando se usan ambas implementaciones del control de potencia en lazo cerrado; de nuevo los resultados se han obtenido para varios valores de SINR objetivo. De acuerdo a los resultados se puede concluir que ambas soluciones presentan comportamientos similares, apareciendo diferencias notables sólo para valores de SINR objetivo muy altos (20 dB).

En general, se puede concluir que el modelo propuesto presenta un comportamiento similar a la implementación iterativa, sin la complejidad añadida al tiempo de convergencia. En concreto, para los escenarios analizados, se ha observado que la implementación iterativa requiere entre 4 y 12 ciclos para converger. Además, cabe destacar que los tiempos de convergencia tienen una gran dependencia de la geometría de la red y densidad de usuarios, por lo que el número de iteraciones se incrementaría en escenarios más complejos. Por otro lado, el modelo propuesto obtiene los niveles de potencia en una sola iteración, equivalente a una instancia del problema, por lo que reduce la complejidad de la simulación. El análisis en profundidad de la reducción de complejidad se abordará en trabajos futuros, una vez que el modelo ha sido validado.

B. Análisis de selección de acceso

Tras evaluar el rendimiento del modelo, se pretende analizar las diferencias en niveles de potencia obtenidas mediante los esquemas de lazo abierto y cerrado cuando se aplican diferentes técnicas de selección de acceso. Para obtener los siguientes resultados se ha utilizado el modelo propuesto para el esquema de lazo cerrado, por lo que ambos esquemas de control de potencia tienen la misma complejidad de simulación. A diferencia del análisis anterior, en este caso se despliega un escenario heterogéneo compuesto tanto por estaciones macro como *small-cells*, de acuerdo a los parámetros descritos en la Tabla I. Además, la comparativa se ha realizado usando diferentes técnicas de selección de acceso para analizar las interacciones del control de potencia y dichas estrategias. En concreto, se compara la selección tradicional basada en RSRP con dos configuraciones de CRE, usando niveles de *bias* de 3 y 9 dB, y con DUDe. Todos los resultados se han obtenido para valores de SINR objetivo de 5 dB, aunque se observaron resultados similares con otros valores.

En la Figura 3 se muestra la distribución de la SINR, para diferentes densidades de *small-cells*, cuando se aplica el esquema en lazo abierto. Los resultados en lazo cerrado no se muestran porque, como se mostró en la Figura 1, usando la el modelo propuesto siempre se obtiene el nivel de SINR requerido. Como se puede apreciar, los resultados indican que el esquema en lazo cerrado no asegura la calidad de servicio requerida, sino que presenta un una distribución bastante dispersa, que depende tanto de la densidad de *small-cells* como de la técnica de selección de acceso. En concreto se puede apreciar que las selecciones de acceso RSRP y CRE$_3$ muestran las distribuciones más dispersas.

Finalmente se comparan los niveles de potencia, por PRB, tanto para el esquema en lazo abierto como cerrado. Aunque ambas configuraciones presentan diferentes

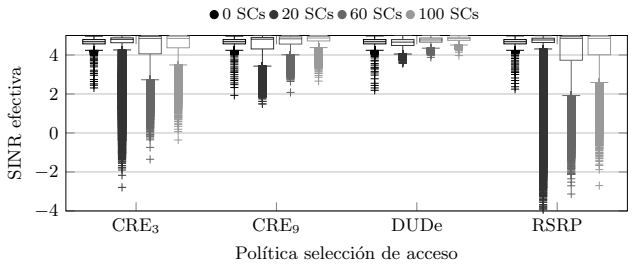
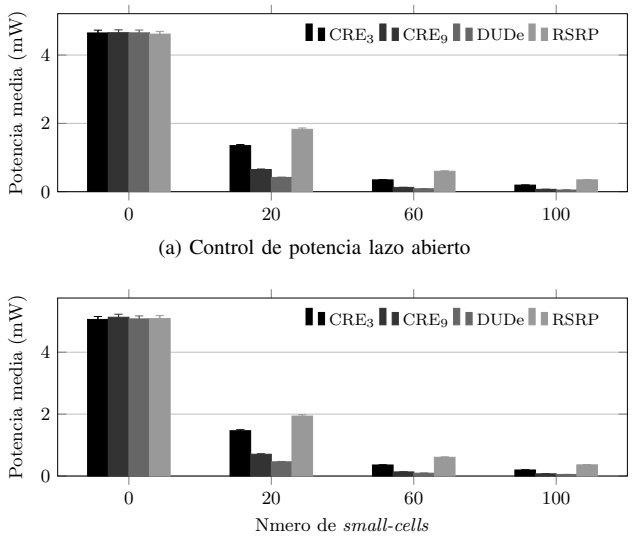


Fig. 3: Distribución de la SINR efectiva usando el esquema en lazo abierto



(a) Control de potencia lazo abierto

(b) Modelo propuesto

Fig. 4: Potencia requerida por PRB para diferentes configuraciones de selección de acceso y densidad de *small-cells*

comportamientos en términos de SINR efectiva, la Figura 4 evidencia que la potencia requerida es similar en ambos casos. Esto es consecuencia de que el esquema en lazo cerrado optimiza el reparto de potencia para reducir los niveles de interferencia. Cabe destacar que esta optimización tendría lugar tanto con el modelo propuesto como con la implementación iterativa.

V. CONCLUSIONES

Las comunicaciones en enlace ascendente en redes celulares no han recibido mucha atención desde la comunidad científica. Sin embargo, su importancia está aumentando notablemente, como consecuencia de la aparición de nuevos servicios y topologias, lo que pone de manifiesto la necesidad de mecanismos que permitan mejorar la gestión de sus recursos. A su vez, la evaluación de las diferentes soluciones precisan de herramientas y modelos de simulación capaces de tener en cuenta la complejidad de los nuevos despliegues y topologías de red.

En este trabajo se ha propuesto un modelo analítico para el control de potencia en lazo cerrado del enlace ascendente, que permite reducir la complejidad de simulación de manera notable con respecto a los enfoque tradicionales. Este modelo se ha validado comparando su comportamiento con las técnicas de modelado tradicionales, que están basadas en implementaciones iterativas, y requieren por tanto un mayor tiempo de simulación. Los resultados han puesto de manifiesto que el modelo propuesto proporciona resultados similares a las técnicas tradicionales, reduciendo la complejidad del análisis de forma significativa.

Además, aprovechando la funcionalidad del modelo, se ha analizado la interacción entre la políticas de selección de acceso y esquemas de control de potencia, tanto en lazo abierto como cerrado. Los resultados muestran que ambos aspectos están estrechamente relacionados, con una fuerte dependencia de la geometría de la red.

En trabajos futuros se seguirá haciendo uso del modelo propuesto para acometer el análisis de sistema de redes densas. En concreto, nos centraremos en escenarios con un gran número de dispositivos de usuario, que serán más relevantes en despliegues futuros. Por otro lado, se llevará a cabo un análisis en detalle de la reducción de la complejidad de simulación que permite el modelo. También se pretende extender su definición para considerar técnicas de *scheduling* y cooperación entre celdas.

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 de los proyectos ADVICE: *Dynamic provisioning of connectivity in high density 5G wireless scenarios* (TEC2015-71329-C2-1-R) y FIERCE: *Future Internet Enabled Resilient Cities* (RTI2018-093475-A-100).

REFERENCIAS

[1] V. Frascolla, F. Miatton, G. K. Tran, K. Takinami, A. D. Domenico, E. C. Strinati, K. Koslowski, T. Haustein, K. Sakaguchi, S. Barbarossa, and S. Barberis, "5G-MiEdge: Design, standardization and deployment of 5G phase II technologies: MEC and mmWaves joint development for Tokyo 2020 Olympic games," in *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*, Sep. 2017, pp. 54–59.

[2] M. Weichold, M. Hamdi, M. Z. Shakir, M. Abdallah, G. K. Karagiannidis, and M. Ismail, *Cognitive Radio Oriented Wireless Networks: 10th International Conference, CROWNCOM 2015, Doha, Qatar, April 21-23, 2015, Revised Selected Papers*, 1st ed. Springer Publishing Company, Incorporated, 2015.

[3] X. Chu, D. Lopez-Perez, Y. Yang, and F. Gunnarsson, *Heterogeneous Cellular Networks: Theory, Simulation and Deployment*. New York, NY, USA: Cambridge University Press, 2013.

[4] A. Simonsson and A. Furuskar, "Uplink Power Control in LTE - Overview and Performance. Principles and Benefits of Utilizing rather than Compensating for SINR Variations," in *2008 IEEE 68th Vehicular Technology Conference*, Sept 2008, pp. 1–5.

[5] Y. Wang and S. Venkatraman, "Uplink power control in LTE heterogeneous networks," in *2012 IEEE Globecom Workshops*, Dec 2012, pp. 592–597.

[6] J. Turkka, Olivia, M. D. Villaluz, and S. Foo, "Optimization of LTE uplink performance in multivendor heterogeneous networks," in *2018 International Conference on Information Networking (ICOIN)*, Jan 2018, pp. 374–379.

[7] K. Safjan, S. Strzyż, K. I. Pedersen, J. Steiner, and C. Rosa, "Automatic methods for HetNet uplink power control optimization under fractional load," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sept 2013, pp. 3056–3060.

[8] A. Haider, S.-H. Lee, S.-H. Hwang, D. I. Kim, and J. H. Na, "Uplink open loop power control for LTE HetNet," in *2016 URSI Asia-Pacific Radio Science Conference (URSI AP-RASC)*, Aug 2016, pp. 83–85.

[9] K. Safjan, S. Strzyż, K. I. Pedersen, J. Steiner, and C. Rosa, "Open Loop Power Control parameter settings impact on LTE HetNet uplink performance," in *2013 IEEE International Conference on Communications Workshops (ICC)*, June 2013, pp. 1134–1138.

[10] H. Martikainen, I. Viering, and B. Wegmann, "Dynamic range aware LTE uplink P0 optimization in HetNet," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[11] K. L. Clarkson, K. G. Hampel, and J. D. Hobby, "Modeling UpLink Power Control with Outage Probabilities," in *2007 IEEE 66th Vehicular Technology Conference*, Sep. 2007, pp. 799–803.

[12] R. Patachaianand and K. Sandrasegaran, "System-Level Modeling and Simulation of Uplink WCDMA," in *Fifth International Conference on Information Technology: New Generations (itng 2008)*, April 2008, pp. 1071–1076.

[13] H. Tabassum, F. Yilmaz, Z. Dawy, and M. Alouini, "A Statistical Model of Uplink Inter-Cell Interference with Slow and Fast Power Control Mechanisms," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3953–3966, Sep. 2013.

[14] M. Di Renzo and P. Guan, "Stochastic Geometry Modeling and System-Level Analysis of Uplink Heterogeneous Cellular Networks With Multi-Antenna Base Stations," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2453–2476, June 2016.

[15] F. Wang and W. Wang, "Analytical modeling of uplink power control in two-tier femtocell networks," in *2015 Wireless Telecommunications Symposium (WTS)*, April 2015, pp. 1–6.

[16] S. Essassi, M. Siala, R. Hamila, M. O. Hasna, and S. Cherif, "Power control and RB allocation for LTE uplink," in *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Sept 2016, pp. 417–421.

[17] 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.

[18] B. Muhammad and A. Mohammed, "Uplink closed loop power control for LTE system," in *2010 6th International Conference on Emerging Technologies (ICET)*, Oct. 2010, pp. 88–93.

[19] 3rd Generation Partnership Project (3GPP), "Technical Specification Group Radio Access Network; Further Advancements for E-UTRA," 2017. [Online]. Available: http://www.3gpp.org/dynareport/36814.htm


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Lightweight Testbed for Machine Learning Evaluation in 5G Networks

Carlos Hernández-Chulde, Cristina Cervelló-Pastor
Department of Network Engineering,
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
{carlos.hernandez, cristina}@entel.upc.edu

***Abstract*—The adoption of Software Define Networking, Network Function Virtualization and Machine Learning will play a key role in the control and management of fifth-generation (5G) networks in order to meet the specific requirements of vertical industries and the stringent requirements of 5G. Machine learning could be applied in 5G networks to deal with issues such as traffic prediction, routing optimization and resource management. To evaluate the adoption of machine learning in 5G networks, an adequate testing environment is required. In this paper, we introduce a lightweight testbed, which utilizes the benefits of container lightweight virtualization technology to create machine learning network functions over the well-known Mininet network emulator. As a use case of this testbed, we present an experimental real-time bandwidth prediction using the Long Short Term Memory recurrent neural network.**

***Keywords*—5G, SDN, NFV, machine learning, containers**

I. INTRODUCTION

The fifth generation (5G) of communication networks will bring with new requirements, such as high data rates, high traffic densities, low latency and high reliability, and use cases such as the Internet of Things (IoT) and critical communication applications. These requirements and use cases impose new challenges, which demand efficient, intelligent and agile network management. Additionally, 5G will create an ecosystem that increases innovation opportunities for new applications in vertical industries such as manufacturing, healthcare, media and entertainment, financial services, public safety, the automotive industry, public transportation, energy utilities, food and agriculture, and city management. Each of these has a specific set of requirements in latency, throughput, availability, reliability, coverage, mobility, ans so on. 5G will provide a flexible network that caters to such varied requirements. Network flexibility implies a high degree of softwarization, virtualization and automation [1]. From the network perspective, Software Defined Networking (SDN) is considered to be the materialization of the softwarization concept, and Network Function Virtualization (NFV) of the virtualization paradigm [2]. One key component in enabling network flexibility is network slicing, as it allows us to create tailored logical networks on top of a common shared physical infrastructure in order to efficiently satisfy the specific needs of each vertical industry. A network slice involves a set of network functions and resources that are required to run these network functions. SDN and NFV can provide the programmability, flexibility and modularity that are necessary to create network slices [3].

Since SDN and NFV allow network functions to run in software instead of being tightly coupled with hardware, they provide flexibility and reconfigurability to the network. Thus, network functions can be modified, updated and placed at any location in the network. However, the dynamic behavior of network functions introduces complexities and makes the provisioning, management and control of network slices impractical in a manual way. In this dynamic environment, continuous monitoring and network analytics become compulsory to understand the network behavior. Similarly, providing the network with automation capabilities is essential for network operation and management. Network automation reduces operational costs, avoids human error and accelerates the service time to market.

Besides, the application of machine learning (ML) to network analytics provides the network with learning and decision-making capabilities. ML techniques can extract relevant information from the network data and then utilize this knowledge for autonomic network control and management, as well as service provisioning. Based on historical and real-time data, ML mechanisms can predict network behavior and adapt it to the new network conditions by allocating the required amount of network resources without overprovisioning. ML can also be used for energy-saving optimization. If the current demand is low, it may be possible to switch off some elements or migrate services to locations with lower energy costs in order to optimize energy consumption. ML may be

effectively applied in automatic network orchestration and network management, making self-organizing networks feasible. In other words, ML is a key enabler of automation and contributes to addressing the problem of deploying network intelligence. In this context, SDN and NFV combined with ML are key enablers of 5G networks [4].

In this respect, it is worth mentioning that standardization entities are working in this field. The 3rd Generation Partnership Project (3GPP) has introduced a Network Data Analytics Function (NWDAF) in the 5G System Architecture. NWDAF is defined as an operator-managed network analytics logical function that can provide slice-level network data analytics to a network function [5]. The European Telecommunications Standards Institute (ETSI) has created an Industry Specification Group (ISG) called Experiential Networked Intelligence (ENI). The ENI system is an innovative context-aware entity that enables intelligent service operation and management applying technologies, such as big data analysis and artificial intelligence mechanisms to adjust offered services based on changes in user needs, environmental conditions and business goals [6].

In this scenario, as ML has recently received much attention as a key enabler of the control and management of 5G networks, researchers need tools to design, test and evaluate it. Researchers face various difficulties when testing ML applications due to infrastructure limitations, the expense or difficulty in building physical testbeds, or the unavailability of emulation platforms. Thus, in this paper we present an emulation test platform that is able to emulate ML as network functions using Mininet and Docker containers to facilitate the development and testing of ML applications in 5G networks. Network functions are executed inside Docker containers that are interconnected through the underlying Mininet-based emulation environment.

The remainder of this paper is structured as follows. In Section II, theoretical background is reviewed. In Section III, we introduce the testbed architecture and detail its components. Section IV presents the results of experimentation results consisting of traffic prediction using the Long Short Term Memory (LSTM) recurrent neural network as a ML technique. Finally, in Section V, conclusions and plans for future work are presented.

II. BACKGROUND

A. Network Functions Virtualization (NFV)

NFV transforms the way in which operators design and manage networks by employing virtualization technology [7]. NFV decouples specialized network functions from hardware and implements them as Virtual Network Functions (VNFs). VNFs are implemented in software and deployed on commercial off-the-shelf (COTS) servers [8]. Multiple VNFs can be connected in order to create complex network services (NSs), which are managed by a management and orchestration system (MANO).

By separating network functions from hardware, NFV offers several advantages over traditional network architectures: (1) reduced equipment footprint and power consumption, as it is possible to collapse multiple network functions into a single physical server; (2) rapid service development and deployment, making network upgrade tasks easier; (3) longer hardware life cycles; and (4) reduced maintenance costs. These benefits mean that NFV enhances flexibility and scalability while reducing CApital and OPerational EXpenditure (CAPEX and OPEX) [7].

B. Containers

Since NFV involves implementing network functions in software, virtualization technologies such as virtual machines (VMs) and containers play an important role in VNFs' development. Prior to the deployment of VNFs in production environments, VNFs must be tested in confined, lightweight environments [9]. Researchers and developers use these environments in the development and prototyping of new NSs. In these environments workloads run as software instances over VMs or containers.

Containers and VMs provide application isolation and bundle applications with all of their dependencies in a self-contained unit that can run anywhere. Both share physical computing resources, allowing for efficient use in terms of energy consumption and cost. Although the goals of containers and VMs are similar, the approach to achieving them differs. While VMs provide hardware virtualizations, containers provide operating-system-level virtualization.

Containers are a more lightweight virtualization technology than VMs; unlike VMs, containers do not require a hypervisor, as VMs do. A VM also needs its own operating system, which means that each VM runs a full copy of an operating system, regardless of whether the operating system is the same on two or more VMs. This adds an overhead, as starting an operating system occupies time, memory and storage.

Containers run on the top of the host operating system, sharing the kernel. Applications running in containers share operating-system-level architecture that provides them with basic services. Containers require an underlying operating system that provides the basic services to all of the containerized applications. By sharing operating system resources, the need to replicate operating system code is significantly reduced, which means that a server can run multiple containerized applications with a single operating system installation. Therefore, containers are very lightweight in terms of size and starting time. In other words, this means that by using containers, we can run more application instances on a single server than we can with VMs.

Containers utilizes two kernel features, such as namespaces and control groups (cgroups), to create virtual environments on top of an operating system. Namespaces provide a layer of isolation by limiting what a container can view and access, such as processes trees, networking resources or file system. When a container runs, the kernel creates a separate namespace that the container will use. Thus, this container's access is limited to that namespace. In contrast, cgroups provide resource allocation. With

cgroups, the kernel create groups of processes for resource management purposes. Cgroups allow granular control over resources by limiting or prioritizing system resources such as CPU time, system memory, network bandwidth, or combinations of these. In this sense, cgroups ensure that the containers use the resources that they require [10].

C. GPU Usage in ML

As mentioned in Section I, ML will play a crucial role in the operation and management of 5G networks. Applying ML in network analytics enables the intelligent use of network-generated data. ML will provide the network with insights into traffic patterns, available resources, potential security threats and user behavior, allowing the network to proactively adapt or change its behavior based on previous knowledge of these issues.

In 5G, the number of devices connected to the network is expected to grow exponentially. Therefore, the amount of data collected for network control and management will also increase. Moreover, it is envisioned that 5G will involve a combination of different technologies such as heterogeneous networks, cloud computing or edge computing. In this complex ecosystem with large volumes and varied types of data, the application of ML in network analytics will require a significant computational power, which Graphical Processing Units (GPUs) can provide.

The execution of ML workloads can be accelerated by using GPUs. A GPU has more numerous and smaller cores than a CPU. As GPUs have many cores and each core performs rapid calculations simultaneously, they are highly suitable for parallel processing. Thus, the use of GPUs in ML is a cost-effective and high-performance option in comparison to traditional CPUs.

III. TESTBED ARCHITECTURE

The application of ML to provide the network with a certain degree of intelligence has attracted the attention of several standards bodies and industry forums. ML techniques enable the network to make autonomous decisions by processing large amounts of network data. As mentioned in Section I, 3GPP has included a dedicated function called NWDAF in 5G system architecture for the purposes of data collection and data analytics.

At this point, it is worth mentioning that 3GPP's 5G system architecture is service based. In a Service-based Architecture (SBA), the architecture elements are defined as network functions that offer their services via a common bus known as Service-Based Interface (SBI). Network functions that are allowed to make use of the provided services can directly communicate with each other as originators or consumers. The SBA model takes advantage of the latest virtualization and software technologies, such as containers, to offer modularity, extensibility, reusability and self-containment in network functions. NWDAF is one key function within SBA, facilitating access to network data analytics. Consumer network functions decide how the data analytics provided by NWDAF are used to improve the network performance. For example, the Policy Control Function (PCF) may use per slice data analytics in its policy decisions, or the Network Slice Selection Function (NSSF) may use the load-level analytic information for slice selection.

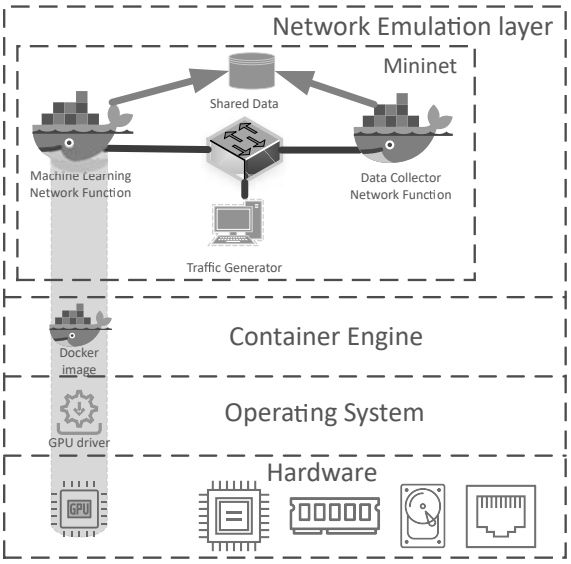


Fig. 1. Testbed architecture

In addition, the application of ML in network analytics requires both a module to monitor and collect data from the network and a module to apply ML techniques to extract knowledge from the data collected. The data collection module's role is to gather and store sufficient information from different sources to understand the current state of the network. It also performs data preprocessing to ensure that only useful data is stored. Relevant collected data may include network configuration, traffic data, control and management data, application and service-level data, and even external information, such as social networks [11]. The collected data are transformed into knowledge via ML in the second module. This module is the key component, as it is responsible for choosing the best ML algorithm that fits a certain problem or use case. In this module, based on real-time and historical network data, ML techniques can bring intelligence to the network by providing useful insights about its current and future states. The outcome of this module can be used by network controller or management systems to make decisions (either automatically or through human intervention) in order to optimize the use of network resources and enhance the provision of NSs.

In this context, it is necessary to have a platform ready for the development, prototyping and evaluation of network functions that provide network data analytics services such as NWDAF. The main objective of this work is to integrate the concepts and technologies described in Section II into a testbed, using the well-known Mininet network emulator and one of the most commonly-used container engines, Docker. The testbed architecture is ilustrated in Fig. 1.

In this testbed, data analytics network functions run as containerized applications within Docker containers.

Table I
TESTBED COMPONENTS

Component	Testbed Component
Network Emulator	Containernet
GPU Support for Containers	NVIDIA Container Runtime
Container Engine	Docker
Host Operating System	Ubuntu Bionic Beaver
GPU	NVIDIA GPU

There are two types of containers: one for data collection and one for ML application. The latter container type executes ML algorithms using GPUs. Since the used GPU is an NVIDIA GPU, the Docker containers use NVIDIA Container Runtime to access the GPU. NVIDIA Container Runtime simplifies the process of building and deploying containerized GPU-accelerated applications and guarantees the best performance on NVIDIA GPUs. Similarly, to provide interconnection between Docker containers on the top of Mininet, a fork of Mininet called Containernet is used [12]. Containernet extends the Mininet network emulator to allow the use of standard Docker containers as Mininet virtual hosts within the emulated network. In addition, Containernet allows the user to add or remove containers from the emulated network and to change resource limitation at runtime. Finally, any traffic generation tool, such as Iperf, can be used to generate traffic. The traffic generator host can be either a Mininet host or a Docker container.

This testbed, therefore, provides a framework for developing, testing and evaluating the application of ML in 5G networks in a simple, flexible and lightweight manner. Table I summarizes the components of this testbed.

IV. EXPERIMENTAL EVALUATION

We conducted an experiment to validate whether our testbed is lightweight and easy to use. The experiment consisted of predicting the traffic of a network via the use of ML; specifically, we used LSTM to do this. The experiment was carried out on a single physical machine which featured a CPU Intel(R) Core(TM) i9-9900K 3.60 GHz, 64 GB of RAM, running Ubuntu 18.04. The GPU used was an NVIDIA GeForce RTX 2080 with 2944 built-in cores and 8 GB of GDDR6 dedicated memory.

A. Traffic Prediction and LSTM

Network traffic prediction is an important issue in network operations and management, especially with regard to such diverse and complex networks as 5G networks. The aim of traffic prediction is to forecast the volume of future traffic by analyzing historical traffic information. Based on the results of traffic prediction, the network can make decisions in advance and adopt suitable preemptive actions to ensure its smooth operation, before a network overload occurs. These actions may include proactive routing policies or the provision of network resources.

Traffic prediction has been addressed via time series forecasting (TSF) [13]. Recent advances in deep learning have demonstrated that Recurrent Neural Networks (RNN) are powerful tools for TSF [14]. In our experimentation we used LSTM RNN for traffic prediction.

Time series traffic forecasting uses past traffic measurements to forecast future traffic patterns. For example, given a traffic measurement $x(t)$ at a time t, one can obtain a time series of $\{x(t), t = 1, 2, \ldots\}$. Traffic prediction consists of estimating the traffic at a future time $x(t+m)$ given n previous measurements, i. e.,

$$x(t+m) = f(\{x(n), n = 1, 2, \ldots, t\}) \tag{1}$$

LSTM is a special case and the most commonly used type of RNN. It is capable of learning long-term dependencies, which means that it can remember information that was previously learned. LSTM comprises multiple layers formed by one or more memory cells. A cell is responsible for memorizing values over time. Each cell is composed of three basic units: the input, output and forget gates that control information flow in an LSTM cell. The gates decide whether to forget, keep, update or output previously acquired information. LSTM is the most successful model for predicting long-term time series [15].

The input vector of our LSTM traffic prediction neural network corresponds to the recent traffic measurements, i.e., $x = [x(t), x(t-1), \ldots, x(t-n)]$, while the output vector is the predicted traffic in a future time $y = [x(t+1), x(t+2), \ldots, x(t+m)]$. Since LSTM networks retain past memory, traffic prediction for time interval $[t+1, t+m]$ is not only determined by the recent traffic measurements in $[t-n, t]$ but also indirectly by traffic measurements before $t-n$ through the memory cells.

B. Experimental Results

In our experiment, we deployed a simple topology over Mininet that consisted of a traffic generator host and the corresponding traffic sink, the data collector function and the ML function. All of these components were connected via a Mininet switch. Using Iperf, the traffic generator host generated traffic based on the dataset described later in this section. The data collector and ML network functions were Docker containers. One of the advantages of using Docker is that it offers us a repository of container images called the Docker hub. In this repository, we can find many containerized applications ready for use. As we intended to test ML algorithms, we used a container image that includes TensorFlow and CUDA. TensorFlow is an open-source platform for ML that provides a complete and flexible set of tools and libraries for ML development, whereas CUDA is a parallel computing platform which allows to harness the power of the NVIDIA GPUs, accelerating the ML workload execution.

To emulate the data collection function, we developed a Python script that periodically collected statistics from Mininet's switch interfaces and stored the collected data in a shared volume. The ML function accessed the collected data that were stored in the shared volume and used these data to train the LSTM RNN. Once the ML model had been trained, it was stored in the shared volume for later use in real-time traffic prediction. The training task is an

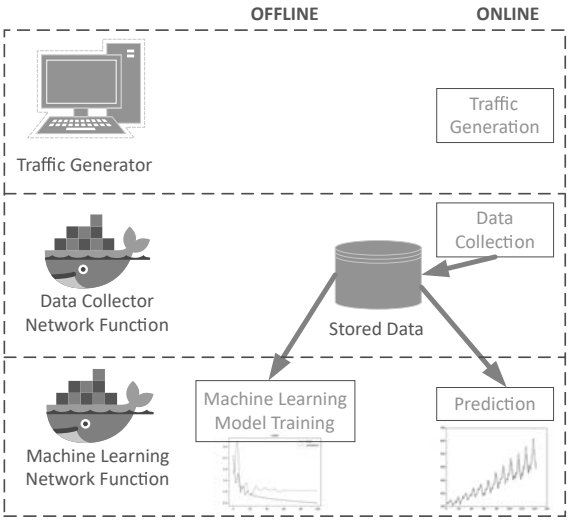
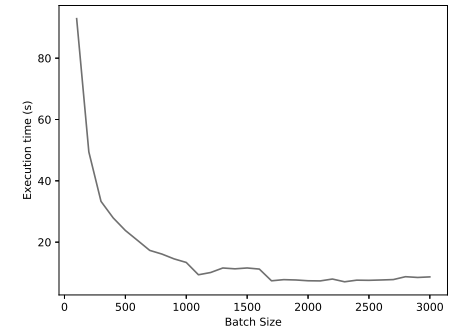
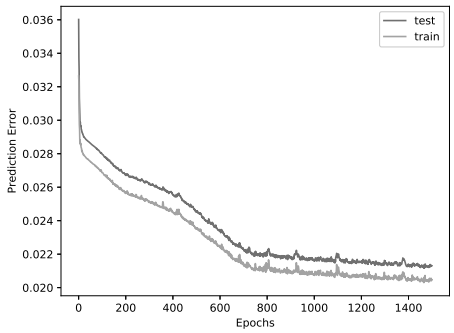


Fig. 2. Experiment details



(a) Batch size



(b) Epochs

Fig. 3. LSTM parameters

offline task that uses historical data for training LSTM model and can periodically retrain the model with the collected data, while the prediction task is an online task that makes prediction when a new traffic measurement is received. We also developed training and prediction tasks as Python scripts. It is worth mentioning that containerized network functions are executed on demand, which means that they run only for as long as it takes to execute the Python scripts execution, thus optimizing the use of computational resources. Fig. 2 presents the details of the experimental testbed.

The process of training our ML model is described below. We used a dataset from [16]. The dataset contained information about the traffic generated on a cellular network and provided hourly data on traffic statistics for each base station. The dataset consisted of (1) a base station identifier, (2) the date and time in UNIX format, (3) the number of users associated with the base station, (4) packets and (5) bytes transferred by the base station at the indicated time. In our case, we took the information of bytes and the date and time of two base stations to predict the traffic that each base station would use in the future.

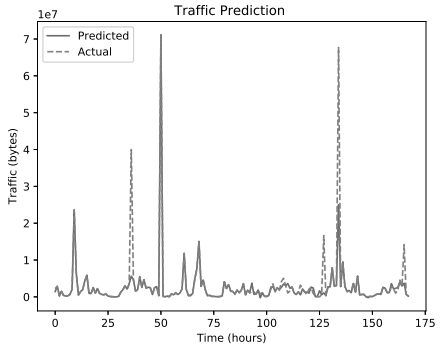
Specifically, we forecasted the traffic demand for each base station in the next hour, based on 24 past measurements ($n = 24$ (1 day)).

We generated training samples using a sliding window-based approach [14]. For example, to predict the traffic in the next hour ($m = 1$) based on the past 24 traffic measurements ($n = 24$), we used every consecutive 25 measurements as one training sample. The first 24 measurements became the input vector, and the 25th measurement in the training sample was used as the output label.

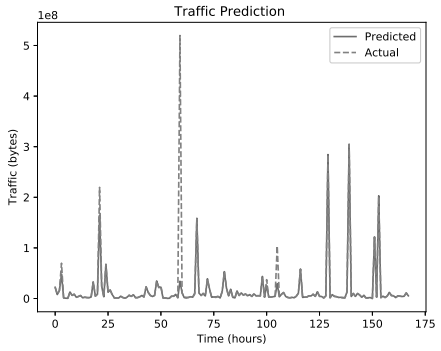
As the dataset was week-long, we reproduced the same data for the previous six months for training purposes. The dataset was divided in 80% training and 20% testing. Using the trained model, we predicted future traffic. Given that the LSTM architecture is characterized by the number of epochs and the batch size, we performed a set of experiments in order to identify the optimal values for these parameters to minimize the prediction error and execution time. The batch size is the number of training samples used in each iteration. We chose the value for the batch size that minimized the execution time. Fig. 3(a) shows that a batch size of above 1700 minimizes the execution time. In addition, the number of epochs determines the maximum number of passes over the training dataset. Different values for the number of epochs were tested in order to identify the optimal one that minimizes the prediction error. Thus, in the Fig. 3(b) it is evident that after 700 epochs, there is not a considerable improvement in prediction error, so this value was chosen in the LSTM.

We configured the LSTM network with one hidden layer, 100 neurons, an Adam optimizer with default values, 700 epochs and a batch size of 1700. Fig. 4 presents the traffic prediction results for the two base stations. The prediction values are very similar to the actual values, so this model was considered as a valid model for traffic prediction in this dataset.

In order to validate the computational overhead, we conducted offline training and online prediction, both in the CPU and GPU. In addition, to evaluate the overhead introduced by containers, we performed the same tasks on Docker containers and directly on the host. The training and prediction overhead in terms of processing time are presented in Table II. From the results in this table, it is evident that the training time is longer than the prediction time which is very short. However, this is not a problem



(a) Base station 1



(b) Base station 2

Fig. 4. Traffic prediction results

Table II
PROCESSING TIME FOR TRAINING AND PREDICTION

Processing Time (s)				
	Host		Container	
	CPU	GPU	CPU	GPU
Training	14.350	9.195	14.390	9.312
Prediction	0.062	0.067	0.064	0.069

in traffic prediction, because training is an offline task and once the training is completed, the trained model can be used for real-time prediction.

As expected, the processing times for training and prediction on the containers and the host were similar; this is because containers can access the hardware directly through the operating system. Using containers does not lead to a virtualization overhead unlike in VMs with the hypervisor. Finally, using the GPU reduced training runtime, since GPUs allow parallel computing over a large number of cores, running thousands of threads at a time. GPU and CPU prediction times are quite similar, as prediction is a small workload and does not require a large number of threads. It is evident that the prediction time on the CPU is slightly lower than on the GPU, due to higher frequency of the CPU cores; the frequency in the CPU is 3600 MHz, whereas GPU's frequency is 1515 MHz.

V. CONCLUSIONS AND FUTURE WORK

This work presents the use of Docker containers and Mininet to build a lightweight testbed with the aim to evaluate the application of ML in 5G networks. In this testbed, the functions that perform network analytics using ML run as containerized NFVs. This paper also describes how containers can run ML algorithms on GPUs.

In our future work, we intend to integrate the testbed with an SDN controller and MANO system to test a comprehensive network ecosystem, in which the output of ML network functions will assist in the decision-making process to apply adequate policies and configuration parameters in the network. Likewise, we will use this testbed to assess the introduction of a distributed network analytics architecture for 5G networks applying distributed ML approaches.

REFERENCES

[1] A. Bosneag and M. X. Wang, "Intelligent network management mechanisms as a step towards 5G," in *2017 8th International Conference on the Network of the Future (NOF)*, Nov 2017, pp. 52–57.

[2] M. Condoluci and T. Mahmoodi, "Softwarization and virtualization in 5G mobile networks: Benefits, trends and challenges," *Computer Networks*, vol. 146, pp. 65–84, 2018.

[3] J. Ordonez-Lucena *et al.*, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, May 2017.

[4] T. S. Buda *et al.*, "Can machine learning aid in delivering new use cases and scenarios in 5G?" in *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, April 2016, pp. 1279–1284.

[5] 3GPP, "System Architecture for the 5G System," 3rd Generation Partnership Project (3GPP), Technical Specification (TS), Dec 2018.

[6] ETSI, "Experiential networked intelligence (ENI); ENI use cases," Experiential Networked Intelligence ETSI ISG, Group Report, Apr 2018.

[7] T. N. Tavares *et al.*, "NIEP: NFV Infrastructure Emulation Platform," in *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, May 2018, pp. 173–180.

[8] ETSI, "Network Functions Virtualisation (NFV); Architectural Framework," Network Functions Virtualisation ETSI ISG, Group Specification, Dec 2014.

[9] S. van Rossem *et al.*, "Monitoring and debugging using an SDK for NFV-powered telecom applications," in *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, Nov 2016.

[10] RedHat. Introduction to Linux Containers. [Accessed: May 24, 2019]. [Online]. Available: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux_atomic_host/7/html/overview_of_containers_in_red_hat_systems/introduction_to_linux_containers

[11] A. Mestres *et al.*, "Knowledge-defined networking," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 47, no. 3, pp. 2–10, Sep. 2017.

[12] M. Peuster, H. Karl, and S. van Rossem, "Medicine: Rapid prototyping of production-ready network services in multi-pop environments," in *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, Nov 2016, pp. 148–153.

[13] R. Boutaba *et al.*, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," *Journal of Internet Services and Applications*, vol. 9, no. 1, p. 16, Jun 2018.

[14] L. Mei *et al.*, "Realtime Mobile Bandwidth Prediction Using LSTM Neural Network," in *Passive and Active Measurement*, D. Choffnes and M. Barcellos, Eds. Cham: Springer International Publishing, 2019, pp. 34–47.

[15] A. Pelekanou *et al.*, "Provisioning of 5G services employing machine learning techniques," in *2018 International Conference on Optical Network Design and Modeling (ONDM)*, May 2018, pp. 200–205.

[16] [Online]. Available: https://github.com/caesar0301/city-cellular-traffic-map. [Accessed: Jun 1, 2019].

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Diseño de laboratorio de prácticas para las asignaturas de Redes Seguras y Seguridad en Comunicaciones de un Máster de Ciberseguridad

Francisco J. Nóvoa[1], Diego Fernández[1], Raúl R. Rubio[2], Miguel Rodríguez[2], Fidel Cacheda[1], Víctor Carneiro[1] y Carlos Dafonte[1]

1 - Departamento de Ciencias de la Computación y las Tecnologías de la Información y las Comunicaciones, Universidade da Coruña 2 - Departamento de Ingeniería Telemática. Universidade de Vigo

Universidade da Coruña, CITIC y Universidade de Vigo

Facultad de Informática. Campus de Elviña. Universidade da Coruña. 15071. A Coruña

Esc. de Ing. en Telecomunicación. Campus Lagoas-Marcosende, C/ Maxwell, s/n, 36310 Vigo, Pontevedra

{fjnovoa, diego.fernandez}@udc.es, {rrubio, miguel}@det.uvigo.es, {fidel, viccar, dafonte}@udc.es

***Resumen*- En el curso 2018-2019, la Facultad de Informática de la Universidade da Coruña y la Escuela de Ingeniería de Telecomunicación de la Universidade de Vigo han puesto en marcha el primer Máster Interuniversitario de Ciberseguridad acreditado por ANECA en Galicia. Se trata de una titulación eminentemente técnica y orientada a la práctica, donde la Ingeniería Telemática tiene una especial relevancia al contar con dos asignaturas de seis créditos en el primer cuatrimestre denominadas "Seguridad en Comunicaciones" y "Redes Seguras". Entre los múltiples retos que se han tenido que afrontar, se encuentra el diseño y puesta en marcha de los laboratorios de prácticas asociados a estas materias. En este documento se explica cuál ha sido el proceso de preparación, planificación, diseño, implementación, operación y optimización del entorno de prácticas para las asignaturas.**

***Palabras Clave*- ciberseguridad, laboratorio, docencia**

I. INTRODUCCIÓN

En septiembre de 2018, comenzó en la Facultad de Informática de la Universidade da Coruña (UDC) y la Escuela de Telecomunicaciones de la Universidade de Vigo, el primer Máster Universitario en CiberSeguridad (MUniCS). Se trata de un máster oficial acreditado por ANECA que se imparte simultáneamente en los dos centros y su plan de estudios [1] es de 90 créditos.

La titulación tiene un marcado carácter técnico. En ella se pueden encontrar asignaturas como "Redes Seguras", "Seguridad en Comunicaciones" o "Fortificación de Sistemas Operativos" que tienen una clara orientación hacia la seguridad defensiva, mientras que otras como "Test de Intrusión" presentan fundamentalmente técnicas ofensivas. Otras materias combinan ambos aspectos. En el diseño de la docencia de todas las materias hemos intentado mantener el equilibrio entre teoría y práctica con el objetivo de que los alumnos adquieran sólidos fundamentos y habilidades que puedan poner en práctica de forma inmediata al incorporarse al ámbito laboral.

Concretamente, la asignatura "Redes Seguras" [2] tiene como objetivo principal que los estudiantes aprendan a diseñar e implementar infraestructuras de red que sean capaces de proporcionar los servicios de seguridad necesarios en un entorno corporativo. Deben conocer las arquitecturas de seguridad de referencia y ser capaces de desplegarlas y administrarlas de forma segura utilizando diferentes tecnologías como *firewalls*, sistemas de detección y prevención de intrusiones, VPNs, etc. La materia está concebida para que las prácticas de laboratorio tengan una importancia capital en el proceso de aprendizaje.

En lo que respecta a "Seguridad en Comunicaciones" [3] se realiza una revisión por capas de la arquitectura de comunicaciones de Internet, mostrando sus principales debilidades desde el punto de vista de la seguridad y proporcionando las técnicas y herramientas necesarias para mitigarlas. Los estudiantes deben conocer en detalle los protocolos de red que aportan seguridad a la transmisión de la información, y las implicaciones derivadas del lugar que ocupan dentro de la arquitectura en la que se organiza el software de comunicaciones.

Una vez diseñadas las guías docentes de estas materias, los profesores participantes en las mismas tomaron la decisión de diseñar e implantar un laboratorio de prácticas basado en dispositivos físicos (switches, *routers*, firewalls) que permita que los alumnos adquieran las habilidades prácticas planteadas y que los objetivos de aprendizaje puedan ser alcanzados por los alumnos en el mayor grado posible.

Los docentes diseñamos e implementamos el laboratorio de prácticas siguiendo una metodología clásica reconocida como PPDIOO (*Prepare, Plan, Design, Implementation, Operation and Optimization)* [4], utilizada habitualmente para el desarrollo de proyectos de red en entornos no académicos y en los que los autores tenemos dilatada experiencia [5].

Hemos organizado el resto del artículo siguiendo la estructura propuesta por PPDIOO y finalizamos con las conclusiones extraídas después de la realización de la impartición de esta materia en el curso 2018-2019.

II. Preparación

Durante la fase de preparación establecimos los requisitos organizativos y funcionales. Además, propusimos la arquitectura de alto nivel del entorno así como las tecnologías que darían soporte a dicha arquitectura y, finalmente, buscamos la financiación para abordar este proyecto.

A. Especificación de requisitos

En primer lugar, establecimos los requisitos que debía cumplir el laboratorio, teniendo en cuenta la orientación marcadamente práctica del máster y utilizando como referencia las guías docentes de las materias [2], [3].

Los requerimientos se pueden clasificar en organizativos (no funcionales), que están relacionados con el entorno de implantación, y funcionales, directamente relacionados con las tareas que los estudiantes deben realizar en el entorno de entrenamiento que conforma este laboratorio de prácticas.

Comenzamos abordando las condiciones del ámbito de trabajo. El primer requisito que establecimos fue el de usar la misma infraestructura para las dos materias directamente relacionadas con Ingeniería Telemática, "Redes Seguras" y "Seguridad en Comunicaciones". En segundo lugar, decidimos que los laboratorios deberían ser exactamente iguales en los centros de A Coruña y Vigo, de modo que las prácticas fuesen exactamente reproducibles en ambos escenarios. En tercer lugar, definimos que el laboratorio tenía que ser homogéneo (el uso de diferentes plataformas tecnológicas incrementa la curva de aprendizaje del alumno), modular, escalable y adaptable al número de estudiantes, cuyo número máximo es de 20 por centro.

Posteriormente, revisamos los resultados docentes esperados, definidos en las guías de las materias, para establecer cuáles serían las funcionalidades que deberían proporcionar dichos equipos para cada una de las dos materias.

"Redes Seguras" tiene como objetivo principal el diseño e implementación de redes que proporcionen los servicios de seguridad necesarios en una red empresarial.

Para ello, los alumnos deben ser capaces de diseñar infraestructuras de red seguras, con diferentes topologías y aproximaciones, dependientes del entorno y de los requisitos de seguridad definidos. Al mismo tiempo tendrán que ser capaces que convertir una política de seguridad en configuraciones específicas

Por otro lado, los estudiantes deben comprender el papel que desempeñan los firewalls en una arquitectura de red de seguridad perimetral y como los diferentes tipos de filtrado (basado en paquetes sin estado, basado en paquetes con estado, en capa de aplicación, basado en zonas, entre otros) pueden ser utilizados para controlar el acceso entre las diferentes redes de la organización.

Como complemento al diseño e implementación de redes basadas en seguridad perimetral, es necesario también que los estudiantes se familiaricen con los sistemas de detección y prevención de intrusiones, por lo que deberán ser capaces de implementar este tipo de soluciones en el laboratorio de prácticas.

En la materia de "Seguridad en Comunicaciones" abordamos los protocolos que permiten fortificar la comunicación tanto redes locales como entre puntos remotos. Es por esto que comenzamos estudiando cómo proteger el acceso tanto a redes cableadas, Ethernet, como a redes inalámbricas basadas en la familia de estándares IEEE 802.11, utilizando estándares comúnmente aceptados como IEEE 802.1x o . Analizamos también cómo compartimentar redes utilizando técnicas basadas en VLAN.

Estudiamos también las características de los protocolos relacionados con los planos de control (e.g. enrutamiento) y gestión de red (e.g. SNMP), estudiando sus vulnerabilidades y los mecanismos de fortificación disponibles.

Por último, nos centramos en los aspectos relacionados con la protección del envío de información a través de infraestructuras de red no seguras, mediante *Virtual Private Networks* (VPN), usando diferentes familias de protocolos como IPsec o TLS.

B. Arquitectura de red y tecnologías

Los profesores de las materias hemos decidido utilizar como arquitectura de referencia el modelo de red empresarial propuesto por Cisco [4], puesto que representa adecuadamente una red de datos corporativa. Además este modelo se puede simplificar para adaptarse a entornos de tamaño más reducido. Otra ventaja que presenta es que hace referencia a todas las tecnologías de seguridad que vamos a estudiar en ambas asignaturas y proporciona guías para su integración.

En cuanto a las tecnologías que vamos a utilizar, son:

- Ethernet: conmutación básica, *Spanning-Tree*, VLAN (IEEE 802.1Q), control de acceso basado en identidad (IEEE 802.1x) y técnicas de mitigación de vulnerabilidades para *arp* o DHCP.
- Redes inalámbricas basada en Wi-Fi: implementación de diferentes *frameworks* de seguridad: WPA2, WPA3, soporte para control de acceso basado en identidad (IEEE 802.1x)
- Enrutamiento: tanto en IPv4 como en IPv6. Soporte para protocolos de enrutamiento dinámico como RIP, OSPF o BGP.
- Firewalls: filtrado estático de paquetes, filtrado dinámico de paquetes, filtrado basado en contenidos de capa de aplicación y filtrado basado en zonas.
- VPN basadas en IPsec: despliegue de topologías punto a punto y de acceso remoto (control de acceso basado en IEEE 802.1x).
- VPN basadas en MPLS: como alternativa a IPsec proponemos un escenario alternativo
- Sistemas de detección y prevención de intrusiones (IDS/IPS).
- Monitorización mediante SNMPv3 y *syslog*
- Gestión remota de los dispositivos, basada protocolos seguros

- Implementación del modelo de seguridad de autenticación, autorización y auditoría (AAA).

C. Viabilidad económica

En este punto de la fase de preparación analizamos diferentes posibilidades que existían en el mercado para diseñar una arquitectura que permitiese incorporar todas estas tecnologías, de modo que los costes fuesen asumibles tanto por la Universidade da Coruña como por laa Universidade de Vigo.

Comenzamos estudiando las posibilidades que existían para trabajar con versiones gratuitas de entornos de virtualización de redes como GNS3 (*Graphical Network Simulator*) [6] o Eve-NG (*Emulated Virtual Environment*) [7] pero, en ambos casos, la necesidad de trabajar con imágenes de sistemas operativos bajo licencia generaba la necesidad de adquirir dichas licencias, lo que provocaba un problema puesto que, habitualmente, las licencias se comercializan asociadas al número de serie de un dispositivo concreto y su coste es alto. Además, si bien estos entornos de simulación operan bien a nivel de enrutamiento y seguridad, presentan bastantes limitaciones a la hora de trabajar con switches o switches multicapa.

Consideramos también la posibilidad de utilizar *Cisco Packet Tracer*. Si bien su licencia de uso solamente requiere el registro en la plataforma del programa Cisco Netacad [8], el análisis de tráfico y la limitación de funcionalidades en dispositivos de red y finales no lo hacen adecuado para su uso en la docencia de estas materias. Otra opción de simulación analizada fue Cisco VIRL (*Virtual Internet Routing Lab*) [9]. En este caso, pese a que soluciona los problemas derivados de la adquisición de licencias de los sistemas operativos de los dispositivos de red, el propio software que, actualmente, solamente una versión denominada *Personal Edition*, presenta un coste por licencia de 199 $ por alumno, lo que en un máster cuyos 90 créditos en total tienen coste (suponiendo que todas las materias se superan el primer año) de 2.821,68 €, no es viable económicamente. Además, el pago de estas licencias cada año, por alumno, es muy difícil de gestionar y justificar, lo que de nuevo lleva a la conclusión de ser una opción no viable.

Finalmente, debido a los problemas mencionados, optamos por la compra de dispositivos físicos para el laboratorio. Para reducir los costes que esta adquisición presenta, optamos por adquirir dispositivos reacondicionados con un distribuidor que proporciona garantía de por vida en los equipos y las actualizaciones de los sistemas operativos. También optamos por elegir dispositivos que pueden desempeñar las funciones de *router*, *firewall*, concentrador de VPN e IDS/IPS.

Como resultado de esta fase obtuvimos una arquitectura de alto nivel, diseñada para hasta veinte estudiantes organizados en grupos de prácticas de cuatro alumnos, donde a cada grupo se asignan *switches* de capa 2, *switches* multicapa (conmutación en capa 3) y *routers* con servicios de *firewall,* VPN e IPS/IDS, así como un router inalámbrico. Todos estos dispositivos, en conjunto, deben implementar las tecnologías especificadas en el apartado *B. Arquitecturas de red y tecnologías*.

III. PLANIFICACIÓN

Durante la fase de planificación identificamos los requisitos de los dispositivos en base a la información obtenida en la fase anterior.

En primer lugar definimos los tipos de dispositivos que necesitamos y a continuación establecimos las funcionalidades que debían proporcionar, lo que constituyó el pliego de condiciones a enviar a los proveedores:

- *Switches* Ethernet: conmutación básica, *Spanning-Tree* y mecanismos de protección del mismo, VLAN, (IEEE 802.1Q), control de acceso basado en identidad (IEEE 802.1x) y técnicas de mitigación de vulnerabilidades para arp o DHCP.
- *Switches* Ethernet Multicapa: además, de las características proporcionadas por los "*Switches* Ethernet", deben proporciona conmutación en capa 3 (enrutamiento por hardware), tanto en IPv4 como en IPv6, así como protocolos de enrutamiento dinámico (RIP y OSPF) y filtrado estático de paquetes.
- *Routers*: enrutamiento IPv4 e IPv6, soporte para protocolos de enrutamiento dinámico como RIP, OSPF o BGP; filtrado estático y dinámico de paquetes, filtrado de capa de aplicación y filtrado basado en zonas; VPNs IPsec tanto para configuraciones *site-to-site* como de acceso remoto.
- El *router* Wi-Fi se excluye del proceso de compras puesto que ambos centros disponen ya de dispositivos de estas características disponibles.

Es necesario aclarar que algunas características como el número de puertos, la velocidad de los mismos o el *through-put* no son especialmente relevantes puesto que el objetivo de estos laboratorios es probar y configurar funcionalidades, tecnologías y protocolos, no dar servicio a una organización y, por lo tanto, no se tienen en cuenta.

Enviamos este pliego de condiciones a tres proveedores que realizaron ofertas con una oferta que incluía 10 *switches* Ethernet, 10 *switches* multicapa y 15 *routers* que cumplían las características solicitadas. De los tres presupuestos seleccionamos el que combinaba el mejor precio y además se adecuaba a las formas de pago soportadas por la Universidade da Coruña y la Universidade de Vigo.

Al mismo tiempo que analizábamos presupuestos y confirmábamos el pedido, en la Facultad de Informática de la Universidade da Coruña fue necesario acondicionar el aula para ubicar el equipamiento de forma segura (en racks de 42 U). Además, solicitamos al Servicio de Obras la instalación de cableado estructurado de modo que los estudiantes pueden conectarse a través de consola o SSH a los equipos en los armarios desde su puesto de trabajo.

IV. DISEÑO

Desarrollamos el diseño en base a los requisitos tanto técnicos como organizativos obtenidos en las fases anteriores. Hemos hecho referencia tanto a la selección específica de dispositivos, topología física, diseño lógico y especificación de tecnologías y protocolos a utilizar funcionamiento.

En cuanto al listado específico de equipos, es el siguiente:

- *Switches* Ethernet: WS-C2960G-24TC-L (Cisco Catalyst 2960 24 10/100/1000 4 T/SFP LAN Enterprise Image): 10 unidades.
- *Switches* Multicapa: WS-C3560-24 PS-E (Cisco Catalyst 3560 24 10/100 PoE + 2 SFP): 10 unidades.
- *Routers* multiservicio: Cisco 1941-SEC/K9 (Cisco 1941-SEC/K9 I*ntegrated Services Security Router*): 15 unidades.

En la topología física, se establece cómo se interconectan físicamente los dispositivos, interfaz a interfaz. En este documento se omiten estos detalles puesto que no aportan ningún valor especial. Sin embargo, es necesario resaltar que es una información clave a entregar a los operadores del Servicio de Gestión de Red que en la fase de implementación colocarán los dispositivos de red en el armario y los conectarán.

En cuanto a la topología lógica, para cada grupo de prácticas de 4 alumnos, es la que se puede observar en la Fig. 1. Debido a que, como máximo, podría haber 5 grupos de prácticas por centro, en cada uno habría un máximo 5 topologías como la que se puede ver a continuación.

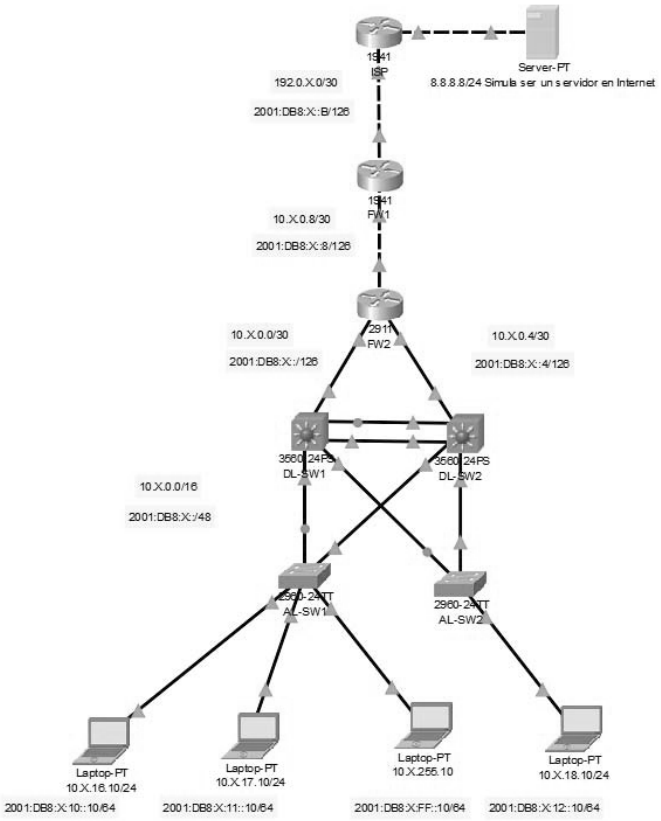


Fig. 1. Topología lógica del entorno de trabajo de un grupo de prácticas

Como se puede observar se propone una topología de red estructurada en tres capas: capa de acceso, basada en switches de capa 2 Catalyst 2960; capa de distribución/núcleo (se agregan las funcionalidades de capa de distribución y núcleo en una sola), basada en switches multicapa Catalyst 3560; y una frontera corporativa formada por dos *firewalls*, implementados mediante *routers* Cisco 1941. El *router* denominado ISP simula ser un router de un proveedor de servicios.

En cuanto a las tecnologías y configuraciones de base que proporcionamos a los alumnos son: la división de puertos de usuario en la capa de acceso en 4 VLANs, la utilización de enlaces troncales basados en IEEE 802.1Q, la operativa automática de *Spanning-Tree Protocol*, con un implementación de tipo *Per VLAN STP* + y enrutamiento dinámico basado en RIP o en OSPF.

Con respecto al direccionamiento IP, hemos previsto que cada grupo de prácticas trabaje con un rango de direcciones IP diferente, con el propósito de interconectar redes de grupos distintos en las prácticas de VPN. La estrategia seguida es la siguiente: cada grupo de prácticas se identifica con número X y el espacio de direccionamiento asignado que tienen es el siguiente:

- Direccionamiento IPv4 privado: 10.X.0.0/16
- Direccionamiento IPv4 público: 192.0.X.0/24
- Direccionamiento IPv6: 2001:DB8:X::/48

V. Implementación

En esta fase instalamos el equipamiento y realizamos la configuración básica para poner en marcha el entorno de prácticas, siguiendo las especificaciones de la fase de diseño.

En concreto la instalación física de los dispositivos de red fue realizada por el Servicio de Gestión de Red y las configuraciones básicas de los dispositivos por los profesores de la materia.

Describimos las dos tareas de forma exacta, proporcionando una guía de instalación y la duración en tiempo estimada al Servicio de Gestión de Red.

En el proceso de implementación de una infraestructura de red en producción se proporcionan guías de *rollback* para devolver la red a su estado de operación anterior si se produce algún fallo durante la fase de implementación. En este caso, al ser una infraestructura nueva orientada a laboratorio de prácticas, estos procedimientos de vuelta atrás no son estrictamente necesarios puesto que un error en el despliegue de la configuración únicamente necesita llevar a cabo las operaciones de eliminación de dicha configuración, que además en estas fases iniciales se consigue con un simple reinicio.

VI. Operación

En una red en producción, en la fase de operación llevamos a cabo las tareas de que mantienen las características operacionales de la red, así como sus niveles de rendimiento. Para ello necesitamos monitorizar y administrar los dispositivos clave de la red, lo que permite detectar desviaciones y corregir fallos de forma proactiva. Se llevan a cabo además tareas de actualización de sistemas operativos y aplicación de medidas de seguridad adicionales sobre nuevas vulnerabilidades descubiertas. Sin embargo, en este entorno, la fase de operación consiste en la realización de los ejercicios prácticos planteados a los alumnos y cuyos títulos se indican a continuación:

1. Configuración de seguridad mínima en *routers* y switches
2. Procedimiento de recuperación de contraseñas
3. Despliegue del escenario de trabajo de referencia
4. Configuración de mecanismos de control de acceso a los dispositivos

5. Control de acceso a la red de área local basado en autenticación 802.1x. Asignación dinámica de VLANs.
6. Fortificación de la capa de acceso: protección frente a los ataques de *arp spoofing* y *DHCP spoofing*
7. Seguridad perimetral
8. Detección de intrusiones
9. VPNs IPsec sitio a sitio
10. VPNs IPsec de acceso remoto

Con esta enumeración de prácticas el lector puede comprender las posibilidades que aportan este diseño y la realización de prácticas en este entorno.

VII. Optimización

En la fase de optimización, normalmente, se llevan a cabo tareas correctivas, mediante tareas de gestión de red proactiva, con el objeto de que los problemas detectados en fases tempranas no afecten al funcionamiento de la red. Si los problemas son leves, es suficiente con ajustar las configuraciones. Pero si los fallos con graves o afectan a gran cantidad de dispositivos puede ser necesario hacer cambios en el diseño.

Para optimizar la operativa y el uso de estos laboratorios de prácticas es vital la información aportada por los alumnos en las encuestas de calidad que son efectuadas desde los centros. En base a esta información estamos planteando como mejoras para el próximo curso las que se detallan a continuación:

- Necesitad de acceso remoto a los laboratorios para incrementar notablemente el número de horas de uso de los equipos. De este modo los alumnos pueden sacar un mayor partido a las inversiones efectuadas por las universidades.
- Proporcionar guías de gestión de configuraciones a los estudiantes para que optimicen sus tiempos de despliegue (realización de prácticas). Estamos considerando en este punto la posibilidad de comenzar a trabajar con paradigmas DevNetOps [10], [11]

VIII. Conclusiones

En este documento se presenta la experiencia vivida por los profesores que imparten materias directamente relacionadas con Ingeniería Telemática en un Máster de Ciberseguridad de reciente creación. Pretendemos aportar a la comunidad docente de nuestro ámbito, no solamente la topología de red que hemos desarrollado y las características de los equipos que hemos adquirido, sino como el uso de metodología de despliegue de red nos permite obtener una solución efectiva y con costes acotados para que los alumnos puedan adquirir experiencia en la administración de dispositivos de seguridad en red.

Aunque los profesores estamos satisfechos con el trabajo realizado por los alumnos durante este año, somos conscientes de que debemos realizar algunas mejoras para optimizar su rendimiento en cursos posteriores. Además de las mejoras indicadas en la fase de optimización, es necesario aportar documentación básica de administración de dispositivos de red y guías de configuración adaptadas a los equipos de los que disponemos puesto que, aunque la mayor parte de los alumnos de este máster proceden de titulaciones de Ingeniería Informática o de Telecomunicación, sus conocimientos sobre administración de redes son muy diferentes. Hemos podido observar algunas situaciones en las que la falta de experiencia ha provocado pérdidas de tiempo o de trabajo. Creemos que esta situación que es subsanable con unas guías y recomendaciones en las que deberían trabajar durante las primeras semanas de curso.

Agradecimientos

Este trabajo ha sido financiado, en parte, por el Ministerio de Economía y Competitividad de España (Proyecto TIN2015-70648-P), por la Xunta de Galicia (Centro singular de investigación de Galicia “acreditación ED431G/01” 2016-2019) y la Unión Europea (European Regional Development Fund - ERDF).

Referencias

[1] Universidade da Coruña y Universidade de Vigo, «MUniCS,» 20 05 2019. [En línea]. Available: https://www.munics.es.

[2] Facultad de Informática. Universidade da Coruña, «Guía docente: Máster Universitario en Ciberseguridad. Redes Seguras,» 31 05 2019. [En línea]. Available: https://guiadocente.udc.es/guia_docent/index.php?centre=614&ensenyament=614530&assignatura=614530006&any_academic=2018_19&idioma_assig=cast&idioma=cast.

[3] Escuela de Ingenieria de Telecomunicación. Universidade de Vigo, «Guía Docente: Máster Universitario en Ciberseguridad. Seguridad en Comunicaaciones.,» [En línea]. Available: https://secretaria.uvigo.gal/docnet-nuevo/guia_docent/?centre=305&ensenyament=V05M175V01&assignatura=V05M175V01103&idioma_assig=cast. [Último acceso: 31 05 2019].

[4] A. Bruno y S. Jordan, «Network Design Methodology,» de *CCDA 200-310 Official Cert Guide, Fifth Edition*, Cisco Press, 2016.

[5] F. J. Novoa, J. Pereira, J. M. Vázquez, A. F. Castro y J. Teijeiro, «Network analysis and design: a customized methodology for a DICOM PACS,» de *MEDINFO*, San Francisco, USA, 2004.

[6] Graphic Network Simulator, «GNS3,» [En línea]. Available: https://gns3.com/. [Último acceso: 31 0 2019].

[7] EVE-NG, «Emulated Virtual Network for Network Security and DevOps Professionals,» [En línea]. Available: https://www.eve-ng.net/. [Último acceso: 31 05 2019].

[8] Cisco Systems, «Cisco Netacad,» [En línea]. Available: https://www.netacad.com. [Último acceso: 31 05 2019].

[9] Cisco Systems, «Cisco Virtual Internet Routing Lab,» [En línea]. Available: https://learningnetworkstore.cisco.com/cisco-virtual-internet-routing-lab. [Último acceso: 31 05 2019].

[10] Juniper Networks, «What is DevNetOps?,» [En línea]. Available: https://www.juniper.net/us/en/products-services/what-is/devnetops/. [Último acceso: 31 05 2019].

[11] H. Preston, «Embrace NetDevOps, Say Goodbye to a "Culture of Fear",» Cisco Systems, 09 10 2017. [En línea]. Available: https://blogs.cisco.com/developer/embrace-netdevops-part-1. [Último acceso: 31 05 2019].

[12] Universidade da Coruña, «Universidade da Coruña: Máster Universitario en Ciberseguridad,» 20 05 2019. [En línea]. Available: https://guiadocente.udc.es/guia_docent/index.php?centre=614&ensenyament=614530&assignatura=614530006&any_academic=2018_19&idioma=cat&idioma_assig=cast.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Modelado basado en Ontologías para Redes de Transporte en Carreteras

Susel Fernandez, Luis Cruz-Piris, Ivan Marsa-Maestre
Departamento de Automática
Universidad de Alcalá
Escuela Politécnica Superior. Campus Universitario, Ctra. Madrid-Barcelona km. 33, 600. 28805. Alcalá de Henares. Madrid
susel.fernandez@uah.es, luis.cruz@uah.es, ivan.marsa@uah.es

***Resumen*- Los sistemas inteligentes de transporte son un conjunto de soluciones tecnológicas que se utilizan para mejorar el rendimiento y la seguridad del transporte por carretera. Un elemento crucial para el éxito de estos sistemas es que los vehículos puedan intercambiar información no solo entre ellos, sino también con otros elementos de la infraestructura vial a través de diferentes aplicaciones. Para el éxito de este intercambio de información, se necesita un marco común de conocimiento que permita la interoperabilidad. En este trabajo se propone un sistema basado en ontologías para proporcionar asistencia en la carretera, que facilite a los conductores la toma de decisiones en diferentes situaciones, teniendo en cuenta la información sobre diferentes elementos relacionados con el tráfico, como pueden ser las rutas, señales y reglas de tráfico y elementos meteorológicos.**

***Palabras Clave*- sistemas inteligentes de transporte, ontologías, redes de tráfico.**

I. Introducción y antecedentes

La continua evolución de los sistemas de transporte inteligentes ha dado paso a una nueva era de sistemas inteligentes interconectados, que ha significado un salto cuantitativo en la seguridad del transporte por carretera. Estos sistemas permiten el intercambio de información entre diferentes aplicaciones, y el análisis posterior de esta información para contribuir a mejorar la seguridad y la comodidad de los conductores en los viajes por carretera.

Debido a su alto grado de expresividad, el uso de ontologías es crucial para garantizar una mayor interoperabilidad entre los agentes de software y las diferentes aplicaciones involucradas en los sistemas de transporte inteligentes. Las ontologías proporcionan un vocabulario común en un dominio determinado y permiten definir, con diferentes niveles de formalidad, el significado de los términos y las relaciones entre ellos [1]. Las ontologías facilitan el diseño de esquemas conceptuales exhaustivos y rigurosos para permitir la comunicación y el intercambio de información entre diferentes sistemas e instituciones.

Hay algunos trabajos previos enfocados en ontologías para sistemas de transporte por carretera. En [2] se presenta una ontología para representar el tráfico en carreteras. Su objetivo fue la construcción de un sistema de información de tráfico fiable que brindara información sobre las carreteras, el tráfico y los escenarios relacionados con los vehículos en las carreteras. También proporciona formas para analizar qué tan crítica es una situación específica. Por ejemplo, una ambulancia puede necesitar conocer el estado de congestión de una zona de peaje. Solicitar esta información es crítico si la ambulancia se está dirigiendo a la escena de un accidente. En cambio, en el caso de un vehículo común que circule sin prisas por una carretera, esta información no sería crítica.

En [3] se propone una representación de alto nivel para los vehículos autónomos y su entorno. El sistema sirve de ayuda a los conductores para tomar decisiones "ilegales" pero prácticas en determinadas circunstancias (por ejemplo, cuando un automóvil dañado no permite la circulación, tomar la decisión de moverse a otro carril cruzando una línea continua para adelantar al vehículo detenido, siempre que el otro carril esté despejado). Esta representación incluye conocimiento topológico y reglas de inferencia para calcular el siguiente movimiento que un vehículo autónomo debería tomar, como asistencia al conductor.

El trabajo propuesto en [4] es un enfoque para crear una descripción genérica de la situación para sistemas avanzados de asistencia al conductor utilizando un razonamiento lógico sobre una base de conocimiento de la situación del tráfico. Contiene múltiples objetos de diferentes tipos, como vehículos y elementos de infraestructura como carreteras, carriles, intersecciones, señales de tráfico, semáforos y relaciones entre ellos. El proceso de inferencia lógica se realiza para verificar e interpretar la situación razonando sobre las reglas de tráfico.

En el trabajo en [5] se propone una ontología para la gestión del tráfico, que agrega ciertos conceptos de tráfico a la ontología general de sensores A3ME [6]. Los conceptos agregados son especializaciones de posición, distancia y clases de sensores de aceleración, y las diferentes acciones que tienen lugar en los movimientos del vehículo.

En [7] se introdujo una base de conocimientos basada en ontologías, con mapas y reglas de tráfico. Se pueden detectar las situaciones de exceso de velocidad y tomar decisiones en las intersecciones para cumplir con las reglas de tráfico. En este trabajo pero no se consideran elementos importantes como las señales de tráfico y las condiciones climatológicas.

La mayoría de los trabajos encontrados en la literatura se centran en describir situaciones de tráfico muy específicas, tales como encontrar estacionamiento, acciones de vehículos de emergencia y situaciones de intersección [8][9], comportamiento del conductor [10]. Pero ninguno de ellos es lo suficientemente general y expresivo como para abarcar una amplia variedad de situaciones de tráfico. Por lo tanto, es necesario desarrollar ontologías en el dominio del tráfico vial lo suficientemente expresivas como para describir cualquier situación de tráfico.

Este trabajo presenta un sistema basado en ontologías para la gestión del transporte por carretera, con el objetivo de proporcionar asistencia al conductor en diferentes situaciones de tráfico. La ontología desarrollada gestiona el conocimiento relacionados con los vehículos y los elementos del entorno que pueden influir en el tráfico vial, como por ejemplo los elementos de la infraestructura, las condiciones climáticas y las reglas de tráfico.

El documento está organizado de la siguiente manera. La sección II presenta la arquitectura del sistema. En la sección III se explican casos de estudio con diferentes escenarios de tráfico. Finalmente, las conclusiones y líneas de trabajo futuro se resumen en la sección IV.

II. ARQUITECTURA DEL SISTEMA

En la Figura 1. se muestra el sistema propuesto para los servicios de asistencia al conductor. En la base de la arquitectura está la ontología [11], desarrollada para el dominio específico del tráfico en carretaras. Para desarrollar el proceso de razonamiento se definen los mecanismos de inferencia lógica, utilizando el razonador *Pellet*. En el nivel superior se encuentran las distintas aplicaciones que acceden a la Información de la ontología a través de consultas SPARQL.

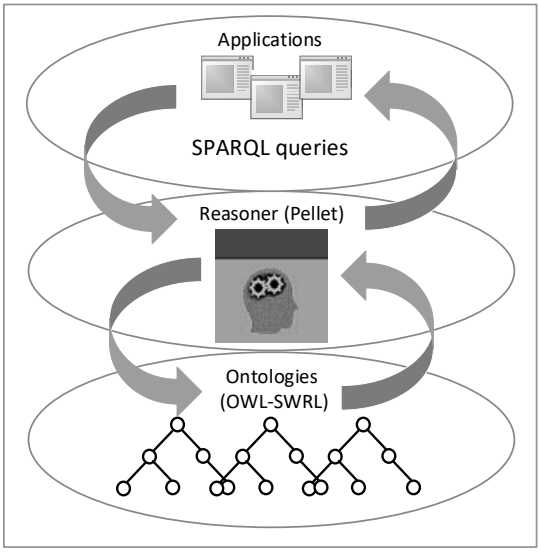


Fig. 1. Arquitectura del sistema.

A. *Descripción de la ontología*

La ontología desarrollada en el sistema permite modelar y relacionar las diferentes entidades de tráfico vial identificadas. La implementación se desarrolló en el lenguaje OWL-RDF [12] utilizando la herramienta *Protégé* [13].

Para una mejor comprensión, presentamos modelo del conocimiento de la ontología dividido en dos grupos de conceptos interrelacionados. El primer grupo contiene los elementos referentes a los vehículos, y el segundo grupo los elementos referentes a la infraestructura vial. El grupo principal está relacionado con los vehículos. Los conceptos de este grupo se muestran en la Figura 2.

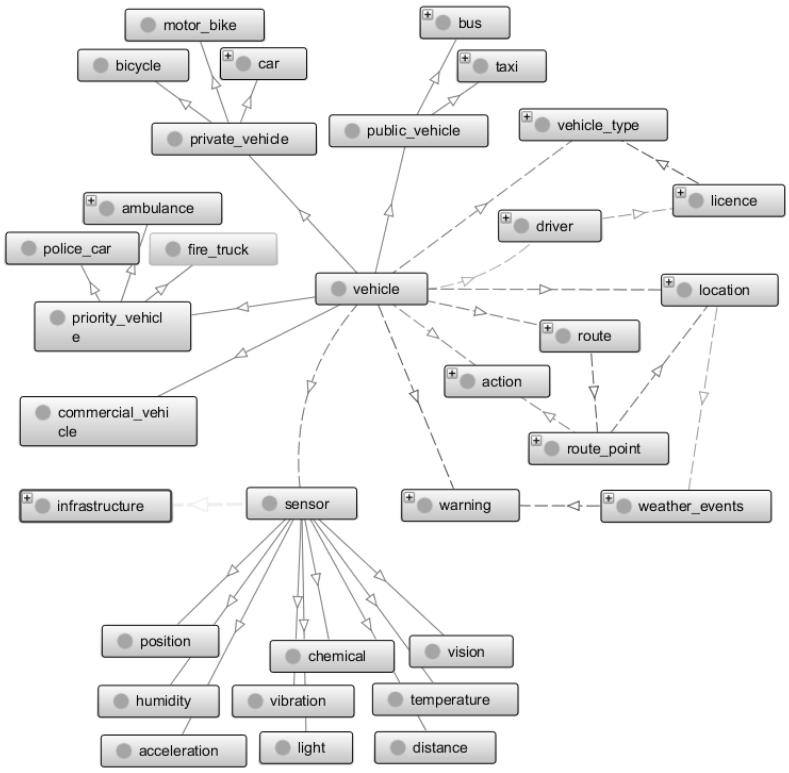


Fig. 2. Conceptos relacionados con vehículos.

La figura muestra la taxonomía de los vehículos, que se pueden clasificar en: vehículos comerciales, vehículos públicos (autobuses y taxis), vehículos privados (automóviles, bicicletas y motocicletas) y vehículos prioritarios (ambulancias, camiones de bomberos y coches de policía). Las diferentes relaciones entre los vehículos y otras entidades se definen también en este grupo. Algunas de estas entidades son: ubicación, que muestra la ubicación exacta (latitud y longitud) de un vehículo, punto de ruta o elemento de infraestructura; información sobre los conductores y los tipos de vehículos que pueden conducir según su tipo de permiso de conducción.

Una de las características fundamentales de este grupo es que cada vehículo tiene asociado un conjunto de acciones a realizar, que pueden variar según la ruta y las señales de tráfico encontradas, así como un conjunto de advertencias según la situación meteorológica en el área.

Con respecto a los sensores, estos pueden ubicarse no solo en los vehículos sino también en diferentes partes de la infraestructura, tales como puentes, carreteras, señales, etc. En la ontología se han definido varios tipos de sensores como: vibración, aceleración, humedad, temperatura, etc. La Figura 3 muestra el segundo grupo, que organiza los elementos relacionados con la infraestructura vial. En este grupo, el concepto más importante representa las carreteras.

Para una mejor gestión de las situaciones de tráfico, dividimos las carreteras en segmentos, conectados a través de intersecciones. Cada segmento contiene carriles, y en cada carril hay señales diferentes, como señales de alto o control de velocidad, semáforos o señales viales. Cada señal tiene una acción asociada a las normas de tráfico correspondientes.

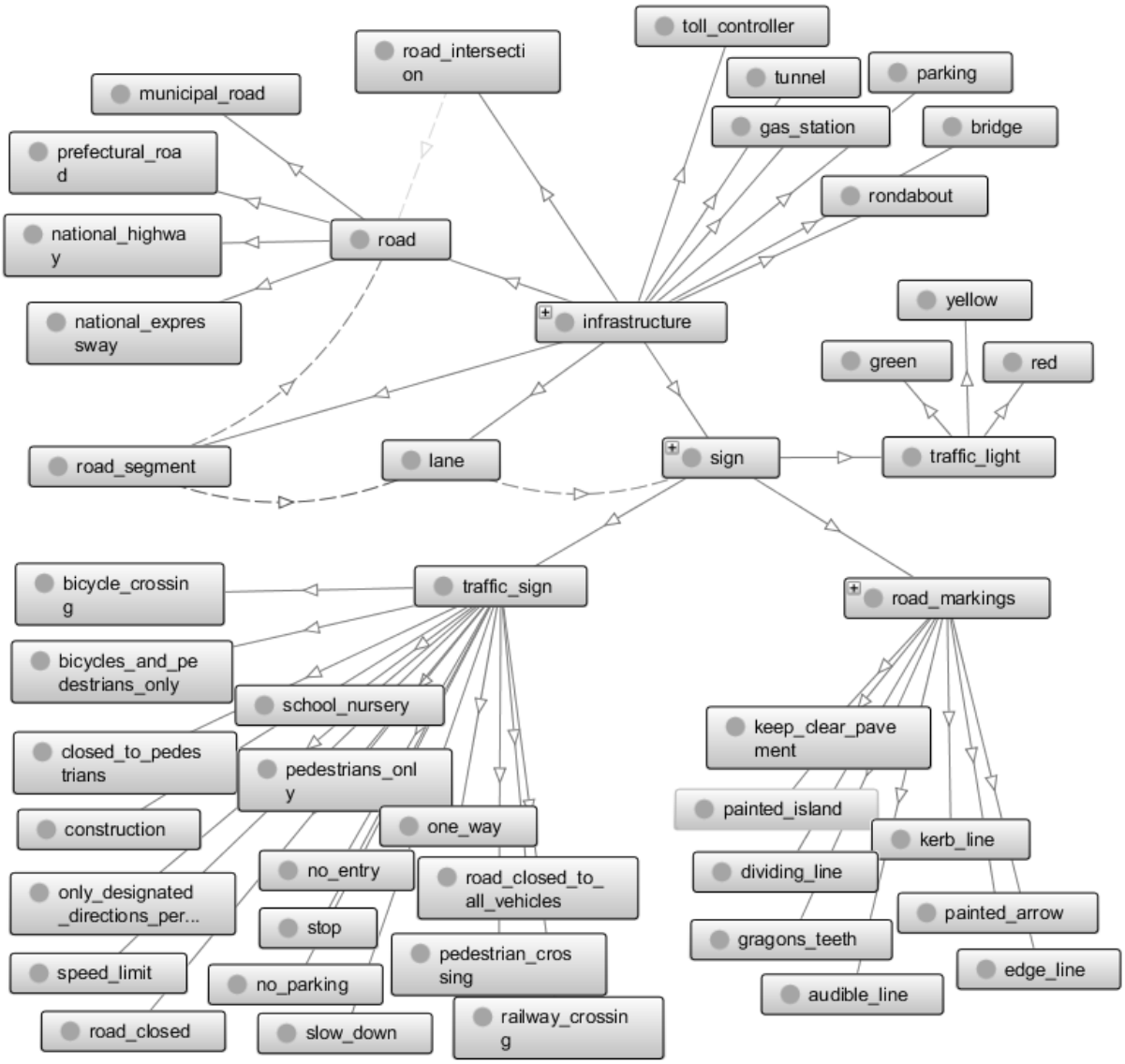


Fig. 3. Conceptos relacionados con la infraestructura de la vía.

B. Mecanismo de Razonamiento

Un aspecto crucial cuando se trabaja con ontologías es el mecanismo de razonamiento, que en la Inteligencia Artificial es simplemente la capacidad de obtener nuevo conocimiento a partir del conocimiento ya disponible mediante estrategias de inferencia. Para razonar con ontologías, se utilizan principalmente tres técnicas: razonamiento con lógica de primer orden, razonamiento con lógica de descripción y razonamiento con reglas.

En este trabajo utilizamos el razonador *Pellet* [14], que es una herramienta para razonar con ontologías, que admite los tres tipos de razonamiento. *Pellet* se implementa en *Java*; está disponible de forma gratuita y permite verificar la consistencia de la ontología.

Las reglas de razonamiento en la ontología del tráfico se han desarrollado utilizando el Lenguaje de Reglas de la Web Semántica (*SWRL*) [15]. En esta ontología, las reglas *SWRL* se utilizan para definir diferentes regulaciones de tráfico y las diferentes acciones que un conductor puede tomar, de acuerdo con la situación actual de la carretera. Entre las reglas definidas en la ontología se encuentran, por ejemplo, aquellas que permiten al razonador inferir la transitividad con respecto a la ubicación de los elementos de tráfico. Esto significa que si un elemento de tráfico (e.g un vehículo o señal de tráfico) está ubicado en un carril, y ese carril está ubicado en un segmento de la carretera, entonces el elemento de tráfico también se encuentra en esa carretera. Otros conjuntos de reglas definidas están dirigidas a determinar la acción que el conductor de un vehículo debería ejecutar dependiendo de determinadas circunstancias, por ejemplo, cuando se circula en el mismo carril que un vehículo prioritario en situación de emergencia o cuando nos encontramos en el mismo segmento que una señal de tráfico concreta.

C. Consultas a la ontología

En el nivel superior del sistema, las diferentes aplicaciones consultan la información almacenada en la ontología para llevar a cabo su ejecución. Como lenguaje de consulta ontológica se ha utilizado *SPARQL* [16].

Un ejemplo simple de consulta *SPARQL* con la ontología sería obtener la lista de vehículos que se encuentran en una ruta determinada.

Otro ejemplo de consulta devolvería todos los puntos asociados con una ruta de un vehículo y la acción que debe realizarse para ir de un punto a otro, considerando su ubicación en el mapa. Esta consulta en concreto resulta muy simple, teniendo en cuenta que por dieño, en la ontología cada punto de ruta está relacionado con el siguiente a través de una acción específica (girar a la izquierda, girar a la derecha o seguir recto), y cada acción depende del tipo de relación (*isAtNorthOf*, *isAtSouthOf*, *isAtWestOf*, *isAtEastOf*) que conecta los segmentos en los que se encuentran los puntos de la ruta. Se ha definido una regla *SWRL* en la ontología que asocia una acción o movimiento para trasladarse de un punto al siguiente punto de la ruta según la relación entre los segmentos en los que se encuentra cada punto de ruta.

III. EXPERIMENTOS

Hemos realizado pruebas de la expresividad de la ontología, realizando consultas para situaciones de tráfico simuladas. En esta sección presentamos parte de los experimentos realizados en diversas situaciones de tráfico simples. El escenario de tráfico definido para los experimentos consta de varias carreteras y sus intersecciones. Cada carretera se divide en varios segmentos, con dos carriles cada uno.

De cada vehículo se conoce su ubicación, velocidad y la ruta que desea seguir. Cada ruta tiene un conjunto de puntos y cada punto de ruta tiene una ubicación (latitud y longitud), así como la información sobre el siguiente punto de la ruta. Se definen previamente una serie de situaciones climatológicas específicas en diferentes puntos del mapa. Consultando con la ontología podemos saber la próxima acción que debe tomar el conductor, dada la posición del vehículo, la ruta elegida y las señales de tráfico ubicadas a lo largo de la ruta. También podemos recibir recomendaciones con respecto a la situación del clima a lo largo de la ruta.

En este trabajo, hemos llamado *Desired_Action* a la acción que el conductor desea llevar a cabo para moverse de un punto al siguiente a lo largo de la ruta, independientemente de las señales de tráfico; *Next_Action* es la acción que el conductor realmente debería tomar en cada punto considerando únicamente las señales de tráfico correspondientes. Hemos definido una serie de advertencias para diferentes situaciones climatológicas que se pueden encontrar en la ruta, como lluvia, nieve, niebla, viento, etc. Cada una de estas advertencias está asociada con una serie de recomendaciones para facilitar la circulación en estas condiciones.

A partir de la posición del vehículo en cada punto, la ubicación de las señales de tráfico y la ruta, las acciones se deducen mediante el razonamiento aplicando diferentes reglas de la ontología, en cada uno de los pasos que se describen a continuación:

1. Localizar en qué segmento de la carretera está ubicado el vehículo y cuál es el siguiente punto de la ruta. Esto se hace considerando la posición (latitud y longitud) y las coordenadas de los puntos de inicio y final de cada segmento.
2. Elegir la acción deseada para ir de un punto a otro dependiendo del tipo de conexión entre los segmentos en los que se encuentran los puntos. Por ejemplo, si el vehículo está en segmento1 y el siguiente punto de la ruta está en segmento2; el segmento2 está ubicado al este del segmento1, entonces la acción que debe tomar el vehículo para ir del punto1 al punto2 es girar a la derecha.
3. Elegir la siguiente acción a ejecutar por el vehículo, teniendo en cuenta únicamente las señales de tráfico. Esta es la misma acción asociada con la siguiente señal de tráfico ubicada en el segmento donde se encuentra el vehículo. Por ejemplo, si el vehículo está en un segmento con una señal de *Stop*, la acción que debe tomar el conductor es detenerse.
4. Si hay alguna condición climática especial en el siguiente segmento de la ruta, entonces se le asigna al vehículo la advertencia correspondiente a esa condición climática.

Los experimentos se realizaron en simulación con 50 vehículos y 20 rutas. Para cada ruta se definieron distintos escenarios de tráfico específicos variando diversos factores como el nivel de congestión, el estado de los semáforos y las condiciones climatológicas en distintos puntos.

Para cada vehículo se definieron a priori las distintas acciones a tomar a lo largo de los diferentes puntos de la ruta y luego se compararon estos resultados con los obtenidos por el sistema para evaluar la expresividad de la ontología. Las medidas utilizadas para evaluar la expresividad de la ontología fueron la *Precisión* y el *Recall*, que representan el nivel de exactitud y completitud de los resutados respectivamente.

Dado un conjunto de referencia *R* y un conjunto resultante *A*, la *Precisión* es un indicador de la exactitud y se define como la razón entre el número de instancias correctas y aquellas que el algoritmo considera que pertenecen al conjunto de instancias correctas (ecuación 1).

$$Precision\ (A, R) = \frac{|R\ \cap A|}{|A|} \quad (1)$$

El *Recall* describe la completitud y se define como la razón entre el número de instancias correctas y todas las instancias que realmente pertenecen an conjunto de instancias correctas (ecuación 2).

$$Recall\ (A, R) = \frac{|R\ \cap A|}{|R|} \quad (2)$$

La Tabla 1 presenta los resultados de los experimentos realizados sobre la expresividad de la ontología en términos de *Precisión* y *Recall*. La tabla muestra que para un total de 50 vehículos y 20 rutas, la media de los valores de *Precisión* obtenidos fue de 0,95 mientras que la media del *Recall* fue de 0,98, lo que demuestra que la ontología es válida para proporcionar la información necesaria para la toma de decisiones en los distintos escenarios de tráfico evaluados.

Tabla 1
RESULTADOS DE LA EXPRESIVIDAD DE LA ONTOLOGÍA EN TERMINOS DE PRECISION Y RECALL

Nº vehículos	***Nº Rutas***	***Media Precisión***	***Media Recall***
50	20	0,95	0,98

En general, los resultados muestran que la ontología es suficientemente expresiva en términos de señales de tráfico, rutas y reglas de tráfico. La ontología permite inferir el conocimiento relacionado con el clima a partir de datos de sensores, sin embargo hay sensores de infraestructura que miden otros datos útiles, como el flujo de multitudes y el flujo de tráfico, que aún no se han tenido en cuenta en la ontología. El procesamiento de los datos de esos sensores mejoraría el trabajo en la optimización de la ruta. Consideramos también que para obtener mejores resultados de cara a la mejora de la conducción se necesita ampliar la ontología incorporando una serie de conceptos y relaciones que permitan tener en cuenta otros factores importantes como el comportamiento de los conductores.

IV. CONCLUSIONES

En este documento se presenta un sistema basado en ontologías para la gestión del transporte por carretera. El objetivo principal de este trabajo es proporcionar asistencia al conductor en diferentes situaciones de tráfico, teniendo en cuenta la ruta, el clima y las reglas de tráfico.

La expresividad de la ontología ha sido probada a través de consultas en diferentes situaciones de tráfico que involucran varias señales y las reglas de tráfico. Los resultados de los escenarios probados han sido satisfactorios, pero aún es necesario enriquecer la ontología para abarcar y relacionar más conocimiento. Como trabajo futuro tenemos la intención de continuar mejorando la expresividad de la ontología, con el procesamiento de datos de más sensores ubicados en la infraestructura, por ejemplo, en puentes, carreteras, ríos, túneles. Esos sensores podrían medir el flujo de multitudes, el flujo de tráfico y muchos otros parámetros que son importantes en la optimización del tráfico. También pretendemos mejorar la expresividad de la ontología, agregando información sobre el comportamiento de los conductores, debido a su importancia en todo el proceso de conducción en carretera. Finalmente, planeamos agregar reglas SWRL que describan múltiples mecanismos de negociación automática entre agentes en diferentes escenarios de tráfico.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto MOON-MOdelado basado en ONtologías para redes complejas. CCG2018-EXP-041, de la Universidad de Alcalá.

REFERENCIAS

[1] Studer, R; Benjamins, R.; Fensel, D. Knowledge Engineering: Principles and Methods. In: Data and Knowledge Engineering, 1998, v.25, n.1-2, pp.161-197

[2] Sérgio Gorender, Ícaro Silva. AN ONTOLOGY FOR A FAULT TOLERANT TRAFFIC INFORMATION SYSTEM. 22nd International Congress of Mechanical Engineering (COBEM 2013). November 3-7, 2013, Ribeirão Preto, SP, Brazil

[3] Evangeline Pollard, Philippe Morignot, Fawzi Nashashibi. An ontology-based model to determine the automation level of an automated vehicle for co-driving. FUSION 2013: 596-603

[4] Michael Hülsen, J. Marius Zöllner, Christian Weiss. Traffic Intersection Situation Description Ontology for Advanced Driver Assistance. In 2011 IEEE Intelligent Vehicles Symposium (IV) Baden-Baden, Germany, June 5-9, 2011

[5] A.J. Bermejo, J. Villadangos, J. J. Astrain, A. Cordoba. Ontology Based Road Traffic Management. Intelligent Distributed Computing VI, G. Fortino et al. eds., SCI 446, pp. 103-108.

[6] Herzog, A.; Jacobi, D.; Buchmann, A. A3ME-An Agent-Based Middleware Approach for Mixed Mode Environments. In Proceeding of Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2008), Valencia, Spain, 29 September-4 October 2008; pp. 191-196.

[7] Zhao, L., Ichise, R., Mita, S., & Sasaki, Y. Ontologies for Advanced Driver Assistance Systems.

[8] Fernandez, S., & Ito, T. (2016, September). Using SSN ontology for automatic traffic light settings on inteligent transportation systems. In 2016 IEEE International Conference on Agents (ICA) (pp. 106-107). IEEE.

[9] Cruz-Piris, L., Rivera, D., Fernandez, S., & Marsa-Maestre, I. (2018). Optimized sensor network and multi-agent decision support for smart traffic light management. Sensors, 18(2), 435.

[10] Fernandez, S., & Ito, T. (2015, October). Driver behavior model based on ontology for intelligent transportation systems. In 2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA) (pp. 227-231). IEEE.

[11] Fernandez, S., Ito, T., & Hadfi, R. (2016). Architecture for intelligent transportation system based in a general traffic ontology. In Computer and Information Science 2015 (pp. 43-55). Springer, Cham.

[12] M. Dean, and G. Schreiber, OWL Web Ontology Language Reference. http: //www.w3.org/TR/2004/REC-owl-ref-20040210/; 2004.

[13] Protégé: http://protege.stanford.edu/

[14] Pellet http://clarkparsia.com/pellet/
[15] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, Submission to W3C, May 2004 http://www.w3.org/Submission/SWRL/
[16] SPARQL http://sparql.org/

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Consideraciones de seguridad en el despliegue de redes IoT

Mario Pérez-Gomariz, Fernando Cerdán-Cartagena, Diego García-Sánchez, Juan Suardíaz-Muro.
Departamento de Tecnologías de la Información y la Comunicación
Universidad Politécnica de Cartagena
30202 Cartagena, España.
mario.perez@edu.upct.es, fernando.cerdan@upct.es, diego.garcia@upct.es, juan.suardiaz@upct.es.

***Resumen*- Las redes IoT (Internet of Things) son una solución cada vez más adoptada para la transferencia de datos en redes de largo alcance. Dentro de estas, LPWAN (Low Power Wide Area Network) es una de las tecnologías que más ha crecido en este ámbito debido a las características que presenta como largo alcance, bajo coste o eficiencia energética. Uno de los factores más importantes y demandados es la seguridad en las redes suponiendo uno de los principales retos para los diseñadores de las mismas.**

En este documento se aborda una visión de la seguridad en redes LoRaWAN teniendo en cuenta, no solo la seguridad del núcleo de la red sino, considerando LoRaWAN como una red integrada con otros sistemas de almacenamiento de la información, monitorización o telecontrol. El documento está dirigido principalmente a desarrolladores y usuarios de redes LoRaWAN. Se analizarán y expondrán buenos hábitos de diseño y despliegue de este tipo de redes.

***Palabras Clave*- jitel, telemática, IoT, LPWAN, LoRa, LoRaWAN, seguridad.**

I. Introducción

LoRaWAN [1] es una tecnología emergente en el desarrollo de redes LPWAN. A pesar de su reciente incorporación al mercado, esta tecnología está siendo cada vez más adoptada por empresas, instituciones y usuarios particulares para el desarrollo de redes IoT.

LoRaWAN es un protocolo orientado al desarrollo de redes de largo alcance y bajo consumo. Está basado en el uso de una modulación patentada LoRa (Long Range) basada en técnicas de espectro ensanchado. De esta forma se consiguen grandes rangos de cobertura a la vez que se dota a las transmisiones de gran robustez protegiéndolas frente a interferencias. LoRa usa bandas de frecuencia ISM de libre uso comprendidas entre 868MHz y 900MHz con bajas tasas de datos. Entre otras características destacan además el bajo consumo de energía, bajo coste o la escalabilidad de las redes [8].

El resto del artículo se organiza como sigue: En la sección II se revisan los mecanismos de seguridad del protocolo LoRaWAN; en la sección III, se describe el despliegue de una red real LoRaWAN en la que se han realizado diferentes ataques para analizar la seguridad de la misma; en la sección IV se desarrolla un conjunto de buenas prácticas de diseño que todo desarrollador debería tener en cuenta para diseñar una red segura y, por último, en el capítulo V se exponen las conclusiones finales.

II. Seguridad en la arquitectura LoRaWAN

Una red LoRaWAN está compuesta principalmente por 4 elementos. Los nodos finales que son los encargados de transmitir información hacia la red. Las puertas de enlace o gateways que reenvían los mensajes procedentes de los nodos hacia el núcleo de red. Un servidor de red encargado de gestionar la red y los dispositivos y un servidor de aplicaciones para la comunicación con otros sistemas o aplicaciones. En la Fig.2 se muestra la topología clásica de una red LoRaWAN.

Las redes LoRaWAN se presentan, a priori, como una solución segura para la implementación de redes LPWAN. Desde un primer momento, el protocolo LoRaWAN ha sido diseñado con la intención de proporcionar seguridad a la red protegiendo la información que circula a través de ella.

En la especificación del protocolo [1], se ha desarrollado una estructura de seguridad a varias capas basada en el uso de mecanismos criptográficos.

A. Claves de cifrado

LoRaWAN incorpora dos capas de seguridad, una a nivel de red y otra a nivel de aplicación. Para ello emplea dos claves simétricas de cifrado AES (Advanced Encryption Standard) de 128 bits.

- *NetworkSessionKey (NwSKey):* Es la clave de cifrado utilizada para garantizar la seguridad a nivel de red.

Con ella se encriptan los mensajes que fluyen entre el nodo y el servidor de la red.

- *ApplicationSessionKey (AppSKey):* Es la clave utilizada para garantizar una comunicación segura extremo a extremo entre el nodo y el servidor de aplicaciones de LoRaWAN. Esta clave es utilizada para cifrar la carga útil del mensaje.

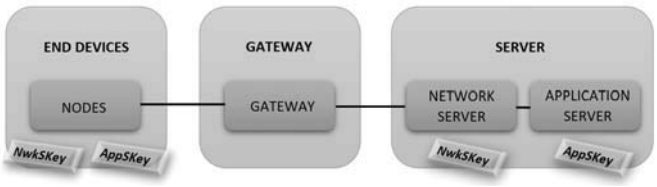


Fig. 1. Topología y manejo de claves de cifrado de una red LoRaWAN.

Como se observa en la Fig.1, los nodos de la red conocen ambas claves de cifrado, sin embargo, la clave NwSKey únicamente es conocida por el servidor de red y la clave AppSKey solo por el servidor de aplicaciones. De esta forma, la información útil del mensaje solo es conocida por el servidor de aplicaciones y no por el de red. Este hecho hace que la red permita configuraciones multi organización ya que la información es protegida extremo a extremo y un administrador de red no podría acceder a los datos de los mensajes.

La forma en la que un nodo se une a la red y obtiene las claves de cifrado se denomina activación de dispositivo.

B. Modos de activación

La activación de un dispositivo consiste en la obtención por parte del nodo de las claves de cifrado (NwSKey y AppSKey) y el identificador de dispositivo (DevAddres) en la red. Existen dos modos diferentes:

- *ABP (Activation By Personalitation):* Es la configuración más sencilla. Todas las claves se almacenan previamente en el dispositivo final, por lo tanto, este puede transmitir mensajes cifrados desde el momento en el que se conecta a la red.

- *OTAA (On The Air Activation):* En este método de activación, los nodos negocian las claves de cifrado con el servidor de red. Cada vez que un nodo se conecta a la red, se inicia un proceso de unión en el que el nodo y la red intercambian una serie de mensajes para la obtención de las claves de cifrado. Para ello se utiliza la clave AppKey que ha de ser proporcionada al nodo previamente. Tras el proceso de unión, tanto los servidores como los nodos conocen las claves y todos los mensajes posteriores al proceso de unión serán cifrados. Cada vez que el dispositivo se vuelva a conectar a la red o se reinicie se iniciará un nuevo proceso de unión y se generarán nuevas claves.

C. Otros mecanismos de seguridad en redes LoRaWAN

Además de los mecanismos de cifrado comentados para proteger la confidencialidad de los datos, LoRaWAN incorpora otros para proteger la integridad de los mensajes. De forma que el receptor pueda identificar al remitente del mensaje y verificar su identidad. Así se puede comprobar que el mensaje no haya sido alterado durante su envío. Existen principalmente dos mecanismos de protección:

- *Códigos MIC (Message Integrity Code):* Todos los mensajes que se transmiten en la red se firman con un código de integridad MIC. Estos códigos son generados a partir de diversos parámetros de la red como se muestra en [6]. Una vez generado, este código es insertado en la trama de LoRaWAN como campo de integridad. Cuando el mensaje llega al servidor, en primer lugar, se comprueba la integridad del mensaje chequeando su código MIC. Si es correcto, el mensaje es aceptado; si no lo es, el mensaje se rechaza. Mediante el código MIC se previene cualquier manipulación del mensaje por parte de un tercero.
- *Frame Counters*: Los códigos MIC se encargan de asegurar la integridad de los mensajes, sin embargo, no proporcionan seguridad ante ataques de repetición. Por lo tanto, un atacante, podría interceptar un mensaje vía radio y retransmitirlo a la red de forma repetitiva. Este ataque puede ser bloqueado de manera sencilla en una red LoRaWAN utilizando contadores de mensajes Frame Counters. Cuando un dispositivo se une a la red, dos contadores FCntUp y FCntDown se establecen a cero y se incrementan cada vez que se produce un mensaje en enlace ascendente o descendente. Si en el intercambio de mensajes se recibe uno cuyo valor de contador sea inferior al esperado, el mensaje será rechazado ya que procederá de un dispositivo malicioso. Además, los contadores de paquetes son utilizados para controlar el número de paquetes perdidos en la red.

Podemos observar como la especificación LoRaWAN, a priori, ofrece un protocolo robusto proporcionando mecanismos de seguridad en los dos pilares básicos de la seguridad en redes; la confidencialidad y la integridad de la información.

El gran auge de esta tecnología ha suscitado gran interés por investigadores, por ello se han realizado numerosos estudios con el fin de analizar la seguridad de las redes LoRaWAN. Se han llevado a cabo ataques en la capa física en [2], análisis de seguridad en los modos de activación de dispositivos en [3, 4, 5, 6] o análisis de seguridad en el manejo de las claves de cifrado por parte de los servidores de red en [6].

En este estudio se ha desplegado una red LoRaWAN y se ha realizado un análisis de seguridad desde una perspectiva global, de forma que, se considere la integración de una red LoRaWAN con otros sistemas habituales. Las redes LoRaWAN no son sistemas aislados, sino que, en una aplicación real, estas redes están integradas e intercambian información con otros sistemas. Los más comunes son bases de datos para el almacenamiento de la información, sistemas de monitorización de datos, telecontrol o plataformas en la nube. Entonces en el despliegue de una red LoRaWAN, no solo debe tenerse en cuenta la seguridad del protocolo del núcleo de la red sino también debe prestarse atención a la forma en la que la red intercambia información con otros sistemas.

III. ANÁLISIS DE SEGURIDAD EN LOS EXTREMOS

Las compañías y organizaciones que desarrollan las principales soluciones comerciales de servidores LoRaWAN, ofrecen un gran número de integraciones con terceros para el tratamiento y visualización de los datos de la red. Las integraciones más habituales son las basadas en aplicación ya que suponen la solución más simple y rápida de proporcionar a nuestras redes nuevos servicios como el almacenamiento de los datos. La mayoría de estas integraciones están basadas en el protocolo HTTP (Hypertext Transfer Protocol).

HTTP es un protocolo de transferencia de hiper texto en el que se transmite la información en texto plano. Por lo tanto, el uso de este protocolo puede provocar fallos en la seguridad del sistema como vamos a comprobar.

A. Escenario de pruebas

Para el análisis de la seguridad, se ha desplegado y puesto en marcha una red LoRaWAN real en la que se realizarán las pruebas. En la Fig.2 se muestra el diseño de la arquitectura de red realizada para la recogida, transmisión y presentación de los datos.

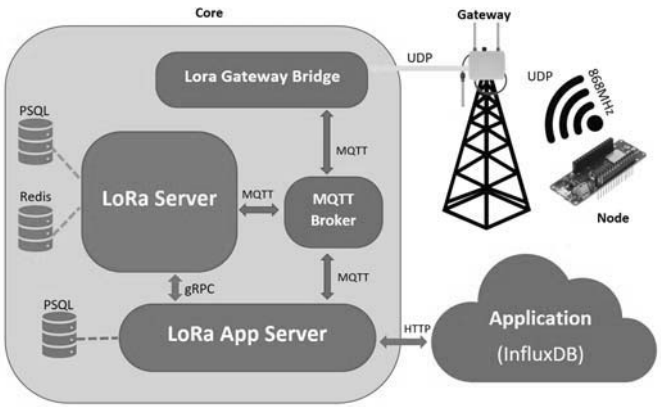


Fig. 2. Arquitectura de red LoRaWAN desplegada.

Como hardware de red se han utilizado nodos finales Arduino MKR1300 y una estación base Multitech modelo MTCDTIP-LEU1-266A-868.

Para el núcleo de la red se han utilizado los servidores del proyecto LoRaServer [9] además de un bróker MQTT (Message Queue Telemetry Transport) [10] para la comunicación entre el gateway y los servidores. Como integración o aplicación final, se ha optado por una base de datos InfluxDB [11] ya que el proyecto LoRaServer admite integración por aplicación con este tipo de base de datos.

El funcionamiento de la red consiste en que los nodos finales transmiten periódicamente datos de temperatura ambiental hacia el núcleo de la red. Entonces los servidores se comunican con una base de datos InfluxDB en la que, finalmente, se almacenan los datos de temperatura.

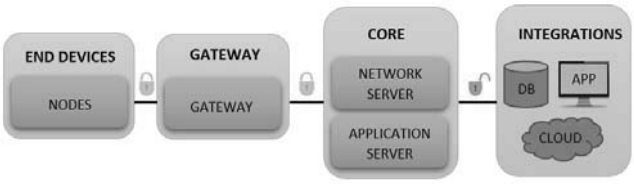


Fig. 3. Seguridad en la integración de servicios en una red LoRaWAN.

En la Fig.3. vemos una de las configuraciones más típicas de una red LoRaWAN en la que todos los datos van cifrados desde los nodos hasta el servidor de aplicaciones; sin embargo, posteriormente los datos se transmiten sin ningún tipo de seguridad hacia la base de datos o plataformas IoT. Este hecho abre una brecha de seguridad en la red y un vector de ataque.

Como se comentó en el apartado anterior, uno de los principales problemas recae en el uso del protocolo HTTP entra la comunicación del servidor de aplicaciones y la base de datos. Por lo tanto, esta configuración de red es vulnerable a un gran número de ciberataques como:

- *Sniffing:* Ataque pasivo que consiste en la captura de paquetes de la red.
- *MITM (Man in the Middle):* Ataque activo que consiste en la intromisión de un tercero en una comunicación con la finalidad de interceptar y modificar los datos.
- *DoS (Denial of Service):* Ataque mediante el cual se le deniega el acceso de un recurso a un dispositivo o conjunto de dispositivos.
- *ARP spoofing:* Consiste en la inundación de la red con mensajes ARP (Address Resolution Protocol) falsos cuyo fin es modificar las tablas ARP de los dispositivos para alterar la dirección destino de un mensaje.

A continuación, se exponen los dos principales problemas que presenta la topología planteada y se mostrarán los ataques realizados a la red.

B. Análisis de la red

En el caso de estudio, la topología planteada presenta dos principales problemas:

1. *Confidencialidad de la información:* Debido a que el protocolo HTTP no implementa ningún mecanismo de cifrado, un atacante puede obtener el contenido de los mensajes tan solo capturando tráfico en la red mediante técnicas de Sniffing. Para comprobarlo, se ha usado la herramienta Wireshark [12]. Hemos capturado tráfico HTTP filtrando los puertos 8086 que es el que utiliza InfluxDB por defecto para recibir las peticiones.

Fig. 4. Paquete interceptado mediante técnicas de sniffing utilizando la herramienta Wireshark entre el servidor de aplicaciones LoRa y la base de datos InfluxDB.

Como podemos apreciar en la Fig.4, se ha conseguido interceptar el paquete en el que viajaba el identificador del dispositivo y el dato de temperatura, que en este caso era de 19 grados.

2. *Integridad de la información:* Además de la confidencialidad de los datos, este escenario presenta vulnerabilidades en cuanto a integridad de la información. No solo es posible acceder a los datos de la red de forma ilícita, sino que también un

atacante podría modificar los datos interceptados alterando el contenido del mensaje. Esto es posible debido a que el protocolo HTTP no implementa por sí solo mecanismos de firmado de los mensajes.

Dada la configuración de la Fig.5, se ha recreado la acción de un atacante que accederá a la red, interceptará los mensajes y los modificará haciendo que en la base de datos se guarden datos erróneos.

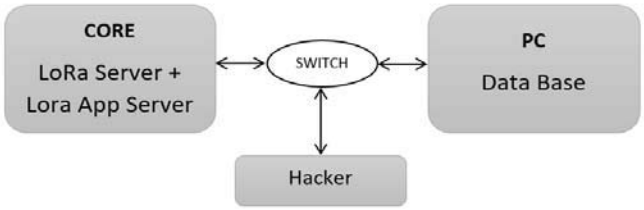


Fig. 5. Topología de red empleada para la realización de un ataque MITM entre el servidor de aplicaciones LoRa y la base de datos InfluxDB.

Para llevar a cabo el experimento, se ha realizado un ataque ARP spoofing en el que, nosotros como atacantes, nos hacemos pasar por la base de datos. De esta forma el servidor de aplicaciones LoRa nos enviará los datos a nosotros. Una vez interceptados los paquetes, los analizamos para localizar los campos que nos interesan, en nuestro caso la temperatura. Una vez localizado el campo, lo modificamos con herramientas como Hexinject [13] o Polymorph [14]. Por último, reenviamos el paquete modificado hacia la base de datos InfluxDB.

El resultado del ataque lo vemos en la Fig.6 en la que se muestra el mensaje alterado que hemos enviado a la base de datos. Posteriormente, accediendo a InfluxDB vemos como se están almacenando datos de temperatura de 22 grados en vez de los 19 reales que están transmitiendo los nodos.

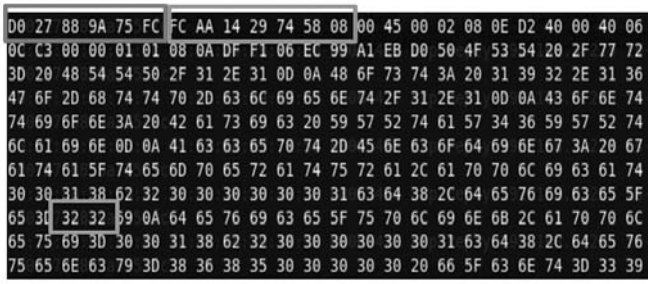


Fig. 6. Intercepción y modificación de datos mediante la herramienta Hexinject. Datos marcados: MAC destino (hex), Mac Origen (hex) y dato de temperatura alterado (hex).

Mediante la realización de los ataques anteriores se ha comprobado que la seguridad de la red puede verse comprometida debido a que el protocolo HTTP no implementa mecanismos criptográficos. A pesar de que es ampliamente conocido que el protocolo HTTP es inseguro, a menudo es utilizado por desarrolladores de redes debido a su facilidad y rapidez de implementación.

Para solventar los problemas de seguridad de HTTP, es necesario implementar este protocolo junto con protocolos criptográficos como TLS (Transport Layer Security) para realizar comunicaciones seguras. Este protocolo utiliza certificados X.509 que deben ser generados por una autoridad certificadora CA. Estos han de ser adquiridos y configurados en los servidores y las bases de datos o aplicaciones. Con su uso, los mensajes intercambiados se cifrarán y verificarán comprobando, en cada transferencia, el contenido del mensaje y el remitente del mismo.

IV. Buenas prácticas de diseño

Toda red debe diseñarse teniendo en cuenta principalmente dos factores, la funcionalidad y la seguridad. En primer lugar, la red debe satisfacer las necesidades funcionales del diseño y en segundo lugar se debe asegurar la confidencialidad de los datos, la identidad de los usuarios y la protección de la infraestructura.

No es posible asegurar una red al cien por ciento frente a ciber ataques. Sin embargo, con buenas prácticas de diseño podemos agregar un mayor grado de seguridad a la red reduciendo vulnerabilidades.

Los ataques más comunes en las redes IoT, y por tanto en las redes LoRaWAN, se producen en los extremos ya que es aquí donde se encuentran los elementos más expuestos de la red. Por tanto, el diseño de un modelo de seguridad de cualquier red LoRaWAN debe asegurar la protección de los tres elementos principales: los nodos o dispositivos finales, el core o núcleo de red y las aplicaciones finales.

A. *Capa física*

Los nodos componen uno de los extremos de la red. Los ataques DoS, la suplantación de sesión o el acceso de nodos fraudulentos a la red son los ataques más comunes en esta capa. Para reducir el impacto de estos y otros ataques en esta capa es importante, en primer lugar, realizar un buen manejo de las claves de los dispositivos ya que estas son las llaves que permiten el acceso a la red. Ante todo, debemos proteger la confidencialidad de las claves de cifrado independientemente del modo de activación que estemos utilizando, OTAA o ABP. Las claves deben ser únicas por dispositivo y ser generadas de forma aleatoria para que en el caso de que una clave sea filtrada, no se comprometa la seguridad de todos los nodos de la red.

Es aconsejable siempre el uso del modo de activación OTAA ya que es el mecanismo más seguro de unión a la red. Esto es debido a que los dispositivos negocian las claves con el servidor y estas se vuelven a generar cada vez que el nodo se vuelva a conectar a la red.

Otro aspecto importante es la elección correcta de los dispositivos finales. En los nodos, además de la lógica de programación, se encuentran almacenadas las claves de cifrado de la sesión. Por lo tanto, si una persona accediera físicamente al nodo podría descargar el código y por tanto las claves del dispositivo. Por ello, es imprescindible proteger, en la medida de lo posible, el acceso físico a la electrónica del nodo ya que, estos suelen situarse en lugares remotos y no siempre se tiene control sobre ellos. Deben seleccionarse dispositivos finales que permitan proteger los programas frente a lectura mediante métodos criptográficos o credenciales.

B. *Capa de red*

La red LoRaWAN está compuesta por servidores de red y apliacación. Esta es la capa donde la acción del desarrollador

tiene menor impacto. Lo más habitual en el despliegue de redes LoRaWAN es el uso de servidores de terceros, bien soluciones comerciales o bien servidores de código abierto. Es importante seleccionar servidores de red testados en el ámbito de la seguridad ya que; tanto el servidor de red como el de aplicación son los últimos responsables del manejo de las claves de cifrado y verificación de los mensajes. Además, es importante mantener los servidores actualizados y utilizar las últimas versiones disponibles en las que continuamente se detectan y corrigen fallos y vulnerabilidades.

En los servidores de la red existen ciertos parámetros, orientados a la seguridad, que pueden ser configurados por el desarrollador. Estos deben ser conocidos y aplicados siempre que sea posible. Por ejemplo, se puede configurar el campo "Frame Counter" desde el servidor de aplicaciones para evitar ataques de repetición tal y como se vio en el apartado II. Además, puede activarse en la red la confirmación de mensajes ACK para detectar ataques de denegación de servicio DoS. De esta forma, los nodos pueden confirmar si los mensajes están llegando a la red. Estos parámetros conforman un grado extra de seguridad en la red, sin embargo, debe considerarse el impacto que el uso de estos, y otros mecanismos de protección, pueden ocasionar en la red. Hay que tener en cuenta que todo elemento de seguridad utilizado para proteger una red implica un gasto extra de los recursos de la misma. Por ejemplo, la activación de la confirmación de mensajes ACK aumentará el tráfico en la red debido a que, por cada mensaje recibido, la red generará otro confirmando los datos. Por lo tanto, siempre debe estudiarse su utilización en cada escenario.

C. Capa de aplicación

Las aplicaciones finales junto con la capa física, son las partes más expuestas de la red y por tanto son las que más ataques sufren. Algunos ejemplos de estos son los mostrados en el apartado III de este documento.

Las tareas de protección de la capa de aplicación se basan en una serie de métodos para proteger el acceso a los dispositivos de la red y el acceso a la información a través de una red local o internet.

Para garantizar la seguridad en esta capa debemos diseñar un modelo de protección dividido en varios niveles de seguridad.

En primer lugar, debemos proteger e impedir el acceso a la red en la que se encuentren los servidores, bases de datos y las aplicaciones. Para ello, debe evitarse el uso de puntos de acceso abiertos. Siempre deben utilizarse métodos de autentificación con contraseñas seguras para acceder a la red. En aconsejable aislar, en la medida de lo posible, la red LoRaWAN del resto de redes corporativas. Además de los servidores, también debe limitarse la exposición del gateway de la red LoRaWAN y el acceso al mismo. En gran número de ocasiones, estos elementos cuentan con interfaces web para facilitar la tarea de configuración y puesta en marcha. En la mayoría de los casos, estas interfaces no son seguras. Una de las principales vulnerabilidades que presentan es la utilización de credenciales por defecto. Además, no siempre estas interfaces cuentan con mecanismos de protección frente a ataques por repetición o fuerza bruta y mediante exploits, que son códigos utilizados para sacar provecho de vulnerabilidades de seguridad de un sistema, es posible obtener las credenciales en cortos periodos de tiempo. Para mitigar estos problemas en primer lugar sería interesante desactivar este tipo de herramientas y si no fuera posible, debemos cambiar las credenciales por defecto y aseguraros que nuestro gateway limita el número de intentos de acceso fallidos.

Las acciones mostradas hasta ahora comprenderían nuestra primera capa de seguridad en la que se limita el acceso a la red y a los dispositivos de la misma. La siguiente capa de protección está enfocada a los datos; y es que, si la primera capa falla y un usuario no autorizado consiguiera acceder a la red o a un dispositivo, debemos impedir que este pueda acceder a la información. Para proteger la información, debemos asegurar los protocolos y los canales de comunicación que usan los servidores y las aplicaciones para intercambiar datos. Para llevar a cabo estas acciones es imprescindible la utilización de protocolos seguros como HTTPS, es decir, el uso del protocolo HTTP junto con el protocolo TLS (Transport Layer Security) en las interfaces web de servidores y aplicaciones. De esta forma todos los datos son cifrados y verificados desde el primer al último elemento de la red añadiendo un grado extra de protección ante ataques como Sniffing, Spoofing o MITM. Como se mencionó en el apartado III, para implementar el protocolo TLS es necesario adquirir un certificado de una autoridad certificadora oficial CA. En el escenario que se plantea en este estudio no es posible la generación de auto certificados firmados por nosotros mismos ya que el servidor rechaza la conexión con la base de datos InfluxDB al no poder validar el certificado ni reconocer a la autoridad que lo firmó.

V. Conclusiones

En este trabajo se ha llevado a cabo un análisis de seguridad de una red LoRaWAN en un escenario real en el que la red está conectada a otros sistemas con los que intercambia información. En primer lugar, se han analizado los métodos de seguridad que implementa el protocolo LoRaWAN en una red aislada. Después, se realizó la comunicación de nuestra red con una base de datos InfluxDB con el fin de almacenar los datos transmitidos por los nodos y recrear, de esta manera, un caso real de uso de una red LoRaWAN. Tras esto, se realizó un análisis práctico de seguridad sometiendo la red a varios ataques. En estos, se detectaron una serie de riesgos que surgen al integrar una red LoRaWAN con otros sistemas. Se ha comprobado que para realizar un despliegue seguro de red es necesario la utilización de certificados X.509 TLS de una autoridad certificadora CA entre las aplicaciones y los servidores de la red ya que, de esta forma, se asegura la confidencialidad y la integridad de la información extremo a extremo. Finalmente, se ha aportado una guía de diseño paso a paso en la que se proponen buenas prácticas de diseño para que tanto usuarios como desarrolladores de redes LoRaWAN puedan incrementar la seguridad y confianza de sus redes. Para ello se han aportado soluciones y recomendaciones que incrementan la seguridad en nodos, gateways, servidores y aplicaciones finales.

Referencias

[1] Julián Fernández-Navajas, "Ejemplo de bibliografía", *Actas de las XIV Jornadas de Ingeniería Telemática*, vol. 1, pp. 1-10, 2019.

[2] Aras, E., Ramachandran, G. S., Lawrence, P., & Hughes, D. (2017, June). Exploring the security vulnerabilities of LoRa. In *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)* (pp. 1-6). IEEE.

[3] Kim, J., & Song, J. (2017). A dual key-based activation scheme for secure LoRaWAN. *Wireless Communications and Mobile Computing*, *2017*.

[4] Na, S., Hwang, D., Shin, W., & Kim, K. H. (2017, January). Scenario and countermeasure for replay attack using join request messages in LoRaWAN. In *2017 International Conference on Information Networking (ICOIN)* (pp. 718-720). IEEE.

[5] Yang, X., Karampatzakis, E., Doerr, C., & Kuipers, F. (2018, April). Security Vulnerabilities in LoRaWAN. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)* (pp. 129-140). IEEE.

[6] Miller, R. (2016). Lora security: Building a secure lora solution. *MWR Labs Whitepaper*.

[7] You, I., Kwon, S., Choudhary, G., Sharma, V., & Seo, J. (2018). An enhanced LoRaWAN security protocol for privacy preservation in IoT with a case study on a smart factory-enabled parking system. *Sensors*, *18*(6), 1888.

[8] Rubio-Aparicio, J., Cerdan-Cartagena, F., Suardiaz-Muro, J., & Ybarra-Moreno, J. (2019). Design and Implementation of a Mixed IoT LPWAN Network Architecture. *Sensors*, *19*(3), 675.

[9] LoRaServer. Disponible online: https://www.loraserver.io/. Accedida el 24/05/2019

[10] MQTT. Disponible online: http://mqtt.org/. Accedida el 24/05/2019

[11] Influxdata. Disponible online: https://www.influxdata.com/. Accedida el 24/05/2019

[12] Wireshark. Disponible online: https://www.wireshark.org/. Accedida el 24/05/2019

[13] Hexinject. Disponible online: http://hexinject.sourceforge.net/. Accedida el 24/05/2019

[14] Poymorph. Disponible online: https://github.com/shramos/polymorph. Accedida el 24/05/2019

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Prueba de concepto de Autoridad de Certificación usando Computación Segura Multiparte

Daniel Morales, Isaac Agudo.
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga
Edificio de Investigación Ada Byron, 29010
d.moralesescalera@lcc.uma.es, isaac@lcc.uma.es

***Resumen*—Este trabajo pretende analizar el paradigma de la Computación Segura Multiparte y sus posibles aplicaciones en el campo de la criptografía. Se plantea como modelo alternativo, más escalable y seguro al uso de módulos hardware de seguridad para aplicaciones que requieran de Terceras Partes Confiables. Concretamente, se ha integrado un protocolo de criptografía RSA multiparte con la librería *certbuilder*, para la creación de certificados X.509. De esta forma se asegura que la creación de los certificados raíz de la Infraestructura de Clave Pública se realiza de forma que la generación de claves y firma de éste se ejecute íntegramente sobre el sistema multiparte, con un modelo de tres partes que trabaja con circuitos aritméticos, sin que ninguna de ellas, de forma aislada, tenga posibilidad de comprometer la clave privada correspondiente. Para comprobar la viabilidad del sistema se han realizado pruebas de generación de certificados con diferentes longitudes de clave, siendo el proceso determinante la creación de las claves. Los elevados tiempos hacen que una aplicación como esta no sea asumible en otros escenarios, pero creemos que para el caso de la creación de los certificados raíz de una infraestructura de clave pública las garantías avanzadas de seguridad compensan el tiempo extra.**

***Palabras Clave*—Computación Segura Multiparte, Criptografía, RSA, Secreto Compartido, Autoridad de Certificación.**

I. INTRODUCCIÓN

Uno de los problemas a los que se enfrenta el modelo actual de Internet es la confiabilidad. Otros problemas como la confidencialidad o la integridad han sido resueltos gracias a la criptografía moderna, mediante protocolos criptográficos simétricos y asimétricos. Sin embargo, en contadas ocasiones, se necesita algún mecanismo que verifique la autenticidad de un actor en Internet.

Este problema se ha solucionado tradicionalmente mediante la integración de Terceras Partes Confiables, actores intermediarios que proporcionan algún tipo de servicio en un entorno no confiable. Sin embargo, esta solución traslada el problema, pues la autenticidad de la Tercera Parte Confiable puede no estar garantizada.

A. *Modelos de Certificación Jerárquica*

Para solucionar el problema presentado, aparecen los modelos de Certificación Jerárquica. Un certificado es un documento digital que identifica a un usuario como tal, proporcionando su información relevante y verificado por una autoridad de certificación. El modelo más extendido es el de certificados de clave pública del estándar X.509. Este tipo de certificado contiene la clave pública del agente a verificar, además de estar firmado digitalmente por una autoridad de certificación mediante su clave privada.

De esta forma, distintas entidades de certificación pueden certificarse unas a otras, dando lugar a una cadena de confianza jerárquica, en cuya cumbre se hallan las autoridades de certificación con potestad para auto-firmar sus propios certificados. Estos certificados se conocen como certificados raíz.

Figura 1. Jerarquía de Autoridades de Certificación

La problemática presentada en este trabajo es que dichas autoridades pueden suponer un punto simple de fallo, de manera que exponer su seguridad puede exponer la seguridad de toda la cadena de confianza que se construye bajo ellas.

B. *Módulos de seguridad hardware*

Una solución extendida para proteger los objetos criptográficos sensibles son los Módulos de Seguridad Hardware. Estos dispositivos protegen las claves dentro de un entorno seguro a nivel de hardware, lo que presenta una capa extra de seguridad ante una posible intrusión en el sistema.

Sin embargo, estos dispositivos pueden presentar diversos inconvenientes, como su elevado coste, con soluciones que parten de los 20000$ [1], además de la exposición a ataques de canal lateral o la pérdida de flexibilidad y escalabilidad. Es por esto que han surgido propuestas a nivel de software, con la problemática de que su nivel de seguridad no alcanza al de la solución hardware.

La propuesta de este trabajo es incluir la funcionalidad de un Módulo de Seguridad, a nivel de software y en un entorno descentralizado, mediante CSM (Computación Segura Multiparte). La ventaja de esta solución es que CSM garantiza la seguridad mediante modelos matemáticos probados.

Aunque las aplicaciones de CSM están a penas empezando a ver la luz, casos como el de Unbound Tech [2] corroboran la viabilidad de este tipo de soluciones. Su propuesta de gestión de claves mediante CSM ha obtenido los niveles de certificación 1 y 2 del estándar FIPS 140-2, que se encarga de acreditar los módulos criptográficos.

II. PARADIGMA DE LA COMPUTACIÓN SEGURA MULTIPARTE

A. *Origen y evolución*

CSM nace como un modelo para evaluar una función entre varios participantes, sustituyendo a una Tercera Parte Confiable por un protocolo, de forma que ninguno conozca los datos de entrada de los demás pero todos obtengan el resultado de la computación.

Sus inicios se remontan a la aparición del protocolo *Garbled Circuit* [3], basado en circuitos booleanos, aunque posteriormente surgieron otras opciones basadas en circuitos aritméticos.

B. *Modelos de seguridad*

A la hora de diseñar un protocolo multiparte se han de cumplir una serie de requisitos para garantizar la seguridad, como privacidad, exactitud, independencia de las entradas, etc. Para abordar estos requisitos se diseñan unos modelos genéricos que definen los umbrales máximos permitidos sobre los que se garantiza la seguridad de un protocolo [4].

B1. Modelo de comunicación: El modelo de comunicación determina las características de los canales por los que los participantes intercambian la información entre ellos. Estos canales pueden ser *unicast* o *broadcast*. Además, pueden asumirse como *seguros*, *autenticados* o *inseguros*. También se determina la temporalidad de la información, que se puede enviar de forma *síncrona* o *asíncrona*.

B2. Modelo de adversario: El objetivo de los adversarios es que la computación no finalice correctamente. Se clasifican en función de las capacidades de acción que tienen sobre el resto de los participantes de la computación. Un *adversario pasivo* es aquel que trata de obtener información de uno o más participantes, pero no puede desviar el flujo de acción natural del protocolo. Un *adversario activo*, por su parte, tiene total control sobre algún participante, por lo que puede modificar su funcionamiento respecto a la ejecución del protocolo. Además, puede ser *estático* o *dinámico* en función de si los participantes corruptos se definen previamente a la computación, o pueden variar a lo largo de la misma.

B3. Modelo de computación: Un protocolo de CSM se define sobre un lenguaje específico o modelo matemático sobre el que se sustentan las operaciones. Los modelos tradicionales de la literatura definen circuitos sobre *campos booleanos*, aunque cada vez es más habitual adoptar soluciones basadas en circuitos aritméticos, que se definen sobre *campos finitos* $(F, +, *)$. Lo bueno de este modelo es que cualquier función computable puede ser expresada como un circuito.

III. ESQUEMAS DE SECRETO COMPARTIDO

Los protocolos que trabajan con circuitos aritméticos suelen basarse en esquemas de secreto compartido. Estos algoritmos criptográficos permiten que un secreto S pueda ser compartido con diversos participantes de forma fragmentada, necesitando de la colaboración de varios de ellos para su reconstrucción. Por ello se conocen como esquemas de umbral (t, n), siendo t el mínimo de fragmentos necesarios para la reconstrucción y n el total de fragmentos existentes.

El caso más extendido de esquema de secreto compartido es el de Shamir [5]. Este esquema construye polinomios aleatorios para dividir el secreto en fragmentos y utiliza la interpolación de Lagrange para su reconstrucción. Aprovecha la propiedad de que para definir un polinomio de grado k se necesitan $k + 1$ puntos. La construcción de estos polinomios se presenta en la Ec.1, donde los $t - 1$ coeficientes $\{a_1, ..., a_{t-1}\}$ se eligen al azar, mientras que a_0 representa al secreto S.

$$f(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{t-1} x^{t-1} \qquad (1)$$

El polinomio asociado al secreto se utiliza para generar puntos de la función y a cada participante se le entrega el valor de un punto. Para reconstruir el secreto se interpola el polinomio con la fórmula de Lagrange (Ec.2) y se calcula su valor para el caso particular $x = 0$.

$$f(x) = \sum_{i=1}^{t} s_i \prod_{j=1}^{t, i \neq j} \frac{x - x_j}{x_i - x_j} \qquad (2)$$

A. *Suma y multiplicación*

Estas operaciones se definen como una serie de cálculos sobre los coeficientes de los polinomios que representan a los secretos compartidos.

En el caso de la suma se trata de una operación sencilla ya que no necesita compartir secretos adicionales. Para construir el polinomio suma que represente la suma de dos secretos, simplemente se suman los coeficientes de los polinomios previamente compartidos que tienen el mismo grado. En la Ec. 3 se presenta el polinomio $h(x)$ como resultado de sumar dos polinomios $f(x)$ y $g(x)$.

$$h(x) = (s_f + s_g) + (r_{1_f} + r_{1_g})x + ... + (r_{t-1_f} + r_{t-1_g})x^{t-1} \quad (3)$$

En el caso de la multiplicación se necesita compartir un secreto adicional, lo que eleva el coste de la computación. El motivo es que la multiplicación de dos polinomios de grado $t-1$ da como resultado un polinomio de grado $2t-2$ (Ec. 4), el cual ha de reducirse a grado $t-1$ para poder interpolarse. Para conseguir esto, cuando cada participante multiplica los dos fragmentos originales de los secretos construye con el resultado un nuevo polinomio aleatorio, el cual conduce a otra ronda de compartición de secretos.

$$h(x) = (s_f s_g) + r_1 x + r_2 x^2 + ... + r_{2t-2} x^{2t-2} \quad (4)$$

IV. VIRTUAL IDEAL FUNCTIONALITY FRAMEWORK

VIFF es un framework para desarrollar prototipos de aplicaciones CSM, escrito en Python. La comodidad de utilizar un framework como éste es que oculta la complejidad matemática subyacente. De esta forma, el desarrollador puede centrarse exclusivamente en la funcionalidad que desea implementar.

VIFF presenta una arquitectura de tres capas, en la que cada capa da servicio a la capa superior, siendo la capa de las operaciones CSM (basada en el esquema de Shamir) la que da servicio a las aplicaciones que se pretenden desarrollar, tal y como se muestra en la Fig. 2.

Figura 2. Pila de implementación de VIFF

Para implementar la capa de comunicación de red, VIFF hace uso de una librería de Python llamada *Twisted*, la cual facilita el desarrollo de modelos de comunicación asíncronos mediante un modelo dirigido por eventos.

Algunas consideraciones de seguridad del modelo desarrollado en VIFF son que el modelo de adversario es pasivo, por lo que puede corromper hasta un máximo de 1/2 de los participantes, y que está limitado computacionalmente a un tiempo polinómico.

V. RSA DISTRIBUIDO PARA INFRAESTRUCTURA DE CLAVE PÚBLICA

El objetivo final del trabajo consiste en integrar un protocolo RSA distribuido que implementa las funciones de generación, cifrado y firma, desarrollado en una tesis [6] e implementado en VIFF, con una librería para la generación de certificados de clave pública X.509.

Por simplicidad de integración, se ha optado por usar la librería *certbuilder* [1], que además de estar desarrollada también en Python, tiene un diseño modular que permite fácilmente engarzar los protocolos de CSM para RSA distribuido. Por consiguiente, tanto para generar las claves RSA como para realizar la firma, la librería *certbuilder* delega en un conjunto de nodos CSM (3 nodos para esta solución concreta).

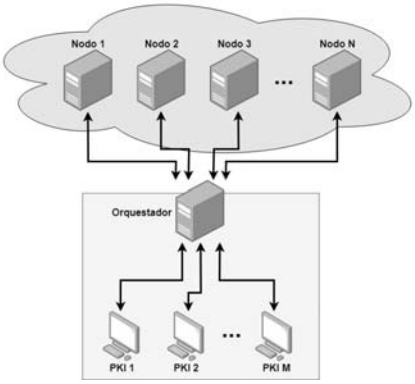


Figura 3. Arquitectura del sistema *certbuilder* multiparte

La topología resultante emula un sistema distribuido en la nube, al que el cliente accede desde una red externa mediante un Orquestador (Fig. 3). Este elemento intermedio es necesario para operaciones de coordinación entre los diferentes nodos, los cuales se encargan de realizar las operaciones CSM.

El trabajo realizado implementa la interfaz necesaria para que el cliente *certbuilder* pueda comunicarse de forma coherente con los nodos CSM, como puede apreciarse en la Fig. 4. Los nodos se despliegan en uno o varios entornos de prestadores de servicios, asignando tres de ellos para cada operación particular.

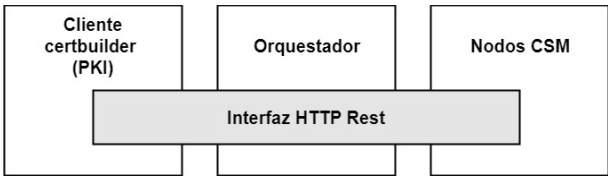


Figura 4. Integración de certbuilder con CSM

A. RSA distribuido con VIFF

Para la generación de las claves se producen varias rondas de generación de los parámetros correspondientes de forma distribuida (Fig. 5), de los cuales se va comprobando

[1] https://github.com/wbond/certbuilder

su validez por los requisitos de primalidad. Si en una determinada fase no se cumplen los requisitos, se vuelve a alguna fase anterior. Si por el contrario, se cumplen, se avanza de fase. Este proceso constituye el principal cuello de botella del protocolo, ya que la generación de los parámetros se hace de forma aleatoria y se ha de volver numerosas veces al inicio del procedimiento, con el gran coste computacional que ello supone por la inclusión del modelo CSM.

En la Ec. 5 puede apreciarse un ejemplo del parámetro público N, generado a partir de los valores privados de cada participante. La seguridad reside en la realización de las operaciones de suma y multiplicación de los parámetros privados mediante CSM, por lo que se necesita comprometer a más de un participante para conocer sus valores.

$$N = (p_1 + p_2 + p_3)(q_1 + q_2 + q_3) \tag{5}$$

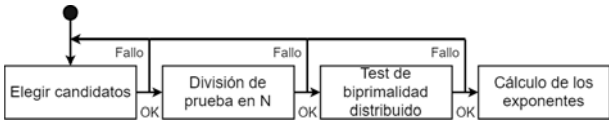


Figura 5. Fases del protocolo RSA distribuido

B. Extensión CSM para certbuilder

Certbuilder hace uso librerías criptográficas para la generación de los certificados. Para no inhabilitar el funcionamiento original de la librería, la solución propuesta ha integrado dos nuevas funciones implementadas en el código fuente, *generate_pair_mpc()* y *build_mpc()*. La primera función se encarga de solicitar al orquestador que inicie el procedimiento de generación de claves en los nodos, entregando como respuesta un identificador asociado al par de claves a generar. El procedimiento es asíncrono, y para solicitar la clave se necesita consultar al orquestador mediante un método HTTP adicional, cuya implementación en el cliente se ha dejado libre. La segunda función calcula el hash del certificado a firmar y se lo envía al orquestador, quien solicita a los nodos que posean la clave a emplear que firmen el hash de forma distribuida. Como resultado, el cliente obtiene el valor de la firma que ha de incorporar al certificado.

C. Interfaz orquestador-nodos

El orquestador coordina a los nodos necesarios para CSM. Pese a que no maneja información crítica en sí misma, presenta un punto de fallo respecto a gestión de acceso a las claves y privilegios. Este aspecto de seguridad se ha omitido en un primer momento, priorizando la funcionalidad del sistema y la seguridad ofrecida en la parte distribuida.

Para realizar varios procesos simultáneos se ha utilizado un sistema de gestión de hebras, donde cada una se encarga de generar un par de claves o de firmar un hash. Cada operación CSM se realiza vinculada a un puerto concreto en cada nodo, por tanto, el orquestador se encarga de solicitar a los nodos sus puertos disponibles, previamente al inicio de cada operación.

Respecto a la generación de las claves, los tres nodos invocan un script de Python en el que se ejecuta el protocolo RSA distribuido previamente expuesto. De esta forma generan los parámetros de las claves, que son guardados localmente en cada nodo. Al finalizar el proceso se genera la clave pública, que es el objeto que se entrega al cliente en formato ASN.1 DER.

El proceso de firma lo implementa otro script de Python basado en el mismo protocolo distribuido. En este caso, el orquestador envía el hash a los tres nodos correspondientes, que devuelven el valor de la firma obtenida.

Cuadro I
GENERACIÓN DE CLAVES RSA CON CSM [6].

Nº bits	Tiempo promedio	Ratio
256	0,97 min	3,83
512	3,78 min	3,89
1024	32,61 min	8,64
2048	120,87 min	3,71

VI. CONCLUSIONES

Pese a que el sistema está en desarrollo[2][3], se han obtenido resultados aceptables en pruebas realizadas en un entorno local.

La ventaja es que ofrece un modelo más escalable, mediante uso por demanda, que no necesita de largos tiempos de amortización del sistema.

Los elevados tiempos de generación de claves pueden ser aceptables para un modelo en el que CSM se utiliza solo en la capa de certificación raíz (Tabla I). Los tiempos de firma, por otro lado, no superan los diez segundos, lo que hace que, una vez generadas las claves, no suponga un coste adicional muy elevado.

Por ello, se contempla la opción de modificar el protocolo CSM empleado para tratar de obtener un sistema más ligero y versátil.

En conclusión, CSM ofrece un modelo de altas garantías de seguridad, pero a costa de grandes tiempos de computación. Es por ello que sus aplicaciones parecen tender a protocolos de computación ligera (tipo IOT) o a procesos críticos que no dependan en gran medidad del tiempo.

REFERENCIAS

[1] Logan Harbaugh. Thales nShield Connect offers enterprise-class key management. September 2009. Available: https://www.networkworld.com/article/2246758/thales-nshield-connect-offers-enterprise-class-key-management.html
[2] Lindell. Unbound receives FIPS 140-2 Level 1 and FIPS 140-2 Level 2 certification, May 2019. Available: https://www.unboundtech.com/unbound-receives-fips-140-2-certification/
[3] A. Chi-Chih Yao, How to Generate and Exchange Secrets (Extended Abstract), 1986, pp. 162-167.
[4] Hirt Martin, Multi-Party Computation: Efficient Protocols, General Adversaries, and Voting, 2001.
[5] Shamir Adi, How to Share a Secret, November 1979.
[6] Mauland Atle, Realizing Distributed RSA using Secure Multiparty Computations, July 2009.

[2]https://github.com/dmoralesescalera/RSA-MPC-server
[3]https://github.com/dmoralesescalera/certbuilder


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Desarrollo de un semáforo inteligente basado en comunicaciones seguras

Manuel Montenegro-Gómez, Isaac Agudo
Departamento de Lenguajes y Ciencias de la Computación,
Universidad de Málaga
{mmg,isaac}@lcc.uma.es

***Resumen*—En los nuevos paradigmas de movilidad surgidos durante los últimos años y en aquellos aún por llegar ha quedado patente la necesidad de modernizar la infraestructura viaria y los elementos de señalización y gestión del tráfico. En el presente trabajo se presenta una propuesta para esta nueva generación de dispositivos de gestión del tráfico: un prototipo de semáforo inteligente conectado que implementa diversas medidas de seguridad. Además de las tradicionales señales luminosas, los usuarios de la vía pueden conocer a través de sus dispositivos el estado del semáforo, además de otra información complementaria a través de la difusión de mensajes BLE firmados con criptografía de curva elíptica. A su vez, el semáforo puede ser gestionado remotamente a través de la tecnología LTE Cat M1 protegida por TLS. Esto abre la puerta, entre otros, a facilitar el tránsito de los vehículos de emergencia cuando estos se acercan a un cruce o modificar el tiempo de los estados del ciclo en función de las necesidades del tráfico.**

***Palabras Clave*—semáforo inteligente, sistemas de transporte inteligente, seguridad en las comunicaciones, bluetooth low energy, curva elíptica, nrf52840, tls.**

I. INTRODUCCIÓN

Todo parece indicar que, en los próximos años, el sector de la automoción será el epicentro de profundas transformaciones tecnológicas. La aparición de nuevos paradigmas de movilidad, una flota de vehículos creciente en número, la implantación de restricciones anticontaminación, la modernización de los vehículos, cada vez con mayor capacidad de asistencia al conductor e incluso de tomar el control, así como el uso de nuevos medios de transporte multimodales compartidos y los vehículos de movilidad personal (VMP) son algunos de las piezas del puzzle que conforman y motivan la investigación y desarrollo de los Sistemas de Transporte Inteligente.

Los medios tradicionales de gestión del tráfico, tales como semáforos y señales viales, tienen ahora la necesidad de aportar una mayor cantidad de información, dotándola de una mayor seguridad, rapidez y tolerancia a fallos, en la medida en que esta información se hace crítica para el funcionamiento de estos nuevos sistemas.

Como ejemplo, la compañía Audi lleva más de dos años prestando un servicio de información de semáforos Audi Traffic Light Information [1] con las funciones Time-to-Green en EEUU, y este año 2019 comienzan las pruebas en Europa. También se ha añadido la funcionalidad Green Light Optimized Speed Advisory (GLOSA), servicio que se apoya en una infraestructura paralela a los semáforos ya existentes y que aporta información de los semáforos cercanos basándose en su posición GPS.

Una alternativa a la propuesta por Audi sería proporcionar a los semáforos y demás señales de tráfico un mecanismo de comunicación directo con los vehículos. En este sentido, el desarrollo de un semáforo inteligente seguro trata de incorporar una solución confiable y económica a este ecosistema, que englobe no solo a los vehículos tradicionales sino a también los VMP.

En el presente trabajo se ha implementado un mecanismo seguro de intercambio de información viaria basado en el protocolo Bluetooth Low Energy versión 5.0, disponible en la mayoría de los teléfonos inteligentes [2], de forma que los usuarios pueda interactuar con las señales inteligentes usando su teléfono. Esto requiere del desarrollo de una infraestructura de clave pública que permita autenticar a los dispositivos desplegados en las vías, siguiendo para ello el mismo camino que los estándares ITS [3] (Intelligent Transport Systems).

Otro requisito central es permitir que determinadas señales viarias puedan estar conectadas a internet. Para ello, se ha analizado el uso del protocolo TLS sobre tecnologías estándar de la 3GPP para Low Power Wide Access Networks (LPWAN). [4]

Para el desarrollo del prototipo se ha utilizado un microcontrolador (nRF52840[1]) de Nordic Semiconductor, con soporte para la version 5.0 de Bluetooth Low Energy y que además cuenta con capacidad de realizar cálculos criptográficos acelerados por hardware.

[1] https://www.nordicsemi.com/Products/Low-power-short-range-wireless/nRF52840

II. SEGURIDAD EN ITS

El ETSI (European Telecommunications Standards Institute) ha elaborado un marco de estandarización sobre las comunicaciones en los ITS que engloba todo tipo de comunicaciones entre los vehículos y su entorno, así como aspectos relativos a la seguridad y a la privacidad. (informe técnico ETSI TR 101 607 [3]). La ETSI también define un conjunto de aplicaciones para la mejora del tráfico basadas en las comunicaciones dedicadas de corto alcance (DSRC por sus siglas en inglés) o los ITS cooperativos (c-ITS).

Además, el Institute of Electrical and Electronics Engineers (IEEE) define el estándar IEEE 802.11p [5] para el acceso inalámbrico en entornos vehiculares (WAVE, por sus siglas en inglés).

Los mensajes de tipo Cooperative Awareness Message (CAM), definidos en la norma europea ETSI EN 302 637-2 [6], están diseñados para ser transmitidos de punto a multipunto y transportan información sobre el vehículo o el entorno. Van además firmados digitalmente, como figura en el estándar [7] e incluyen la información siguiente:

- Versión del protocolo.
- Cabecera: información del dispositivo que generó la firma, hash del certificado asociado al dispositivo, marca de tiempo de generación del mensaje e identificación de la aplicación.
- Payload o carga útil del mensaje.
- Firma digital del mensaje mediante el algoritmo ECDSA con la curva NIST P-256.

Para distribuir las claves públicas necesarias para verificar los mensajes CAM, se define también un formato compacto de certificado digital, con un tamaño de 132 octetos con los campos siguientes:

- Versión del certificado.
- Identificador del dispositivo firmante del certificado.
- Información sobre a quién va dirigido el certificado.
- Clave pública de la firma.
- Firma digital, codificada como un par de puntos.

En el cuadro I se reflejan los tamaños de cada unos de los campos de los mensajes mencionados.

Cuadro I: Formato del mensaje CAM

Certificado CAM		Mensaje seguro CAM	
Elemento	Octetos	Elemento	Octetos
version	1	version	1
signer	9	header	24
subject	2	payload	x
public key	44	signature	68
validity	10		
signature	66		

III. ARQUITECTURA DEL SISTEMA

La arquitectura consta de dos planos: el de gestión y el de difusión, tal como se ilustra en la Figura 1.

Plano de difusión

El plano de difusión contiene dos actores. De un lado, se encuentra el dispositivo en modo baliza. Mediante mensajes de difusión BLE se encuentra difundiendo a tiempo real el estado actual del semáforo. Esto incluye

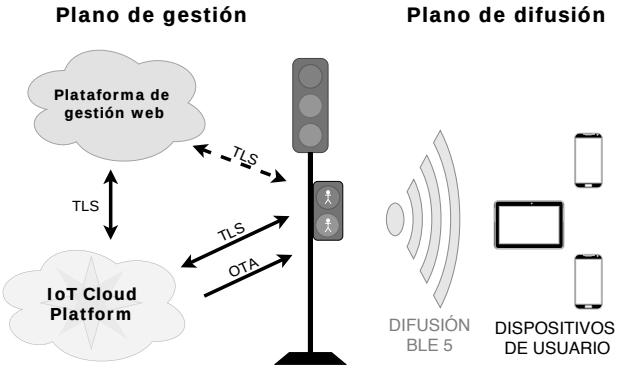


Figura 1: Arquitectura del sistema diseñado

información como el estado actual en el que se encuentra, el tiempo restante para el siguiente cambio de estado o su posición GPS a fin de poder ser ubicado.

Estas tramas son firmadas antes de ser emitidas e incluyen información adicional como longitud de la trama o una marca temporal a fin de asegurar su legitimidad y validez. Por tanto, es el semáforo el encargado de generar la firma.

En el lado de los dispositivos de usuario, estas tramas son recibidas y corresponde al mismo verificar que la firma es correcta. Esto se realiza haciendo uso del certificado que es transmitido en la trama y que corresponde de manera unívoca a cada dispositivo.

Si bien los *certificados* de los dispositivos se envían junto con la trama de difusión, esto limita el tamaño disponible para la carga útil de la trama. Actualmente se están analizando estrategias de distribución más eficientes.

Plano de gestión

Por otro lado, el plano de gestión es el encargado de dotar de lógica al dispositivo. En primer lugar tenemos una plataforma web, encargada de gestionar cada dispositivo de la infraestructura. La plataforma de gestión contiene una base de datos con la ubicación y estado a tiempo real de cada uno de los dispositivos.

La gestión incluye cambios en los tiempos destinados a cada estado del semáforo en función de las características actuales del tráfico, apertura o cierre de semáforos según convenga al paso de un vehículo de emergencia, desactivación o activación de semáforos, etc.

La plataforma web podría comunicarse directamente con los dispositivos de la carretera, pero la tendencia actual sería usar una plataforma IoT intermedia que haga de pasarela entre los semáforos y la plataforma de gestión. Resulta conveniente la utilización de algún estándar de gestión IoT tal como MQTT o CoAP, aunque en nuestro caso particular hemos usado la plataforma de gestión IoT de la empresa *Particle*, que gestiona los dispositivos desplegados y permite aplicar actualizaciones del firmware de los dispositivo si fuera necesario con un mínimo esfuerzo en el desarrollo.

IV. Comunicaciones infraestructura a vehículo

Para el desarrollo de este prototipo se ha optado por BLE frente a las tecnologías tradicionales para c-ITS debido su menor coste y disponibilidad inmediata en un mayor número de dispositivos personales. Dado que por requisitos de diseño la comunicación debe realizarse de punto a multipunto, se ha considerado el uso de mensajes BLE de difusión.

Los mensajes de difusión BLE, *Advertisement*, aparecen en la configuración Bluetooth Low Energy que fue introducida en la versión 4.0 del estándar Bluetooth. Supone una mejora frente a versiones anteriores del protocolo Bluetooth, en las cuales era necesario establecer una conexión punto a punto, para lo cual debía producirse una fase previa de descubrimiento y negociación.

Con la introducción de Bluetooth Low Energy se añadió esta capacidad de transmitir información de difusión, eliminando la necesidad de establecer una conexión previa entre dos dispositivos. Estos paquetes se transmiten en abierto y pueden contener cualquier tipo de información: desde instrucciones para realizar el emparejamiento entre dos dispositivos, hasta información sobre el entorno capturada por los sensores de un dispositivo. La información transmitida puede ser leída por cualquier dispositivo BLE situado en las cercanías y que esté funcionando en modo escáner.

La versión 5.0 del estándar Bluetooth introdujo algunas nuevas características a BLE [8]: mayor tasa de transferencia, alcance más amplio, mayor capacidad de transmisión en difusión e incremento de la coexistencia entre canales de frecuencia. De estas nuevas características, resultan de especial interés para el dispositivo desarrollado las siguientes:

- Alcance amplio o *Long Range*. Para conseguir un mayor alcance de la aplicación sin comprometer el consumo energético, se ha introducido codificación de tipo Forward Error Correction (FEC) en la capa física del estándar Bluetooth. Al utilizar esta característica, la información es codificada antes de ser enviada. Esto se traduce en una mayor sensibilidad a cambio de una tasa de transmisión menor.
- Mensajes de difusión extendidos o *Extended Advertising*. En esta versión de la especificación Bluetooth se ha rediseñado la capa de transporte. Existen tres canales para los mensajes de difusión (canales 37, 38 y 39) que tienen una capacidad máxima de 31 bytes. En esta nueva versión, cuando el mensaje ocupa más de 31 bytes, hasta un límite de 255 bytes, se negocia un canal de datos para transmitir el mensaje, y los datos son transmitidos en el canal elegido.

Si bien no todos los dispositivos móviles inteligentes tienen soporte actualmente para BLE versión 5.0 con estas dos extensiones (se ha confirmado el soporte en Google Pixel 3, Samsung S10+ y OnePlus 6), cabe esperar que el número de móviles con soporte para estas dos extensiones aumente conforme salgan al mercado nuevos terminales.

En el ámbito del presente trabajo solo se considera la emisión de información desde la infraestructura hacia los usuarios. Más concretamente, únicamente se trata la comunicación desde el semáforo inteligente a los usuarios de la vía. En este supuesto, la información facilitada es pública, por lo que no se considera determinante la protección de la privacidad ni la confidencialidad en el canal de difusión.

La solución que se propone consta de una sola trama en la que se incluye tanto la carga útil, como la firma de los datos y un certificado compacto del dispositivo. La estructura de los mensajes se encuentra representada en la Figura 2.

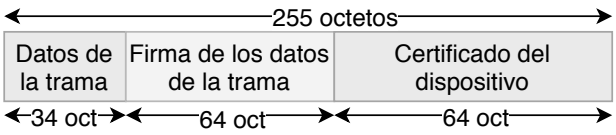


Figura 2: Estructura de mensaje advertising BLE 5

Como se puede ver, la trama consta de tres campos principalmente:

- Datos de la trama: donde se encuentran los datos a transmitir. Aquí se encuentra codificado en hexadecimal los siguientes campos: el identificador del tipo de dipositivo (1 oct.), las coordenadas GPS (8 oct.), la dirección hacia la que el semáforo está apuntando (4 oct.), el rumbo hacia el que la señal lumínica del semáforo afecta (vehículos que van a seguir rectos, vehículos que van a girar a derecha o izquierda, etc. 4 oct.), el estado actual (1 oct.) y el tiempo restante en segundos del estado actual (1 oct.).
- Firma de los datos de la trama: se trata de una firma del campo anterior mediante el algoritmo ECDSA. Se emplea curva elíptica de tipo Secp256k1.
- Certificado del dispositivo: se trata de un certificado digital reducido único para cada dispositivo generado por la CA de la plataforma.

La estructura del certificado reducido diseñado en este trabajo se puede encontrar en la Figura 3.

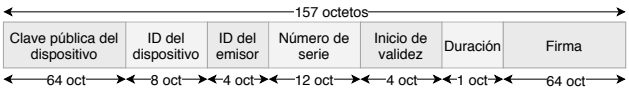


Figura 3: Trama del certificado reducido

Este formato de certificado, al igual que el propuesto en c-ITS, es una alternativa más compacta al estándar X.509 de UIT [10]. Aún así, en nuestro caso este mensaje ocupa 157 octetos frente a los 132 octetos del certificado de los mensajes CAM. Esta diferencia se debe principalmente al tipo de representación elegida para la clave pública, que es representada de forma comprimida (únicamente un punto de la curva y el signo del otro punto) usando 33 octetos en lugar de 64 octetos. En nuestro caso, la SDK criptográfica del nRF52840 utiliza una representación mediante dos puntos, es decir, se necesitan 64 octetos para almacenar

la clave pública. Como puede verse en la Figura 3, el certificado consta de 7 campos:

- Clave pública del dispositivo: Se trata de la clave pública generada por el dispositivo.
- ID del dispositivo: Es un identificador único del dispositivo que está emitiendo los mensajes de difusión. Es asignado por la plataforma de gestión de la infraestructura. En nuestro caso, se ha optado por que el identificador corresponda a las coordenadas GPS del dispositivo, pues no pueden coincidir dos dispositivos en las mismas coordenadas exactas. Las coordenadas GPS se codifican usualmente en 4 bytes.
- ID del emisor: Se trata de un identificador único de la CA que emite el certificado.
- Número de serie: Identifica al certificado en si. Es asignado por la CA.
- Inicio de validez: Es la marca temporal del momento a partir del cual el certificado es válido.
- Duración: Tiempo, medido en días, durante el cual es certificado es válido.
- Firma del certificado reducido: Es la firma digital de todos los campos anteriores.

V. Gestión remota del dispositivo usando LPWAN

LTE Cat M1 es una tecnología de telecomunicación LPWAN (Low Power Wide Area Network) cuyo estándar ha sido desarrollado por el 3GPP (3rd Generation Partnership Project). La especificación para esta tecnología se incluye en el 3GPP Release 13 [4], e incluye las tecnologías LPWAN LTE Cat NB1 y Narrow Band IoT.

Aunque ambas son tecnologías de bajo consumo, orientadas al IoT (Internet of Things), existen diferencias entre ambas [9]. Dada la volatilidad de los datos transmitidos en los ITS, una de las características más valoradas es el menor retardo posible. Es por esto que se ha decidido usar la tecnología LTE Cat M1, con una latencia de 50 a 100ms, frente a NB-IoT, cuya latencia es de varios segundos.

Particle provee servicios de gestión de los dispositivos IoT que fabrica a través de su nube de gestión, simplificando la tarea de gestionar un gran número de dispositivos y sus respectivas comunicaciones.

En nuestro prototipo[2] se usa Particle Cloud [11] como puente entre la plataforma de gestión web y los dispositivos desplegados. La nube de Particle ofrece una API y una conexión segura mediante TLS, para que la plataforma de gestión web acceda a la información de los sensores. Por otro lado, la comunicación entre la infraestructura y los dispositivos se realiza de igual manera mediante TLS.

El esquema de las conexiones entre los elementos hardware del prototipo se encuentra en la Figura 4.

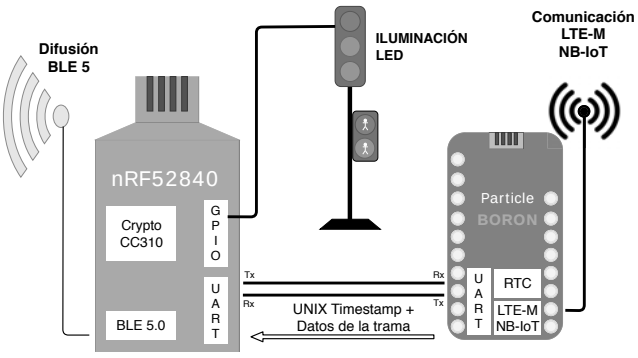


Figura 4: Esquema de conexiones elementos hardware del semáforo

VI. Conclusiones y trabajo futuro

Se puede afirmar que se han cumplido todos los objetivos iniciales del trabajo: el tamaño de trama de los mensajes de difusión BLE extendidos nos permite incluir en ellos la misma información que en el estándar c-ITS, con un nivel de seguridad equivalente. Además, las pruebas de campo nos permiten asegurar que la sobrecarga del sistema, gracias al uso del acelerador hardware criptográfico, es mínima.

Con respecto a la solución propuesta por Audi, el presente trabajo aporta una solución en la que las comunicaciones son realizadas directamente V2X (Vehicle-to-everything), sin la necesidad de una infraestructura de comunicación entre ambos.

La posibilidad de gestión remota segura del semáforo abre las puertas a nuevas aplicaciones que no requieran de un despliegue de fibra dedicado para la interconexión de estos con los centros de control.

Queda pendiente el diseño de un mecanismo para la distribución eficiente de los certificados de dispositivos. Algunas posibilidades que se barajan son:

- Puntos de distribución en las vías: A lo largo del territorio se sitúan balizas encargadas de difundir los certificados de los dispositivos de la zona.
- Distribuidos por el mismo dispositivo: El dispositivo se encarga de transmitir mensajes de difusión con los certificados, intercalados con los mensajes con información del estado.

[2]https://github.com/nicslabdev/MOTAM-nRF52_Beacons

Referencias

[1] El servicio de información de semáforos Audi Traffic Light Information llega a Europa. Fuente: http://prensa.audi.es/ 30/05/2019.
[2] Android Bluetooth Conectivity. Android Open Source Project. https://source.android.com/devices/bluetooth.
[3] ETSI TR 101 607 Technical Report. ETSI.
[4] Release 13 of 3GPP standard. https://www.3gpp.org/release-13.
[5] "IEEE Std. 802.11p-2010, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", IEEE Std 802.11, 2010.
[6] ETSI EN 302 637-2 European Standard. ETSI.
[7] ETSI TS 103 097 V1.2.1 Technical Specification. ETSI.
[8] Bluetooth 5 CS - Core Specification. Bluetooth SIG. 12/2016
[9] Differences between NB-IoT and LTE-M. Accent Systems. https://accent-systems.com/blog/differences-nb-iot-lte-m/.
[10] Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. IETF.
[11] Particle Device Cloud API. Particle Docs. https://docs.particle.io/reference/device-cloud/api/

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Mecanismo para la evaluación de la sincronización en la presentación de contenidos multi-pantalla

Dani Marfil, Fernando Boronat, Jair López, Almanzor Sapena.
Departamento de Comunicaciones,
Universitat Politècnica de València, Campus de Gandia
C/Paraninf 1, 46730, Grao de Gandia, Valencia (Spain)
{damarre@dcom, fboronat@dcom, jailogu@epsg, alsapie@mat}.upv.es

***Resumen*- El consumo simultáneo de contenidos relacionados entre sí en un mismo o en múltiples dispositivos (es decir, escenarios multi-pantalla) es una situación habitual en la actualidad. Para que la experiencia de usuario sea satisfactoria en aplicaciones multi-pantalla, es necesario establecer mecanismos de sincronización que permitan que la reproducción de todos los contenidos esté sincronizada. Dichos mecanismos normalmente realizan ciertas estimaciones, en tiempo de ejecución, de la latencia de reproducción. Por lo que la precisión del grado de sincronización medida por los mismos es difícil de evaluar. En este trabajo, se expone un método para evaluar dicha precisión utilizando técnicas de visión artificial. Así, se puede evaluar, en el extremo de presentación del contenido, la precisión obtenida con los mecanismos de sincronización implementados en dichos sistemas o aplicaciones. Adicionalmente, se han realizado pruebas experimentales en un caso de uso típico, como es un *videowall*, que demuestran la utilidad del mecanismo propuesto.**

***Palabras Clave*- latencia, multimedia, OCR, sincronización, tratamiento de imagen, visión artificial**

I. INTRODUCCIÓN

Actualmente existe una gran variedad de dispositivos de consumo multimedia, con especificaciones y rendimientos heterogéneos. Además, debido a los diferentes elementos involucrados en la transmisión (*broadcast* y/o *broadband*) de los contenidos, el retardo que se acumula a lo largo de la cadena de transmisión varía según el tipo de tecnología utilizada. Por ello, con el fin de que el usuario final pueda consumir múltiples contenidos (relacionados) de forma simultánea y sincronizada, deben establecerse mecanismos que permitan ajustar los procesos de reproducción en cada uno de los dispositivos involucrados, para que, de esta manera, presenten al usuario, de forma sincronizada, fotogramas que hayan sido grabados o generados en el mismo instante de tiempo, bien por un mismo dispositivo de captura o bien por varios dispositivos.

En [1] se presentan los resultados de un estudio centrado en las preferencias, hábitos de consumo y las expectativas de más de 1000 usuarios españoles con relación a servicios de TV híbridos (con consumo de contenidos recibidos por redes broadcast –DVB-T- y broadband –redes IP-). En dicho estudio se identificó el problema de la sincronización como uno de los más importantes a la hora de proporcionar una calidad de experiencia satisfactoria a los usuarios finales.

En [2] se definen los diferentes tipos de sincronización existentes. La sincronización intra-flujo facilita que, dentro del mismo flujo, la información discurra de manera ordenada (p.ej., presentar ordenadamente y equiespaciados en el tiempo los fotogramas en un flujo de vídeo, para una visualización coherente). La sincronización inter-flujo facilita que en un contenido los diferentes flujos involucrados estén presentados y asociados correctamente (p.ej., que el audio y la imagen de un contenido estén reproduciéndose en paralelo de forma coherente). La sincronización inter-dispositivo (*Inter DEvice Synchronisation* o IDES), facilita que varios dispositivos independientes (pero físicamente cerca) reproduzcan, de forma simultánea, el mismo contenido o contenidos relacionados. Finalmente, la sincronización inter-destinatario (*Inter Destination Media Synchronisation* o IDMS), facilita que varios dispositivos en destinos separados geográficamente reproduzcan de forma sincronizada el mismo contenido o contenidos relacionados. Generalmente, los mecanismos de sincronización intra- e inter-flujo ya están implementados en cualquier reproductor y suelen funcionar bien. En este trabajo se presenta una herramienta diseñada para calcular la precisión de sincronismo alcanzado en entornos IDES, aunque también serviría para entornos IDMS simulados en laboratorio.

Normalmente, los mecanismos de sincronización son módulos software que realizan cálculos y estimaciones del retardo de reproducción en el momento de la recepción o la decodificación del contenido. En dichos instantes, aún existe una latencia hasta la presentación del contenido en pantalla que, a priori, es desconocida y, por tanto, debe ser estimada por dichos mecanimos (Fig. 1).

Este tipo de mecanismos se ha utilizado por los autores, por ejemplo, en [3]. Aunque los resultados relativos al nivel de sincronismo alcanzado que se exponen en dicho trabajo son satisfactorios (tanto objetiva como subjetivamente), al basarse en estimaciones de la latencia hasta el momento de la presentación de los contenidos en pantalla, se desconoce su precisión exacta.

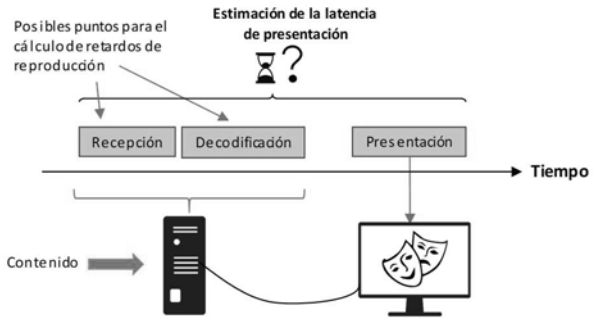


Fig. 1. Puntos en el proceso de reproducción en los que normalmente se calculan los valores de asincronía en soluciones basadas en software con estimación de la latencia de presentación

Es por ello, que se necesitan otros mecanismos de cálculo de asincronías que proporcionen unas medidas más exactas (es decir, no basados en estimaciones) del grado de sincronización obtenido, con el fin de obtener resultados lo más precisos posible. Lo que se persigue con el mecanismo propuesto en este artículo, es poder realizar medidas del nivel de sincronismo alcanzado en el momento de la visualización del contenido en la pantalla, es decir, en el instante de la presentación del contenido. Para ello, se elegirá contenido lo más realista posible y se adaptará para que incluya cierta información visual que, a través de herramientas de visión artificial, pueda ser reconocida e interpretada correctamente. A cada fotograma del vídeo se le deberá insertar, para que se visualice de forma superpuesta, información relacionada con el *timing* del vídeo, como, por ejemplo, una cadena de texto incluyendo el número de dicho fotograma dentro de la secuencia de vídeo. De esta manera, es posible calcular el valor de asincronía existente entre las presentaciones de los contenidos en las diferentes pantallas de los dispositivos involucrados, comparando el número de fotograma que se está visualizando en cada uno de ellos.

Esto se puede realizar de forma automatizada utilizando las numerosas técnicas y mecanismos para el tratamiento y procesado de imágenes existentes hoy en día (p.ej., desde filtros hasta técnicas más complejas para reconocer objetos, etc.). En concreto, las técnicas de reconocimiento óptico de caracteres (*Optical Character Recognition*, OCR) son mecanismos que permiten detectar cadenas de caracteres que puedan existir en imágenes o fotografías. Esto permite interpretar el texto de una forma mucho más eficiente, siendo su aplicabilidad muy variada (p.ej., desde transcribir un libro físico de una manera rápida y automatizada hasta detectar y almacenar la matrícula de un vehículo que ha excedido los límites de velocidad).

En este artículo, se presenta un mecanismo con las siguientes características:

- Realiza el cálculo en instantes de reproducción (más fiable) y no en instantes anteriores como los métodos tradicionales (basados en estimaciones y, por tanto, menos fiables). Permite realizar un cálculo más preciso y fiable del grado de asincronía alcanzado entre dispositivos, en aquellos casos en los que los mecanismos utilizados no proporcionen una precisión satisfactoria o cuyas estimaciones realizadas no sean fiables.
- Está basado en técnicas de OCR para medir la asincronía en la presentación de contenidos relacionados cuando son mostrados en diferentes pantallas conectadas a uno o varios dispositivos.
- Permite realizar una comparativa con el valor de asincronía que los dispositivos involucrados han estimado que existía en cada momento al realizar los ajustes correspondientes en los procesos de reproducción. Con esto, se puede comprobar (y validar) el nivel de precisión alcanzado con dichas estimaciones y, en caso de que el nivel o precisión alcanzada no sea suficiente, realizar los ajustes correspondientes con el fin de mejorar el mecanismo de sincronización empleado (p.ej., corregir o ajustar los parámetros encargados de estimar la latencia asociada al proceso de reproducción).

El artículo sigue la siguiente estructura: en la Sección II se hace referencia al estado del arte y trabajos existentes relacionados, tanto con técnicas de sincronización o cálculo de latencias en sistemas inter-dispositivo/destinatario a partir del procesado de imágenes, como con técnicas de tratamiento y procesado de imagen OCR. En la Sección III, se presenta el mecanismo basado en técnicas OCR propuesto para el cálculo de precisión de la sincronización alcanzada entre dispositivos. En la Sección IV, se presenta un caso de uso para la valoración y validación del mecanismo de sincronización utilizado en un sistema propio de *videowall*. Finalmente, en la Sección V, se exponen las conclusiones sobre la utilidad del mecanismo propuesto y los resultados obtenidos, así como posible trabajo relacionado a realizar en el futuro.

II. Estado del Arte

En esta sección se resumirán los principales trabajos existentes relacionados, incluyendo mecanismos de medida de la sincronización o retardos entre dispositivos con herramientas de procesado de imagen. Además, se presentarán algunos trabajos o soluciones que hacen uso de técnicas de visión artificial para interpretar y analizar cadenas de texto.

A. Mecanismos de evaluación de sincronización basados en el procesado de imágenes

Respecto a la adopción de técnicas de procesado de imagen en este campo, en [4] se propone una solución para el cálculo de retardos entre los diferentes dispositivos involucrados en una videoconferencia. Dicha solución consiste en la generación y análisis de imágenes detectables por ordenador, concretamente códigos QR con información temporal. Estas imágenes son analizadas por otros equipos con dispositivos de entrada de vídeo para poder calcular el retardo existente en la

videoconferencia. El procedimiento para calcular dicho retardo puede resultar incómodo, pues para poder medirlo, los extremos (es decir, los participantes en la videoconferencia), deben apuntar con su webcam a la pantalla de sus equipos, que es donde estarán visualizándose los códigos QR (el del propio usuario y el del usuario remoto). De esta forma, el mecanismo que se propone en dicho trabajo puede detectar ambos códigos y obtener la diferencia de tiempos entre ambos. El prototipo que se implementa en dicho artículo sólo es compatible con MacOSX y, por consiguiente, no es multiplataforma. Además, requiere de una calibración previa para eliminar el retardo inducido por el propio procesado de imagen que realiza dicho mecanismo.

En [5], se presenta un mecanismo para evaluar, a través del análisis de imágenes y de audio, el nivel de sincronismo alcanzado utilizando el estándar DVB-CSS (Digital Video Broadcasting - Companion Screens and Streams [6]). Para tal fin, en ese trabajo se proporciona un vídeo que incluye ciertos pitidos y flashes en instantes conocidos. Este tipo de eventos audiovisuales son recogidos por un microcontrolador Arduino Due[1], el cual está conectado por USB a un PC con el rol de TV o de pantalla complementaria (del inglés, *Companion Screen*). El cálculo de los retardos (asincronías) se calcula obteniendo la diferencia entre los instantes de tiempo en los que se espera que haya pitidos o flashes y los instantes de tiempo en los que ocurren dichos eventos.

Por otro lado, en [7] se presenta un mecanismo para el cálculo de retardos de vídeo extremo-a-extremo. Dicho mecanismo consiste en la inserción en cada fotograma de una marca temporal codificada en formato de código de barras. En ese trabajo se utiliza el *framework* GStreamer [8], por lo que la generación y posterior detección de las marcas de tiempo insertadas se lleva a cabo dentro del propio proceso de reproducción, a través de un elemento de GStreamer implementado para dicha finalidad (en concreto, el elemento *videodetect*[2]).

B. Técnicas de reconocimiento de caracteres OCR

Las técnicas OCR de reconocimiento de caracteres llevan investigándose desde hace décadas. Ya en 1990, en [9] se recopiló una gran variedad de mecanismos existentes para el reconocimiento de diferentes formatos de caracteres (p.ej., para una o más fuentes de texto específicas, para texto escrito a mano, etc.). De acuerdo con dicho trabajo, según la manera de analizar la imagen, las técnicas OCR pueden clasificarse en dos: 1) a través de la utilización de plantillas, en las que el texto a analizar se compara con unos prototipos de caracteres previamente almacenados; y 2) a través del análisis de los parámetros del carácter a reconocer y técnicas de emparejamiento (*matching techniques*). Cabe resaltar que la segunda opción es la más utilizada y consiste, principalmente, en la extracción de parámetros significativos del carácter analizado y su posterior comparación con parámetros de caracteres ideales. Tras esta comparación, se asume que el carácter ha sido reconocido cuando sus parámetros son muy similares a los de uno de los caracteres ideales. El grado de similitud obtenido proporciona el nivel de confianza con el que se ha interpretado el carácter.

Actualmente, también se emplean técnicas más avanzadas para el reconocimiento de texto. Como ejemplo, el trabajo en [10] describe el reconocimiento de escritos a mano mediante el uso de redes neuronales. En dicho trabajo, los caracteres se redimensionan en áreas de 60x40 píxeles y son introducidos en la red neuronal. Cabe destacar que, en dicho trabajo, la red neuronal ha sido entrenada para el alfabeto y lenguaje inglés con más de 19.000 muestras, alcanzando una precisión del 95,69%.

Para el reconocimiento de caracteres, existen numerosas librerías (muchas de carácter *open-source*) que pueden implementarse en diferentes lenguajes de programación, tales como, por ejemplo, python[3] o javascript (nodejs[4]), así como en otros entornos como Matlab[5], que es una herramienta de cómputo numérico con un lenguaje de programación propio.

III. MECANISMO PARA EL CÁLCULO DEL SINCRONISMO ALCANZADO EN ENTORNOS MULTI-DISPOSITIVO

En esta Sección, se presenta un mecanismo no intrusivo para el cálculo, mediante técnicas de visión artificial, del nivel de sincronización adquirido en el instante de presentación de varios dispositivos que deben estar reproduciendo el mismo contenido, o bien contenidos relacionados, de forma sincronizada.

Este mecanismo es capaz de obtener, a partir de un dispositivo de entrada de vídeo (p.ej., una webcam), el número de fotograma que está siendo presentado en cada uno de los dispositivos involucrados. Esta información que se analiza a partir de imagenes obtenidas mediante una cámara de vídeo, junto con la información relativa a la tasa de fotogramas por segundo (fps) del contenido, permite calcular el valor de la asincronía máxima existente entre los dispositivos.

Con esta información, se puede calcular el nivel de asincronía entre los *N* dispositivos involucrados de la siguiente manera (Ec. 1): dado un instante de tiempo *t* (en segundos), la asincronía máxima para dicho instante será la diferencia entre los números de fotogramas máximo y mínimo que se estén reproduciendo en dicho instante en los dispositivos ($\max(n_{trama})$, $\min(n_{trama})$, respectivamente) multiplicada por la duración de un fotograma de dicho contenido ($\frac{1}{fps}$, siendo *fps* la tasa del vídeo en fotogramas por segundo).

$$Asincronía_{max}(t) = (\max(n_{trama}, t) - \min(n_{trama}, t)) * \frac{1}{fps} \ [segundos] \quad (1)$$

Cabe señalar que existe un margen de error en el resultado que se obtiene a partir de la Ec. 1 debido a que no se pueden medir asincronías con un valor menor que la correspondiente a la duración de un fotograma. Dicho error, denominado Error de Precisión (EP), se puede calcular (en segundos) a partir de la Ec. 2 de la siguiente forma:

[1] https://store.arduino.cc/due
[2] https://www.freedesktop.org/software/gstreamer-sdk/data/docs/2012.5/gst-plugins-bad-plugins-0.10/gst-plugins-bad-plugins-videodetect.html
[3] https://pypi.org/project/pytesseract/
[4] https://www.npmjs.com/package/ocr
[5] https://www.mathworks.com/help/vision/optical-character-recognition-ocr.html

$$EP = \pm \frac{1}{fps} \ [segundos] \quad (2)$$

Por tanto, si tras realizarse los cálculos pertinentes, se obtiene que la diferencia entre número de fotogramas de los dispositivos respecto a la referencia es de *n* fotogramas, esto implicaría un nivel de asincronía (en segundos) cuyo valor se encuentra dentro del intervalo definido en la Ec. 3:

$$Asincronía \in \left\{\frac{n}{fps} - EP, \frac{n}{fps} + EP\right\} [segundos] \quad (3)$$

A pesar del EP, el mecanismo puede considerarse suficiente, al ser éste una manera de evaluar la precisión de sincronismo alcanzada en los sistemas evaluados. Es decir, de evaluar y validar el comportamiento de las técnicas de sincronización que ya están implementadas en las aplicaciones y los dispositivos objetos de la evaluación.

El mecanismo se puede dividir en dos fases: 1) la preparación del contenido; y 2) la medición de asincronías en tiempos de presentación. Es por esto que, con el fin de emplear el mecanismo propuesto, se debe utilizar un contenido que disponga de marcas o referencias superpuestas. Por tanto, se puede utilizar cualquier tipo de contenido siempre y cuando haya un proceso previo encargado de la inserción de dicha información (la primera fase). Debido a esto, el contenido generado específicamente para el mecanismo propuesto no debería ser utilizado por usuarios, ya que contará con elementos extraños (*artifacts*), como las cadenas de texto superpuestas tapando parte del contenido, que pueden resultar molestos y empeorar la experiencia de consumo.

A. Fase 1: Preparación del contenido

Durante esta fase, el contenido debe prepararse, de forma que tenga elementos identificables por el mecanismo propuesto. Una manera simple de conseguir esto es superponer el número de fotograma asociado a cada fotograma de vídeo. Entre muchas otras, una de las herramientas que lo permite de forma muy sencilla es *ffmpeg*[6]. En la Fig. 2 se muestra el comando *ffmpeg* para insertar el número de fotograma en un vídeo.

```
ffmpeg -i {contenido_original} -filter_complex
"drawtext=
fontfile=/usr/share/fonts/truetype/freefont/FreeSerif.ttf:
text='frame %{n}': x=100: y=50: fontsize=80:
fontcolor=white@1.0: box=1: boxcolor=black@1.0"
{contenido_generado}
```

Fig. 2. Comando utilizado para la inserción del número de fotograma con la herramienta *ffmpeg*.

Donde *{contenido_original}* es la ubicación y nombre del contenido al que se le va a superponer el número de fotograma de vídeo, *{contenido_generado}* es la ubicación y nombre del fichero generado con los números de fotograma superpuestos. Se utiliza el parámetro *filter_complex* para indicar qué se va a superponer en el vídeo. En el ejemplo, se está insertando el texto "frame *n*", siendo *n* el número de fotograma y la palabra *frame* como identificador de que el número que acompañe a dicha cadena de caracteres corresponde con el número de trama, ya que los autores consideran como muy poco probable que un contenido audiovisual vaya a presentar de forma original esta palabra. Además, se utilizan otros parámetros que permiten configurar la posición donde se insertará este texto (variables *x* e *y* en el comando), o el tamaño y color del texto y si debe tener un fondo (variables *fontsize*, *fontcolor*, *box* y *boxcolor*, respectivamente). La Fig. 3 muestra cómo quedaría el resultado final, tras la generación del contenido a partir del comando de la Fig. 2. Se puede apreciar que se ha superpuesto la cadena de texto "frame 936" en color blanco sobre un fondo negro. Dicho formato condicionará los filtros de imagen a aplicar, tal y como se explica más adelante.

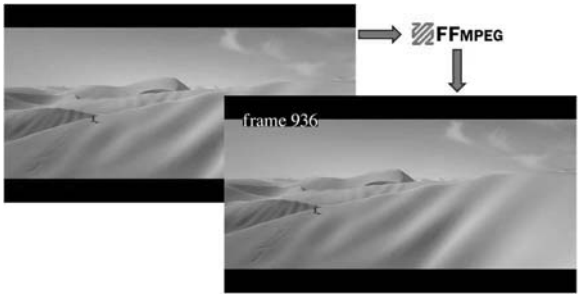


Fig. 3. Contenido con el número de fotograma insertado a través de la herramienta *ffmpeg*.

B. Fase 2: Medición de asincronías en tiempos de presentación

Una vez que el contenido ya cuenta con el número de fotograma superpuesto, se lleva a cabo un calibrado manual previo al inicio de la evaluación del sistema, con el fin de situar el dispositivo de captura de imagen correctamente y determinar la sección a recortar de la imagen obtenida, ya que, dependiendo del dispositivo de captura de imagen empleado, la distancia al sistema a evaluar y la sección a recortar de la imagen obtenida puede variar (esto es, depende directamente de la óptica y la resolución del dispositivo de captura de imagen utilizado). Tras finalizar el calibrado, se iniciará el sistema objeto de evaluación, y, por tanto, la reproducción de dicho contenido en los dispositivos involucrados. A partir de dicho instante ya se puede empezar a capturar vídeo o imágenes de forma periódica, y medir y comparar los valores de los fotogramas que están siendo presentados en las pantallas de cada uno de los dispositivos involucrados.

Dependiendo del rendimiento del dispositivo que esté a cargo del análisis de las imágenes, podrá realizarse bien en tiempo real, o bien almacenando dichas capturas de la entrada de vídeo para su análisis posterior. Por ejemplo, La Fig. 4 muestra un esquema de los posibles dispositivos involucrados para el mecanismo propuesto, donde se puede observar un PC con una entrada de vídeo (una cámara) conectada. Esta cámara registra la imagen de las pantallas involucradas en el sistema para el que se quiere obtener la asincronía máxima. EL PC al que está conectada la cámara se encarga de analizar las imágenes obtenidas y, a continuación, calcular la asincronía máxima entre los *n* dispositivos del sistema.

[6] https://ffmpeg.org/

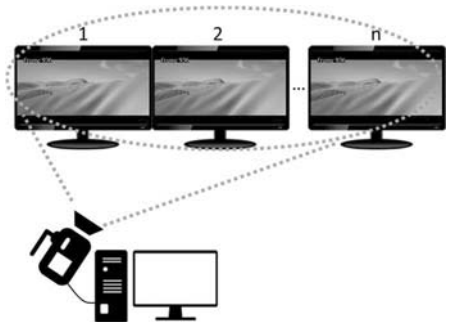


Fig. 4. Esquema de los dispositivos involucrados para el cálculo del nivel de sincronismo alcanzado.

La manera de evaluar el nivel de sincronismo alcanzado es la siguiente: mientras las pantallas del sistema a evaluar están reproduciendo el contenido preparado en la fase 1, el sistema de medida capta, a través de un dispositivo de entrada de vídeo (esto es, una cámara) dichos números y calcula y registra la asincronía máxima en cada instante mediante la Ec. 1.

Con el objetivo de agilizar el procesado de la imagen a evaluar, se realiza un recorte en la imagen captada para almacenar solamente las regiones de interés de cada pantalla en las que estén visibles los números de fotograma. Seguidamente, se aplican una serie de filtros que permiten descartar, de la imagen obtenida, aquellas regiones que no van a aportar información.

Los filtros adoptados para el procesado de imagen, previo al proceso de detección de caracteres, son los típicamente utilizados en técnicas OCR, como el filtro *top-hat*[7] o el filtro de *erosión*[8], los cuales permiten destacar los números de fotogramas superpuestos y ocultar el resto de información de la imagen. Concretamente, el filtro *top-hat* permite resaltar regiones más claras respecto a las oscuras, por lo que, permitirá destacar la zona en la que se encuentra la información del número de fotograma (al ser el texto de color blanco sobre fondo negro). El filtro de *erosión* permite "limpiar" la imagen eliminando elementos diferentes a un tamaño configurado. Esto permite eliminar zonas que puedan dar lugar a falsos positivos, es decir, evitar la detección de caracteres donde no los hay. Tras este último paso, la imagen procesada ya estará preparada para utilizar el mecanismo OCR con el fin de detectar los números de fotograma que existan en la misma. Finalmente, tras obtener la información de los números de fotograma, se podrá hacer uso de la Ec. 1, con la que se obtiene la asincronía máxima existente para cada instante. Como resumen, la Fig. 5 muestra las distintas etapas por las que pasa la imagen antes de detectar correctamente la información insertada.

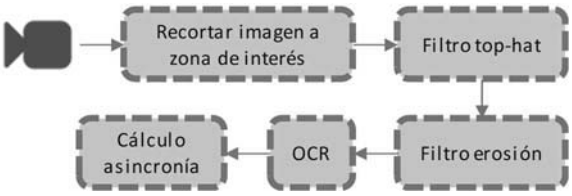


Fig. 5. Distintas etapas del tratamiento de la imagen para detectar la asincronía máxima entre dispositivos de un sistema.

IV. Validación del mecanismo propuesto: Caso de uso

En esta Sección se va a utilizar el mecanismo propuesto para evaluar y validar el mecanismo de sincronización implementado en un caso de uso de un sistema propio de pantallas múltiples, formando un *videowall*, basado en dispositivos Raspberry Pi (en adelante, RPi) y presentado en [11]. En concreto, para esta evaluación se ha configurado un sistema *videowall* de 2x2, es decir, compuesto por 4 pantallas distribuidas en dos filas y dos columnas. El mecanismo de sincronización del *videowall* se basa en el cálculo y estimación de asincronías realizados por el software de reproducción, ejecutándose en las propias RPi, en el momento de la recepción del contenido. En ese momento, se realiza una estimación de la latencia asociada al proceso reproductor desde dichos instantes hasta la presentación del contenido en la pantalla. El sistema propuesto en este artículo nos permitirá, además, comprobar el nivel de precisión alcanzado en el sistema de *videowall* bajo estudio.

Los dispositivos involucrados para obtener las imágenes y calcular la asincronía son un PC (Windows 10, procesador Intel Core i7 6700 @ 3.40GHz, 8GB RAM y HDD de 1 TB) y una webcam Logitech HD C270 (con una resolución de 720p y una tasa de 30fps) situada a unos 2m del sistema a medir.

El contenido utilizado ha sido el tráiler del vídeo Sintel[9], cuya codificación es H.264 + AAC, con una resolución de 1920x1080 y una tasa de 25 fotogramas por segundo (esto es, 1 fotograma cada 40ms). Se ha elegido este vídeo, y no un vídeo más uniforme (p.ej., una carta de ajuste), porque se desea evaluar la sincronización con vídeos que sean realistas, es decir, que sean similares a los que se pueden encontrar en casos reales en un *videowall* (con cambios de escena, movimientos, etc., que puedan afectar al uso de recursos y de capacidad de procesamiento en los dispositivos que ejecutan los procesos de sincronización). Puesto que en el *videowall* se van a visualizar diferentes partes del contenido en cada pantalla, se ha realizado la superposición del número de fotograma en 4 puntos diferentes para que en todo momento se visualice esta información en las 4 pantallas. En la Fig. 6 puede observarse el resultado.

Fig. 6. Contenido preparado para la obtención de la precisión alcanzada en tiempos de presentación.

Para poder comparar los valores medios de asincronía obtenidos a partir del software del propio sistema *videowall*

[7] https://es.mathworks.com/help/images/ref/imtophat.html
[8] https://es.mathworks.com/help/images/ref/imerode.html
[9] https://durian.blender.org/download/

con los que se obtienen a partir del mecanismo propuesto en este artículo, se han llevado a cabo 10 sesiones de aproximadamente 5 minutos, con el objetivo de obtener el valor medio de sincronismo alcanzado según el software del *videowall* (es decir, durante la recepción del contenido y estimando la latencia de los procesos de reproducción).

Tras analizar los valores almacenados por el software del *videowall*, se ha obtenido un valor de asincronía media de 33ms, siendo el tiempo entre fotogramas de 40ms. Por tanto, si dichos valores han sido estimados correctamente, el resultado que se debería obtener a través del mecanismo presentado en este artículo debe ser un valor medio que esté entre 0ms ± $\frac{1}{fps}$ ms. Para este caso específico será, un valor entre 0ms y ±40ms (Ec. 2 y 3). Ello implica que, durante el análisis con el mecanismo propuesto, debería obtenerse de promedio hasta un fotograma de diferencia entre los dispositivos involucrados. Se han realizado las capturas de imágenes durante la presentación de contenidos y se ha utilizado la herramienta Matlab para el procesado de dichas imágenes y el cálculo de las asincronías entre dispositivos según la Ec. 1.

Las Fig. 7 y 8 muestran el detalle de las imágenes tomadas por el dispositivo de entrada de vídeo del *videowall* 2x2 en funcionamiento con el contenido con fotogramas insertados (Fig. 7) y el resultado de procesado previo a la detección del número de fotograma de la imagen capturada por el dispositivo de entrada de vídeo (Fig. 8).

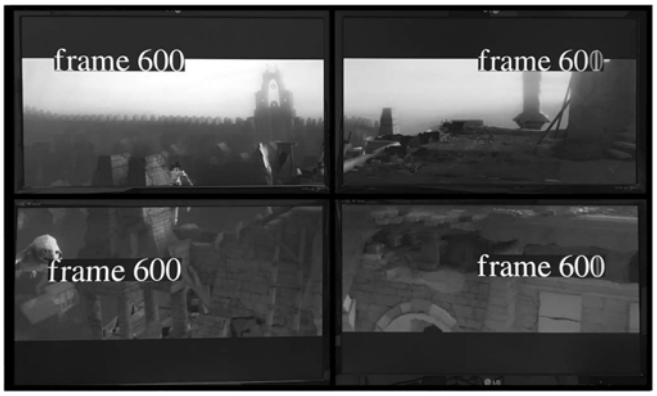


Fig. 7. Detalle del sistema *videowall* 2x2 reproduciendo el contenido con el número de fotograma insertado.

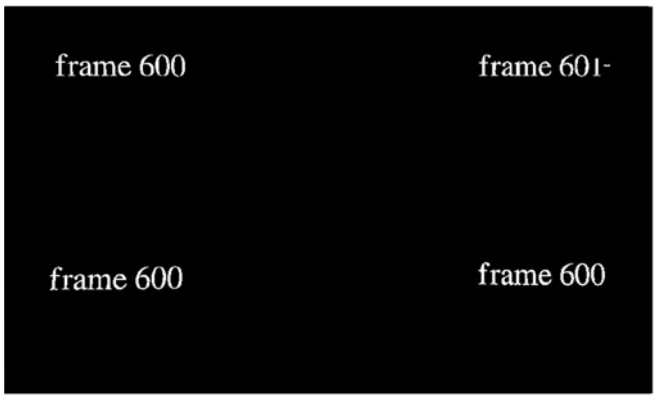


Fig. 8. Detalle de la imagen procesada previa a la etapa de detección de caracteres[10].

Para comprobar el funcionamiento correcto del mecanismo, se han tomado aproximadamente 100 muestras (imágenes), a razón de 1 muestra por segundo, de las cuales alrededor de 10 han sido descartadas al no haberse obtenido de forma clara e inequívoca el valor del número de fotograma tras el análisis de la imagen.

Numéricamente, el error cuadrático medio[11] (ECM) obtenido de la asincronía máxima ha sido de 0.3023 fotogramas^2. Por tanto, se puede obtener su raíz (RECM) y, a continuación, convertir el resultado en unidades de tiempo (Ec. 4):

$$RECM = \sqrt{ECM} * \frac{1}{fps} = 0.022s \quad (4)$$

Como puede observarse, el resultado numérico corrobora que los resultados son similares a los que los dispositivos involucrados en el *videowall* han calculado en tiempos de recepción, es decir, son coherentes con el nivel de sincronismo que se estima en tiempos de recepción por parte de los procesos de reproducción en los dispositivos involucrados (RPi).

Por lo tanto, el funcionamiento del sistema puede considerarse como válido y satisfactorio, puesto que el valor obtenido de 22ms está dentro de los niveles esperados, con una diferencia respecto a los valores obtenidos por el propio sistema *videowall* de 11ms, la cual puede considerarse despreciable.

En definitiva, utilizando el sistema propuesto en este artículo se han podido validar de forma objetiva las estimaciones realizadas por las técnicas de sincronización implementadas en el software del sistema de *videowall* presentado en [11].

V. Conclusiones

En este artículo se ha presentado un mecanismo que permite medir de forma objetiva el nivel de sincronismo alcanzado en tiempos de presentación entre dos o más dispositivos que se supone que deben estar reproduciendo contenidos relacionados de forma sincronizada. Este sistema, además, es útil para validar técnicas de sincronización

[10] Se puede observar cómo en este ejemplo, en la pantalla superior izquierda se acaba detectando el número 601 frente al 600. Esta decisión es criterio de la herramienta utilizada de Matlab, cuya mejora no se aborda en este artículo.

[11] Se toman valores cuadráticos para evitar que valores negativos y positivos se compensen entre sí a la hora de calcular valores medios.

implementadas en aplicaciones que requieran del consumo de contenidos relacionados en una o más pantallas (escenario multi-pantalla). Debido a que, normalmente, dichas técnicas se basan en medidas llevadas a cabo durante la recepción o decodificación del contenido, y no durante la presentación del mismo, deben realizar estimaciones de la latencia asociada al proceso de reproducción hasta la visualización final del contenido en la pantalla. Con el mecanismo que se presenta en este artículo se puede medir de forma objetiva y más fidedigna el grado de sincronización adquirido en dichas aplicaciones y, de forma indirecta, validar los cálculos y estimaciones realizados en las técnicas de sincronización que se incluyen en ellas. Como demostración de la utilidad del sistema propuesto, se ha empleado para validar la técnica de sincronización empleada en un sistema de *videowall* basado en dispositivos de bajo coste (RPis) desarrollado en el propio grupo de investigación. Se ha corroborado que los valores de sincronización medidos por el propio software de los dispositivos involucrados, a partir de las estimaciones realizadas por la solución de sincronización implementada, y el valor obtenido por el sistema propuesto son similares. Como trabajo futuro, se pretende actualizar y optimizar el mecanismo de detección de caracteres para que el número de imágenes descartadas sea el menor posible.

AGRADECIMIENTOS

Este trabajo ha sido financiado, parcialmente, por la Generalitat Valenciana, bajo el programa de Subvenciones para Grupos de Investigación Consolidables, AICO/2017, con referencia AICO/2017/059.

REFERENCIAS

[1] F. Boronat, M. Montagud, D. Marfil, and C. Luzon, "Hybrid Broadcast/Broadband TV Services and Media Synchronization: Demands, Preferences and Expectations of Spanish Consumers," *IEEE Trans. Broadcast.*, vol. 64, no. 1, 2018.

[2] M. Montagud, F. Boronat, H. Stokking, and P. Cesar, "Design, development and assessment of control schemes for IDMS in a standardized RTCP-based solution," *Comput. Networks*, vol. 70, pp. 240–259, Sep. 2014.

[3] F. Boronat, D. Marfil, M. Montagud, and J. Pastor, "HbbTV-Compliant Platform for Hybrid Media Delivery and Synchronization on Single-and Multi-Device Scenarios," *IEEE Trans. Broadcast.*, vol. 64, no. 3, 2018.

[4] J. Jansen, "VideoLat: An Extensible Tool for Multimedia Delay Measurements," in *Proceedings of the ACM International Conference on Multimedia - MM '14*, 2014, pp. 683–686.

[5] M. Hammond, J. Kramskoy, and British Broadcasting Corporation, "Measuring synchronisation timing accuracy for DVB Compainion Screen Synchronisation TVs and Companions," 2015. [Online]. Available: https://github.com/bbc/dvbcss-synctiming. [Accessed: 21-May-2019].

[6] Digital Video Broadcasting, "ETSI TS 106 286-1. Companion Screens and Streams; Part 2: Content Identification and Media Synchronization." .

[7] M. A. Montagud Climent, F. Boronat, and P. S. César Garcia, "A customizable open-source framework for measuring and equalizing e2e delays in shared video watching," in *ACM TVX*, 2014, pp. 1–2.

[8] GStreamer, "GStreamer Framework." [Online]. Available: https://gstreamer.freedesktop.org/.

[9] V. K. Govindan and A. P. Shivaprasad, "Character recognition - A review," *Pattern Recognit.*, vol. 23, no. 7, pp. 671–683, Jan. 1990.

[10] A. Yousaf *et al.*, "Size invariant handwritten character recognition using single layer feedforward backpropagation neural networks," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies, iCoMET 2019*, 2019, pp. 1–7.

[11] P. Salvador, F. Boronat, M. Montagud, and D. Marfil, "Sistema videowall de bajo coste basado en Raspberry Pi, personalizable y configurable dinámica y remotamente vía Web," in *XIII Jornadas de Ingenieria Telematica - JITEL2017*, 2017, pp. 318–325.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Estimación del tiempo de respuesta de *proxies* web en redes comunitarias utilizando algoritmos de factorización matricial

Diego Bores Quijano*, Miguel L. Bote Lorenzo*, Eduardo Gómez Sánchez*, Roc Meseguer Pallarés**
*Departamento de Teoría de la Señal, Comunicaciones e Ingeniería Telemática, Universidad de Valladolid
ETSI de Telecomunicación, Paseo de Belén 15, 47011 Valladolid
**Departamento de Arquitectura de Computadores, Universitat Politècnica de Catalunya
Campus Nord - Edificios D6/C6, C. Jordi Girona 1-3, 08034 Barcelona
diego@gsic.uva.es, migbot@tel.uva.es, edugom@tel.uva.es, meseguer@ac.upc.edu

***Resumen*—En este artículo se evalúa la precisión de 3 algoritmos de factorización matricial para estimar el tiempo de respuesta de los *proxies* web en redes comunitarias a partir medidas de dicho indicador de calidad de servicio que son tomadas por los clientes y compartidas entre ellos. Para ello se ha llevado a cabo una serie de experimentos utilizando los datos obtenidos de la emulación de una red comunitaria formada por 8 clientes y 5 *proxies*. Los resultados muestran que 2 de los algoritmos consiguen hacer las estimaciones con una elevada precisión, siendo el mejor el algoritmo de factorización matricial adaptativa, incluso cuando el número de *proxies* sondeados por cada cliente para obtener las medidas de tiempo de respuesta es bajo.**

***Palabras Clave*—redes comunitarias, *proxy* web, calidad de servicio, algoritmos de factorización matricial**

I. INTRODUCCIÓN

Las redes comunitarias [1] son redes distribuidas, descentralizadas y a gran escala. Estas redes permiten que las comunidades locales creen una infraestructura de red propia, y ofrecen a sus usuarios conexión con Internet de forma asequible, incluso en áreas rurales. En Europa hay varias redes comunitarias de gran tamaño, como es el caso de la Ninux [2] y Funkfeuer [3], desplegadas en Italia y Austria respectivamente. En España, la red comunitaria Guifi.net [4] cuenta con más de 30.000 nodos situados principalmente en Cataluña.

El acceso a la Web en las redes comunitarias se hace típicamente a través de *proxies*, siendo los usuarios los que eligen su *proxy* preferido de entre todos los disponibles en la red [5]. Esto implica que, para acceder a un servidor web externo, suele ser necesario recorrer una ruta dentro de la red comunitaria que atraviesa varios enlaces y nodos hasta llegar al *proxy* seleccionado. Dicho *proxy* será el encargado de realizar la petición al servidor externo en nombre del usuario.

Del hecho de que sean los usuarios quienes eligen el *proxy* que desean utilizar surge la posibilidad de congestión en estos nodos de salida [5]. Por lo tanto, podría ocurrir que mientras unos *proxies* estuviesen subocupados, otros acabaran saturados por la demanda, dando lugar a una degradación del rendimiento del acceso a la Web. Así, la elección del *proxy* por parte de los usuarios puede afectar a la calidad de servicio que éstos perciben.

Los mecanismos de selección automática de *proxies* web en redes comunitarias que es posible encontrar en la literatura como [5], [6] requieren que, cada cierto tiempo, los clientes lleven a cabo sondeos de un número determinado de *proxies*. Esto permite a los clientes obtener medidas indicativas de la calidad de servicio que cabe esperar de cada uno de los *proxies*. Algunos ejemplos de estas medidas son la latencia y el tiempo de respuesta [6].

Cuanto mayor es el número de *proxies* web de los cuales se dispone de medidas indicativas de calidad de servicio, mayores son las posibilidades del cliente de seleccionar uno adecuado. Sin embargo, el sondeo de un elevado número de *proxies* por parte de los clientes puede suponer una elevada carga tanto para unos como para otros, además de un incremento notable de tráfico en la red comunitaria [6]. Para evitar este problema, cabe la posibilidad de que cada cliente sondee un número reducido de *proxies* y comparta con otros clientes las medidas indicativas de calidad de servicio que obtenga. Los clientes que reciben dichas medidas pueden utilizarlas para estimar las medidas que recabarían si ellos mismos sondearan los *proxies* correspondientes. Con las medidas obtenidas de sus propios sondeos y las estimaciones realizadas a partir de los sondeos hechos por otros clientes, cada cliente tiene la posibilidad de elegir entre un número mayor de *proxies* sin necesidad de realizar sondeos adicionales.

Para que esta aproximación para la selección de *proxies*

web pueda ser efectiva, es imprescindible que las estimaciones que hagan los clientes sean precisas. En trabajos de la literatura como [7], [8], [9] se ha comprobado que los algoritmos de factorización matricial pueden obtener estimaciones precisas de medidas de calidad de servicio en problemas semejantes al planteado en este artículo. Por ejemplo, algunos de estos algoritmos se estudiaron en [9] para la estimación tanto del tiempo de respuesta como de la tasa de transferencia ofrecidos por servicios web a partir de las medidas obtenidas mediante sondeos de los clientes y que después eran compartidas entre ellos. Hasta donde saben los autores, estos algoritmos no han sido estudiados para la estimación de medidas de calidad de servicio en el contexto de las redes comunitarias.

El objetivo de este artículo es estudiar la posibilidad de utilizar algoritmos de factorización matricial para estimar el tiempo de respuesta de *proxies* web en redes comunitarias. Para ello se evalúa la precisión de 3 algoritmos de factorización matricial bien conocidos cuando estos se utilizan para llevar a cabo la estimación de los tiempos de respuesta de un conjunto de 5 *proxies* web a partir de los sondeos realizados cada 10 segundos durante 2 días por 8 clientes.

La estructura del resto del artículo es la siguiente. En la sección II se discute el trabajo relacionado que es posible encontrar en la literatura. A continuación, la sección III presenta el marco experimental que ha sido utilizado en nuestro estudio con el objetivo de facilitar su reproducibilidad. La sección IV presenta y analiza los resultados de los experimentos. Finalmente, en la sección V se recogen las principales conclusiones que es posible obtener de este estudio y se avanzan las principales líneas de trabajo futuro.

II. TRABAJO RELACIONADO

La monitorización siempre ha sido un elemento clave para asegurar el rendimiento de un sistema distribuido complejo, como una red comunitaria. Es un primer paso para controlar la calidad del servicio, detectar anomalías o tomar decisiones sobre la asignación de recursos. En el caso de los *proxies* web de una red comunitaria esta monitorización es especialmente útil para la selección de un determinado *proxy* por parte de un cliente. Los clientes pueden seleccionar el *proxy* adecuado de acuerdo con alguna métrica de calidad de servicio. Además, esto está relacionado con el problema que supone que un gran número de clientes puedan navegar por la Web aprovechando la capacidad agregada de un número reducido de *proxies*.

La visibilidad completa entre clientes y *proxies* es una primera aproximación a esta monitorización de soporte para la selección. Conceptualmente es una matriz de todos los *proxies* medidos por todos los clientes con una métrica de calidad de servicio dada, como puede ser el tiempo de respuesta. Un ejemplo típico es la basada en la infraestructura de una red de distribución de contenidos como la de Google [10]. Aunque estas medidas no son sencillas de obtener ni precisas sin implicar a *proxies*, infraestructura y especialmente a los clientes [11]. Esta visibilidad completa también se puede conseguir de forma activa por parte de los clientes. Una primera aproximación es la fuerza bruta, donde el cliente en solitario sondea activamente todos los *proxies* [12]. Aunque esta solución puede ser útil para una red pequeña no es escalable a una red comunitaria y tiene un gran coste en sondeos a los *proxies*.

Para conseguir escalabilidad y reducir el coste de la monitorización, una aproximación típica es reducir la cantidad de *proxies*: probando solo los más cercanos [6] o seleccionando algunos aleatoriamente [13]. Aunque con estas aproximaciones tenemos una visión muy reducida, perdiendo la visibilidad completa.

La visibilidad completa también se puede conseguir cooperando entre los diferentes clientes y compartiendo las visiones parciales de cada uno. Una estrategia típica son las coordenadas virtuales. Un ejemplo típico es Vivaldi [14], que es un sistema que permite la estimación de latencias entre todos los nodos que participan en él. En [15] se presenta un sistema basado en Vivaldi que permite estimar latencias entre un nodo del sistema, por ejemplo un cliente, y otro externo, por ejemplo un *proxy*. Esta aproximación a la visibilidad completa tiene dos problemas. El primero, las medidas que Vivaldi necesita son algo costosas por las continuas pruebas activas por parte de los clientes y no son precisas [11], [16]. El segundo problema es la métrica usada; Vivaldi usa la latencia de red entre nodos. En el caso concreto de la selección de *proxy* la latencia no refleja algunos aspectos importantes, como la carga del propio *proxy*, y la carga de la conexión a Internet del *proxy* [6]. En [5] se propone un sistema basado en Vivaldi donde los clientes comparten tanto los nuevo valores de latencia entre los clientes, como la latencia de los *proxies* probados.

Hay otras estrategias de cooperación no basadas en coordenadas virtuales como la basada en reutilizar los paquetes de sondeo. En [17], los clientes intermedios capturan los sondeos de otro cliente a un *proxy* para calcular su propia información sin hacer un sondeo por su cuenta. Esta aproximación presenta un importante problema: solo se pueden reutilizar los sondeos de otros clientes que pasan por el propio cliente. En las redes comunitarias eso no es lo habitual; los clientes están en los extremos de la topología y no actúan como encaminadores, por lo que no pasan peticiones por ellos [4].

Otra estrategia de cooperación se basa en el sondeo aleatorio de algunos *proxies*. En [18], los clientes sondean el último *proxy* usado y aleatoriamente dos más. Los clientes comparten estos valores solo con los clientes más cercanos topológicamente. El problema de esta estrategia es que no garantiza la obtención de los valores de todas las celdas de la matriz de visibilidad correspondientes a las combinaciones de clientes y *proxies*. Este caso usa la idea de la semejanza entre los valores de sondeo del mismo *proxy* de clientes topológicamente cercanos para rellenar las celdas vacías. Concretamente, se rellenan directamente con los valores medidos por los clientes más cercanos.

Para evitar estas celdas vacías en la matriz de visibilidad se pueden usar técnicas de estimación de estos valores.

En [19] se usan técnicas de aprendizaje profundo (*deep learning*) para realizar obtenerlas. El problema de esta aproximación es el coste computacional de los algoritmos de aprendizaje profundo, por lo que tienen que ejecutarse en equipos dedicados con el problema añadido de la recolección y transferencia a este equipo de la información necesaria para la predicción.

El filtrado colaborativo es una aproximación para la estimación habitualmente mucho menos costosa computacionalmente, lo cual comporta la ventaja de poder ejecutarse en los propios clientes y encaminadores de una red comunitaria. En [20] se propone el uso de un algoritmo de filtrado colaborativo para la recomendación o la selección de servicios. Se basa en el coeficiente de correlación de Spearman entre los vectores de los valores de los sondeos de los clientes. Adicionalmente se reduce el conjunto de clientes usando solo el subconjunto de clientes más cercanos en distancia. Por otra parte, en [21] se trata de obtener los valores desconocidos de la matriz mediante un mecanismo de filtrado colaborativo basado el algoritmo de *Kernel Least Mean Square (KLMS)*. El problema de esta propuesta es también el elevado coste computación. Otro ejemplo es el sistema de recomendación basado en filtrado colaborativo propuesto en [22]. En este caso los valores de la matriz no son solo medias de calidad de servicio, también utiliza las preferencias de los usuarios, como por ejemplo el precio.

La factorización matricial es también una técnica de filtrado colaborativo que ha sido aplicada con éxito para la estimación de los valores desconocidos de la matriz de visibilidad. En [7] se propone un algoritmo de factorización extendida matricial, en el que usan diferentes técnicas de cálculo de similitud tanto entre clientes como entre servicios para obtener un conjunto de clientes y servicios cercanos. La estimación se basa en la idea de una percepción similar de servicios cercanos por parte de clientes cercanos. En [8] se hace una propuesta similar pero en este caso fusionan información global con información cercana. El valor a estimar depende de el conjunto de una información del usuario, la información de los vecinos del usuario, y alguna información global de la matriz. En [9] se propone un algoritmo de factorización matricial adaptativa y en línea, que además tiene en cuenta la evolución temporal de la métrica de rendimiento del servicio. Esta propuesta aporta dos contribuciones interesantes: en linea, con lo que puede ejecutarse en tiempo real en los clientes, y la evolución temporal, con lo que estima siguiendo la tendencia temporal. Hasta donde saben los autores, en la literatura no se ha estudiado la estimación de los valores de la matriz de visibilidad con algoritmos de factorización matricial para llevar a cabo la monitorización y selección de *proxies* en el contexto de una red comunitaria con las restricciones de recursos de este tipo de redes.

III. MARCO EXPERIMENTAL

A. Conjunto de datos

Para la obtención de datos con los que realizar los experimentos, se emuló el funcionamiento de una red comunitaria en la plataforma PlanetLab [23]. Esto se hizo desplegando 8 clientes y 5 *proxies* [24] web en diferentes nodos de PlanetLab distribuidos por Europa y Oriente Próximo tal y como se muestra en la tabla I. Si bien es cierto que algunas redes comunitarias pueden llegar a tener un número de clientes y *proxies* mucho mayor, este conjunto de datos puede considerarse suficiente para estudiar el potencial de los algoritmos de factorización para la estimación del tiempo de respuesta.

Tabla I
DISTRIBUCIÓN DE CLIENTES Y PROXIES EN DISTINTOS NODOS DE LA PLATAFORMA PLANETLAB.

	Nodo
Cliente 1	cse-yellow.cse.chalmers.se
Cliente 2	mars.planetlab.haw-hamburg.de
Cliente 3	pl2.uni-rostock.de
Cliente 4	planet3.cs.huji.ac.il
Cliente 5	planetlab3.di.unito.it
Cliente 6	planetlab-1.ing.unimo.it
Cliente 7	planetlab-2.cs.ucy.ac.cy
Cliente 8	ple43.planet-lab.eu
Proxy 1	ple3.planet-lab.eu
Proxy 2	planet4.cs.huji.ac.il
Proxy 3	cse-white.cse.chalmers.se
Proxy 4	planetlab1.informatik.uni-kl.de
Proxy 5	planetlab1.cs.aueb.gr

Cada cliente sondeó a todos los *proxies* cada 10 segundos durante 2 días. En todos los sondeos se solicitó al *proxy* el mismo fichero[1] y, siempre que se consiguió hacer la descarga, se registró el tiempo de respuesta, entendido como el tiempo que transcurre desde el cliente realiza la petición hasta que éste recibe el último *byte* de la respuesta. De acuerdo con [6], esta medida tiene una correlación alta con el rendimiento real de los *proxies*.

La tabla II recoge el número de medidas del tiempo de respuesta obtenidas por cada cliente para cada *proxy*. En todos los casos se observa que la cantidad de medidas registradas es inferior a los 17.280 sondeos que se realizaron por cada pareja de cliente y *proxy*. En algunos casos las medidas no pudieron realizarse debido a problemas de comunicación entre el cliente y el *proxy*, lo cual puede deducirse de la elevada diferencia en el número de medidas recogidas del *proxy* 5 por los clientes 4 a 6 respecto al resto de clientes. En otros casos el *proxy* no pudo acceder al recurso solicitado por el cliente, con lo que tampoco fue posible medir el tiempo de respuesta. Estos problemas de comunicación tanto entre cliente y *proxy* como entre *proxy* y servidor del recurso solicitado se dan también en las redes comunitarias.

En la figura 1 se muestra la distribución de las medidas recogidas para cada pareja de cliente y *proxy*. En ella se puede ver que las distribuciones de las medidas no son normales en la mayoría de los casos. También se observa que algunas distribuciones presentan una variabilidad significativa. Ambos hechos suponen *a priori* una mayor dificultad para la estimación de los tiempos de respuesta.

[1] http://ovh.net/files/1Mb.dat

Tabla II
NÚMERO DE MEDIDAS DEL TIEMPO DE RESPUESTA RECOGIDAS POR CADA CLIENTE Y PROXY.

	Proxy 1	*Proxy* 2	*Proxy* 3	*Proxy* 4	*Proxy* 5	Total
Cliente 1	15.476	15.474	15.474	15.476	11.664	73.564
Cliente 2	15.486	15.484	15.485	15.486	11.666	73.607
Cliente 3	15.474	15.474	15.473	15.474	11.665	73.560
Cliente 4	14.697	14.697	14.697	14.697	14.697	73.485
Cliente 5	14.742	14.742	14.742	14.742	14.742	73.710
Cliente 6	14.678	14.678	14.678	14.678	14.678	73.390
Cliente 7	15.472	15.470	15.470	15.471	11.658	73.541
Cliente 8	15.502	15.500	15.500	15.501	11.667	73.670
Total	121.519	121.519	121.525	121.525	102.437	588.527

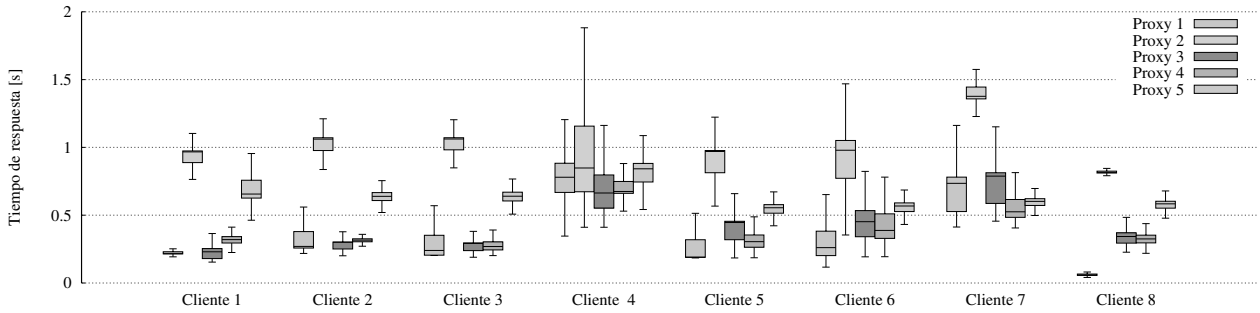


Figura 1. Diagramas de cajas del tiempo de respuesta medido por cada cliente para cada *proxy*.

B. Algoritmos de factorización matricial para la estimación del tiempo de respuesta

Considérese un conjunto de c clientes y p *proxies* web en una red comunitaria. Si, de manera periódica, cada cliente sondea un número de *proxies* inferior a p y compartiera con el resto de clientes las medidas indicativas de calidad de servicio obtenidas, cada cliente podrá crear una matriz de medidas $M \in \mathbb{R}^{c \times p}$. Cada elemento de esta matriz $m_{i,j}$ se corresponde con la medida de indicativa de calidad de servicio entre el cliente c_i y el *proxy* p_j. Dicho elemento tomará un valor conocido en caso de que el cliente c_i haya sondeado p_j exitosamente y un valor desconocido en caso contrario. Se asume que los valores de la matriz presentarán un alto nivel de correlación dado que se espera que dos clientes diferentes que tengan condiciones de acceso de red similares a un mismo *proxy* obtengan medidas similares al sondearlo.

Los valores desconocidos pueden ser estimados mediante la factorización de la matriz M aprovechando dicha correlación. Así, por ejemplo, el algoritmo básico de factorización matricial [25] intenta encontrar las matrices $C \in \mathbb{R}^{d \times c}$ y $P \in \mathbb{R}^{d \times p}$ de forma que la matriz $\hat{M} = C^T P$ de rango d minimice la distancia euclídea a la matriz objetivo M calculada teniendo en cuenta sólo los elementos con valores conocidos. De este modo, a cada elemento $m_{i,j}$ con valor desconocido le corresponde la estimación $\hat{m}_{i,j}$

En este artículo se estudia la posibilidad de estimar el tiempo de respuesta de los *proxies* web de una red comunitaria utilizando 3 algoritmos de factorización matricial bien conocidos. Dichos algoritmos son los siguientes:

- AMF (*Adaptative Matrix Factorization*, factorización matricial adaptativa) [9], que realiza una factorización matricial capaz de tener en cuenta la evolución temporal de los valores en la matriz. Este algoritmo emplea técnicas de transformación de datos, aprendizaje en línea y pesos adaptativos.
- NTF (*Non-negative Tensor Factorization*, factorización tensorial no negativa) [26] también tiene en cuenta la evolución temporal de los valores en la matriz. Sin embargo, en lugar de construir un modelo que evolucione en el tiempo, se genera un modelo a partir de un tensor formado por las matrices correspondientes a diferentes instantes.
- PMF (*Probabilistic Matrix Factorization*, factorización matricial probabilística) [27], que no tiene en cuenta la evolución temporal de los valores de la matriz en la matriz.

La implementación de AMF empleada en los experimentos de este artículo se encuentra en [28], mientras que las de NTF y PMF están diponibles en [29].

C. Algoritmo de base y métrica de rendimiento

La precisión de los algoritmos de factorización matricial estudiados en este artículo se compara con la de un algoritmo de base, estableciéndose así un rendimiento mínimo que se espera que sea superado por los primeros. Este algoritmo de base toma el valor medio de todos los elementos con valor conocido de M como estimación para cualquiera de los elementos con valor desconocido de esa misma matriz.

Para evaluar la precisión de los distintos algoritmos se utiliza el error absoluto medio (MAE, *Mean Average Error*) de las estimaciones. Esta métrica es ampliamente utilizada en la literatura (ej. [7], [8]) para medir el rendimiento de algoritmos de estimación en problemas similares al abordado en este artículo.

D. Experimentos

Los experimentos descritos en este artículo estudian la precisión de los distintos algoritmos suponiendo que no varía el número de *proxies* que cada cliente sondea cada 10 segundos. Sin embargo, los *proxies* que en concreto son sondeados por un cliente dado pueden variar de una ronda de sondeos a otra, puesto que son elegidos aleatoriamente en cada una de ellas. Específicamente, se han llevado a cabo experimentos considerando 2, 3 y 4 *proxies* sondeados por cada cliente. Esto hace que las matrices de medidas que se pueden construir tras cada ronda de sondeos tengan una densidad de valores conocidos del 40 %, 60 % y 80 %, respectivamente. Las medidas que no son incluidas en la matriz de valores conocidos son después utilizadas para evaluar la precisión de las estimaciones.

Para cada densidad de matriz considerada en el artículo se han generado 5 conjuntos diferentes de entrenamiento y test a partir del conjunto de datos inicial. La diferencia entre conjuntos reside en que, para cada ronda de sondeos, los *proxies* que en concreto son sondeados por un cliente dado pueden variar porque son elegidos aleatoriamente.

De este modo, la precisión de todos los algoritmos ha sido evaluada para cada densidad de matriz mediante 5 experimentos de estimación con los conjuntos de datos correspondientes. Para evaluar dicha precisión se han obtenido tanto el valor medio del MAE de las estimaciones hechas en cada ronda como la desviación estándar del MAE entre rondas.

Cada algoritmo utilizado tiene, además, una serie de parámetros que determinan su comportamiento. A continuación se indican los valores que se utilizaron en los experimentos tras llevar a cabo una exploración no sistemática del espacio de posibles valores de los parámetros. Para el caso de AMF, los parámetros son la dimensionalidad de los factores latentes (en nuestro caso 10), la tasa de aprendizaje (0,8), el parámetro de regularización (0,0003), el número máximo de iteraciones (50), el umbral de convergencia (6e-3) y el peso de la media exponencial móvil (0,3). Por su parte, para el caso de NTF los parámetros son, de nuevo, la dimensionalidad de los factores latentes (10), el parámetro de regularización (40), y el número máximo de iteraciones (300). Por último, para PMF los parámetros son la dimensionalidad de los factores latentes(10), el parámetro de regularización (0), el valor inicial de la tasa de aprendizaje (0,01) y el número máximo de iteraciones (600).

IV. RESULTADOS

En la tabla III se muestran los resultados obtenidos al ejecutar los métodos presentados en las subsecciones III.B y III.C sobre el conjunto de datos descrito en la subsección III.A. Para cada método y densidad se muestra el valor medio y la desviación del MAE entre rondas. Como se puede observar, AMF obtiene un rendimiento muy elevado, con errores bajos y poco variables, lo que supone una prueba de que los algoritmos de factorización matricial tienen un gran potencial para estimar desde los clientes el tiempo de respuesta de distintos *proxies*, habiendo sondeado sólo a algunos de ellos y empleando las medidas compartidas por sus vecinos. Además, en la tabla IV se puede apreciar la adecuación al problema de AMF, en gran medida gracias a que incorpora mecanismos de pre- y post-transformación de los datos que son muy adecuados cuando éstos siguen una distribución de probabilidad muy asimétrica, como es el caso. El resultado es considerablemente mejor que el del algoritmo de base, lo que justifica el coste computacional de su utilización (obsérvese que el coste de compartición entre los clientes de las tiempos de respuesta medidos se aplica a los cuatro algoritmos). Finalmente, cabe observar cómo el aumento de la densidad ayuda a mejorar el rendimiento en todos los métodos, aunque en una proporción pequeña, lo que probablemente es debido al escaso número de clientes y *proxies*. En la figura 2 se muestran las estimaciones del tiempo de respuesta que un cliente hace de un *proxy* durante un pequeño intervalo tomado como ejemplo, corroborando que AMF hace estimaciones muy precisas y detecta bien las variaciones más marcadas del tiempo de respuesta.

Para comprender mejor los resultados, las tablas IV y V muestran el desglose del rendimiento de los algoritmos AMF y base, respectivamente, para cada pareja de cliente y proxy, con densidad 40 % (en cada ronda cada cliente sondea a 2 de los 5 *proxies*), aunque las observaciones que se realizan a continuación se repiten para las otras densidades evaluadas. En primer lugar, en la tabla IV puede verse cómo los clientes 1, 2, 3 y 8 obtienen tiempos de respuesta más estables para sus sondeos. Además, estos tiempos son bajos para los *proxies* 1, 3 y 4. Observando el comportamiento de ambos métodos, el AMF obtiene unos resultados de gran precisión para estas parejas cliente y *proxy*. Hay que observar que los otros clientes tienen medidas más altas y con mayor variabilidad, pero AMF aprende para cada cliente cuáles son los clientes más parecidos. Por el contrario, el algoritmo de base trata la información de cualquier otro cliente por igual, cometiendo un alto error para estas parejas (de hecho, en ellas se da el mayor ratio de mejora de AMF sobre el algoritmo de base). Esto refuerza la idea de que este algoritmo de factorización matricial puede ser una buena solución cuando se disponga de muchos clientes, cada uno de los cuales sondee a un conjunto pequeño de *proxies*.

Un detalle particular que se aprecia en la tabla IV es que el cliente 8 obtiene tiempos de respuesta muy bajos del *proxy*. Tanto es así que el rango de estas medidas no se solapa con el de los tiempos de cualquier otra pareja cliente y *proxy*. Esto hace que el resultado sea peor para ambos métodos, ya que en este cliente y para este *proxy* se estima un valor próximo a los medidos en otros clientes. No obstante, aunque afecta a AMF, lo hace en mucha mayor medida con el algoritmo de base. Algo semejante sucede con las medidas que el cliente 7 toma del *proxy* 2, en las que se obtienen valores muy elevados, y con muy poco solape con los medidos en otros clientes. Esto hace que esta pareja de cliente y *proxy* tenga el peor MAE medio, siendo de nuevo una situación mucho más desfavorable para el algoritmo de base.

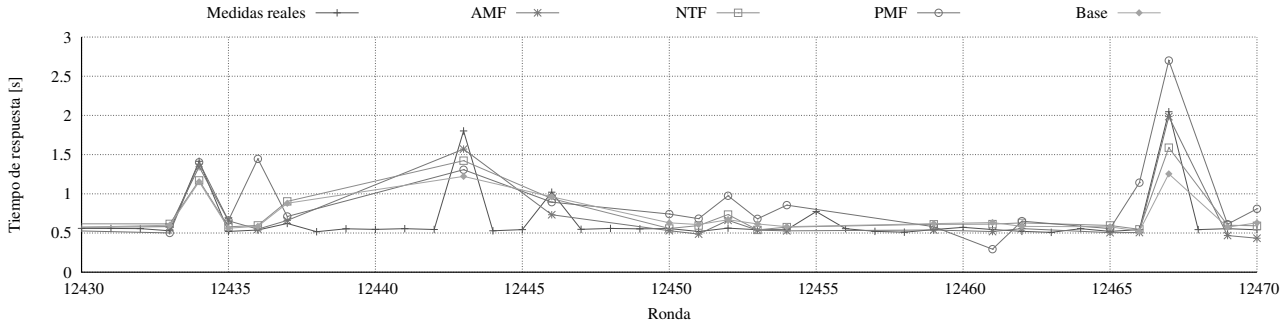


Figura 2. Extracto de la estimación del tiempo de respuesta que el cliente 6 hace del *proxy* 5, utilizando los algoritmos de factorización matricial y de base, frente al tiempo real, con densidad 60 %.

Tabla III
MEDIA Y DESVIACIÓN ESTÁNDAR DEL MAE OBTENIDO A LO LARGO DE LAS RONDAS DEL EXPERIMENTO, PARA AMF, NTF, PMF Y EL ALGORITMO DE BASE, CON DISTINTAS DENSIDADES DEL CONJUNTO DE DATOS

Densidad	AMF	NTF	PMF	Algoritmo de base
40 %	$\mathbf{0,104 \pm 3,8e-4}$	$0,140 \pm 6,18e-3$	$0,570 \pm 9,70e-4$	$0,315 \pm 1,2e-4$
60 %	$\mathbf{0,100 \pm 3,6e-4}$	$0,123 \pm 1,62e-3$	$0,438 \pm 1,05e-3$	$0,313 \pm 4,5e-4$
80 %	$\mathbf{0,096 \pm 5,6e-4}$	$0,116 \pm 3,36e-3$	$0,359 \pm 1,51e-3$	$0,313 \pm 3,2e-4$

Tabla IV
MEDIA DEL MAE OBTENIDO POR AMF PARA CADA PAREJA DE CLIENTE Y PROXY CON DENSIDAD 40 %

	Cliente 1	Cliente 2	Cliente 3	Cliente 4	Cliente 5	Cliente 6	Cliente 7	Cliente 8	Media
Proxy 1	0,050	0,055	0,051	0,170	0,109	0,119	0,108	0,203	0,108 ± 0,00069
Proxy 2	0,136	0,141	0,139	0,138	0,183	0,194	0,209	0,137	0,159 ± 0,00175
Proxy 3	0,043	0,044	0,041	0,163	0,107	0,116	0,100	0,051	0,082 ± 0,00050
Proxy 4	0,057	0,050	0,057	0,158	0,113	0,125	0,087	0,056	0,087 ± 0,00052
Proxy 5	0,059	0,053	0,056	0,151	0,093	0,091	0,073	0,053	0,080 ± 0,00049
Media	0,069	0,069	0,069	0,156	0,122	0,130	0,118	0,102	0,104 ± 0,00038

Tabla V
MEDIA DEL MAE OBTENIDO POR EL ALGORITMO DE BASE PARA CADA PAREJA DE CLIENTE Y PROXY CON DENSIDAD 40 %

	Cliente 1	Cliente 2	Cliente 3	Cliente 4	Cliente 5	Cliente 6	Cliente 7	Cliente 8	Media
Proxy 1	0,383	0,306	0,343	0,257	0,367	0,343	0,180	0,536	0,340 ± 0,00076
Proxy 2	0,454	0,528	0,532	0,366	0,457	0,447	0,927	0,322	0,506 ± 0,00209
Proxy 3	0,388	0,329	0,338	0,180	0,219	0,208	0,226	0,277	0,273 ± 0,00082
Proxy 4	0,291	0,297	0,342	0,205	0,313	0,269	0,152	0,286	0,270 ± 0,00071
Proxy 5	0,153	0,122	0,124	0,321	0,164	0,129	0,112	0,109	0,157 ± 0,00079
Media	0,343	0,327	0,347	0,267	0,307	0,282	0,330	0,316	0,315 ± 0,00012

Por último, la última columna de las tablas IV y V muestra los agregados por *proxy*, incluyendo la desviación estándar. Cabe apreciar que la desviación del error es muy pequeña en todos los casos, aunque es sensiblemente mayor en el *proxy* 2, tanto para AMF como para el algoritmo de base. De hecho, individualmente para cada cliente también la desviación es mayor para este *proxy* (esto no se muestra en la tabla por problemas de espacio). Finalmente, los valores medios del MAE son, con ambos métodos, mayores con el *proxy* 2. La explicación de esto está en la enorme variabilidad de las medidas de este *proxy* tomadas desde los clientes 4, 5 y 6. Aquí se puede apreciar que el hecho de tener clientes muy ruidosos, que obtienen medidas en rangos que se solapan con los de otros clientes (y, que por lo tanto hacen que *a veces* sean parecidos) distorsiona las estimaciones que hacen esos otros clientes.

V. CONCLUSIONES Y TRABAJO FUTURO

El acceso a la Web en redes comunitarias se hace habitualmente a través de *proxies*. Por este motivo, es importante que los clientes seleccionen *proxies* web que puedan ofrecer una calidad de servicio adecuada a los usuarios. Esta selección puede ser llevada a cabo por cada cliente basándose tanto en las medidas indicativas de calidad de servicio recogidas por el propio cliente mediante el sondeo de algunos *proxies* como en las estimaciones de las medidas que el mismo cliente podría obtener de otros *proxies* si los sondeara. Estas estimaciones pueden hacerse empleando algoritmos de factorización matricial a partir de las medidas tomadas por un cliente dado junto con las medidas realizadas por otros clientes, los cuales las comparten con el primero.

En este artículo se ha evaluado la precisión de 3 algoritmos de factorización matricial para la estimación del tiempo de respuesta de los *proxies* web en redes comunita-

rias. Para ello se ha realizado una serie de experimentos de estimación empleando los datos obtenidos de la emulación de una red comunitaria formada por 8 clientes y 5 *proxies* web en la que los primeros sondeaban cada 10 segundos a los segundos. Los resultados muestran el potencial de los algoritmos de factorización matricial, y en concreto de AMF, para la estimación del tiempo de respuesta a partir unas pocas medidas tomadas desde los clientes y compartidas entre ellos.

No obstante, para generalizar estos resultados, en el trabajo futuro está previsto generar un conjunto de datos en el que tanto el número de clientes como el de *proxies* sea mayor, lo que permitiría comprender mejor el efecto de la densidad de medidas y validar las observaciones realizadas a partir de los experimentos de este artículo. Además, se pretende que este nuevo conjunto de datos sea obtenido de clientes y *proxies* web desplegados en una red comunitaria real.

En relación a los algoritmos empleados, se llevará a cabo una exploración sistemática del espacio de valores que pueden tomar los parámetros de entrenamiento. También se comprobará si la técnica de transformación de datos empleada por AMF puede ser usada también para mejorar los resultados de NTF y PMF.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional y la Agencia Estatal de Investigación del Ministerio de Ciencia, Innovación y Universidades a través de los proyectos de investigación TIN2017-85179-C3-2-R y TIN2016-77836-C2-2-R.

REFERENCIAS

[1] Bart Braem, Chris Blondia, Christoph Barz, Henning Rooge, Felix Freitag, Leandro Navarro, Joseph Bonicioli, Stavros Papathanasiou, Pau Escrich, Roger Baig Viñas, Aaron Kaplan, Axel Neumann, Ivan Vilata i Balaguer, Blain Tatum, Malcom Matson, "A case for research with and on community networks", ACM SIGCOMM Computer Communication Review, n. 3, pp. 68-73, 2013. doi: 10.1145/2500098.2500108

[2] Leonardo Maccari, "An analysis of the Ninux wireless community network", Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1-7, 2013. doi: 10.1109/WiMOB.2013.6673332

[3] Funkfeuer, "Initiative für freie Netze", https://www.funkfeuer.at/, Última visita: 9 de mayo de 2019.

[4] Davide Vega, Roger Baig, Llorenç Cerdà-Alabern, Esunly Medina, Roc Meseguer, Leandro Navarro, "A technological overview of the guifi.net community network", Computer Networks, vol. 93, pp.260-278, 2015. doi: 10.1016/j.comnet.2015.09.023

[5] Emmanouil Dimogerontakis, João Neto, Roc Meseguer, Leandro Navarro, Luís Vega, "Client-side routing-agnostic gateway selection for heterogeneous wireless mesh networks", Symposium on Integrated Network and Service Management (IM), pp. 377-385, 2017. doi: 10.23919/INM.2017.7987301

[6] Khulan Batbayar, Emmanouil Dimogerontakis, Roc Meseguer, Leandro Navarro, Esunly Medina, Rodrigo M. Santos, "The RIMO Gateway Selection Approach for Mesh Networks: Towards a Global Internet Access for All", Conference on Ubiquitous Computing and Ingelligence (UCAmI), vol. 2, n. 1258, 2018.

[7] Wei Lo, Jianwei Yin, Shuiguang Deng, Ying Li, Zhaohui Wu, "An extended matrix factorization approach for QoS prediction in service selection", Conference on Services Computing (SCC), pp. 162-169, 2012. doi: 10.1109/SCC.2012.36

[8] Zibin Zheng, Hao Ma, Michael R. Lyu, Irwin King, "Collaborative web service QoS prediction via neighborhood integrated matrix factorization", IEEE Transations on Services Computing", vol. 6, n. 3, pp. 289–299, 2013. doi: 10.1109/TSC.2011.59

[9] Jieming Zhu, Pinjia He, Zibin Zheng, Michael R. Lyu, "Online QoS prediction for runtime service adaptation via adaptive matrix factorization", IEEE Transactions on Parallel and Distributed Systems, vol. 28, n. 10, pp. 2911-2924, 2017. doi: 10.1109/TPDS.2017.2700796

[10] Rupa Krishnan, Harsha V. Madhyastha, Sridhar Srinivasan, Sushant Jain, Arvind Krishnamurthy, Thomas Anderson, Jie Gao "Moving beyond end-to-end path information to optimize CDN performance", SIGCOMM Internet Measurement Conference (IMC), pp. 190-201, 2009. doi: 10.1145/1644893.1644917

[11] Qianwen Yin, Jasleen Kaur, F. Donelson Smith, "Can bandwidth estimation tackle noise at ultra-high speeds?", International Conference on Network Protocols (ICNP), pp. 107–118, 2014. doi: 10.1109/ICNP.2014.31

[12] Stefano Salsano, Fabio Patriarca, Francesco Lo Presti, Pier Luigi Ventre, Valerio Maria Gentile, "Accurate and Efficient Measurements of IP Level Performance to Drive Interface Selection in Heterogeneous Wireless Networks", IEEE Transactions on Mobile Computing, vol. 17, n. 10, pp. 2223-2235, 2018. doi: 10.1109/TMC.2018.2807842

[13] Hong Zhang, Junxue Zhang, Wei Bai, Kai Chen, Mosharaf Chowdhury, "Resilient Datacenter Load Balancing in the Wild", ACM Special Interest Group on Data Communication (SIGCOMM), pp. 253–266, 2017. doi: 10.1145/3098822.3098841

[14] Frank Dabek, Russ Cox, Frans Kaashoek, Robert Morris, "Vivaldi: a decentralized network coordinate system.", ACM Special Interest Group on Data Communication (SIGCOMM), pp. 15-26, 2004. doi: 10.1145/1015467.1015471

[15] Jonathan Ledlie, Margo Seltzer, Peter Pietzuch, "Proxy network coordinates", Target, vol. 22, pp. 25, 2008.

[16] Cristian Lumezanu, Randolph Baden, Neil Spring, Bobby Bhattacharjee, "Triangle inequality variations in the internet", SIGCOMM Internet measurement conference (IMC) pp. 177-183, 2009. doi: 10.1145/1644893.1644914

[17] Bong-Jun Ko, Sisi Liu, Murtaza Zafer, Ho Yin Starsky Wong, Kang-Won Lee, "Gateway selection in hybrid wireless networks through cooperative probing", Symposium on Integrated Network Management (IM), pp. 352–360, 2007.

[18] Khulan Batbayar, Roc Meseguer, Emmanouil Dimogerontakis, Leandro Navarro, Ramin Sadre, "Collaborative informed gateway selection in large-scale and heterogeneous networks", Symposium on Integrated Network Management (IM), pp. 337–345, 2019.

[19] Laisen Nie, Dingde Jiang, Lei Guo, Shui Yu, "Traffic matrix prediction and estimation based on deep learning in large-scale IP backbone networks", Journal of Network and Computer Applications, vol. 76, pp. 16–22, 2016. doi: 10.1016/j.jnca.2016.10.006

[20] Xianrong Zhen, Li Da Xu, Sheng Chai, "QoS recommendation in cloud services", IEEE Access, vol. 5, pp. 5171-5177, 2017. doi: 10.1109/ACCESS.2017.2695657

[21] Xiong Luo, Ji Liu, Dandan Zhang, Xiaohui Chang, "A large-scale web QoS prediction scheme for the Industrial Internet of Things based on a kernel machine learning algorithm", Computer Networks, vol. 101, pp. 81-89, 2016. doi: 10.1016/j.comnet.2016.01.004

[22] G. Vadivelou, "Collaborative filtering based web service recommender system using users' satisfaction on QoS", International Conference on Inventive Computation Technologies (ICICT), pp.1-5, 2016. doi: 10.1109/INVENTIVE.2016.7830110

[23] Brent Chun, David Culler, Timothy Roscoe, Andy Bavier, Larry Peterson, Mike Wawrzoniak, Mic Bowman, "Planetlab: an overlay testbed for broad-coverage services", ACM SIGCOMM Computer Communication Review, vol. 33, n. 3, pp. 3-12, 2003. doi: 10.1145/956993.956995

[24] "Lightweight HTTP, HTTPS, WebSockets Proxy Server in a single Python file", GitHub Repository, https://github.com/abhinavsingh/proxy.py, Última visita: 24 de mayo de 2019.

[25] Robert Bell, Yehuda Koren, Chris Volinsky, "Matrix factorization techniques for recommender systems", Computer, vol. 42, pp. 30-37, 2009. doi: 10.1109/MC.2009.263

[26] Wancai Zhang, Hailong Sun, Xudong Liu, Xiaohui Guo, "Temporal QoS-aware web service recommendation via Non-negative Tensor Factorization", Conference on World Wide Web (WWW), pp. 585-596, 2014. doi: 10.1145/2566486.2568001

[27] Ruslan Salakhutdinov, Andriy Mnih, "Probabilistic Matrix Factorization", Conference on Neural Information Processing Systems (NIPS), pp. 1257-1264, 2007.

[28] "AMF: Adaptive Matrix Factorization for Online QoS Prediction", GitHub Repository, https://github.com/wsdream/AMF, Última visita: 20 de mayo de 2019.

[29] "Towards Open Datasets and Source Code for Web Service Recommendation", GitHub Repository, https://github.com/wsdream/AMF, Última visita: 20 de mayo de 2019.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Combinación de Network Coding Sistemático y Solapamiento en escenarios IoT.

Mihail Zverev[1], Pablo Garrido[1], Ramón Agüero[2], Josu Bilbao[1].
[1]Information and Communication Technologies Area, Ikerlan Technology Research Center,
Paseo José María Arizmendiarrieta, 2, 20500 Mondragón, Gipuzkoa, España.
[2]Dpto. Ingenería Comunicaciones, Universidad de Cantabria,
Avenida Los Castros, s/n, 39005, Santander, Cantabria, España.
[1]{mzverev, pgarrido, jbilbao}@ikerlan.es, [2]ramon@tlmat.unican.es

***Resumen*—Las aplicaciones de las arquitecturas IoT son cada vez más numerosas, y abarcan desde dispositivos *wearables* hasta comunicaciones en entornos vehiculares. Con la consolidación de la Industria 4.0, los entornos IIoT (*Industrial IoT*) son cada vez más comunes. Los enlaces inalámbricos sobre los que se sustentan son generalmente susceptibles a sufrir pérdidas en las transmisiones, y su recuperación puede incrementar notablemente el *retardo*. La fiabilidad (robustez), así como el retardo, son requisitos fundamentales en las comunicaciones IIoT. Una posible estrategia para mejorar el comportamiento de las comunicaciones es la utilización de técnicas *Network Coding* (codificación de red). Su utilización para la mejora del comportamiento de comunicaciones en entornos IIoT no ha recibido hasta la fecha suficiente atención, y no existe ninguna implementación que cuente con una amplia aceptación. En este trabajo se analiza una alternativa interesante para comunicaciones IIoT, basada en la combinación de un esquema *Network Coding* sistemático y una estrategia de solapamiento. Se lleva a cabo un análisis exhaustivo del comportamiento de esta solución, extendiendo notablemente los estudios previos, y se obtienen una serie de conclusiones prácticas de cara a su implementación.**

***Palabras Clave*—TF, CCRS, RD, Network Coding, codificación sistemática, IIoT**

I. INTRODUCCIÓN

Las casas inteligentes, los dispositivos *wearables* y las comunicaciones vehiculares son sólo algunos de los ejemplos de las crecientes aplicaciones IoT. A medida que las fábricas están evolucionando hacia la Industria 4.0, el entorno IIoT (*Industrial IoT*) es cada vez más común. Un despliegue IIoT tiene típicamente múltiples dispositivos interconectados a través de una red, que puede estar caracterizada por sufrir pérdidas de información, siendo la fiabilidad y el bajo retardo requisitos estrictos.

Es bien sabido que los enlaces inalámbricas son propensos a inducir errores en la comunicación. El reenvío de la información perdida (para ser recuperada) supone un aumento de la carga de la red y aumenta el retardo, no sólo en la transmisión que genera el reenvío, sino en todas las comunicaciones de la red.

Existen múltiples técnicas que permiten aumentar la robustez de la comunicación y reducir su retardo, previniendo el reenvío de la información perdida. Dichas técnicas suelen consistir en la codificación del mensaje de tal manera que el receptor pueda recuperar la información original a pesar de las pérdidas. Hasta la fecha la codificación preventiva en IoT no ha recibido suficiente atención. El primer paso en la búsqueda de una implementación de las técnicas de codificación preventiva optimizada para IoT debería ser un estudio de los esquemas de codificación existentes. Este trabajo se centra precisamente en un esquema de codificación que permite responder a estas dos necesidades de las comunicaciones inalámbricas en entornos IIoT.

El resto del trabajo se estructura tal y como sigue. La Sección II proporciona un breve resumen de diferentes esquemas de codificación que se han propuesto en la literatura relacionada, y se identifican aspectos no cubiertos hasta la fecha, que son los que se estudian en este trabajo. En la Sección III se describe la implementación del esquema de codificación que se propone. La Sección IV resume los resultados obtenidos tras las simulaciones realizadas, mientras que en la Sección V, se concluye el trabajo, identificado los aspectos que aparecen como líneas de investigación futuras.

II. ESTADO DEL ARTE

Una técnica que tradicionalmente se ha usado para incrementar la fiabilidad de las comunicaciones es FEC (*Forward Error Correction*). Con ella, los fragmentos de información, o *símbolos*, que envía un transmisor son protegidos por otros símbolos -de reparación, redundantes o codificados- que se envían junto con los originales. Así, las pérdidas se pueden recuperar en el receptor con los símbolos redundantes. Otra forma de aumentar la robustez de la comunicación es el uso de NC (*Network Coding*, codificación de red). Una de las configuraciones

más empleadas en soluciones NC es la conocida como *Intra-flow* (definido en [1]), que comparte varias de las características del FEC, ya que con esta técnica también se envían símbolos codificados para proteger los originales, sólo que en este caso los símbolos codificados podrían ser generados, además de por la fuente. por los nodos intermedios (routers). Así, NC podría generar los símbolos codificados necesarios para cada enlace entre routers consecutivos, mientras que FEC lo hace para toda la comunicación, extremo a extremo. El uso de NC en lugar de FEC supone una reducción de sobrecarga de la red, especialmente apreciable en las redes inalámbricas malladas.

Una técnica NC que ha suscitado mucho interés es RLNC (*Random Linear NC*), introducida en [2] y posteriormente extendida en [3]. Los símbolos originales o *fuente* se agrupan en bloques conocidos como *generaciones*, para posteriormente ser combinados linealmente con el uso de coeficientes aleatorios en un Campo de Galois $GF(2^q)$. Cada combinación lineal de los símbolos fuente se conoce como *símbolo codificado*, y se generan tantos como sean necesarios para cubrir las pérdidas en el próximo enlace. Al recibir los nodos intermedios los símbolos codificados suficientes para recuperar la información original, pueden volver a hacer una combinación lineal con coeficientes aleatorios para generar los símbolos necesarios en el próximo enlace, incluso sin tener que decodificarlos. Es decir, los nodos pueden *recodificar sin decodificar*. El receptor podrá recuperar el mensaje original siempre y cuando reciba suficientes símbolos codificados, independientemente del número de recodificaciones por las que ha pasado el mensaje. Esta técnica de codificación también puede aplicarse a FEC (codificación en los extremos de comunicación, sin recodificación), siendo este esquema conocido como RLC (*Random Linear Coding*).

RLNC introduce una complejidad computacional que sería interesante evitar en el ámbito IoT, por las limitaciones de los dispositivos. Una manera de reducir dicha complejidad es bajar la densidad de la codificación, dando lugar a técnicas SNC (*Sparse NC*). Para codificar se usan sólo *algunos* símbolos fuente en lugar de todos, lo que implica que un símbolo codificado sólo incluye una parte de toda la generación [4]. Una de las técnicas SNC es el NC sistemático (SysNC) [5] que, a diferencia del RLNC clásico, envía los símbolos fuente sin codificar junto con los símbolos codificados necesarios. Gracias a ello, gran parte de la información puede usarse nada más ser recibida, ya que no es necesario decodificarla.

Se dice que el RLNC clásico es un esquema de codificación por bloques, puesto que divide la información en unos bloques (generaciones) que no se solapan. El *solapamiento* tiene sus ventajas, fáciles de ver con SysNC: al solapar las generaciones, se reduce el periodo existente entre símbolos codificados, pudiéndose recuperar los símbolos perdidos con mayor frecuencia. Al solapar las generaciones se puede definir además una ventana deslizante de codificación, que cubra los símbolos que pertenecen a la generación más nueva. Un ejemplo de SysNC con una ventana de codificación y decodificación deslizante es el protocolo *Caterpillar RLNC* (CRLNC), presentado en [6] y extendido con ARQ (*Automatic Repeat reQuest*) en [7]. CRLNC ofrece una probabilidad de decodificación muy cercana a SysNC por bloques, pero con un retardo significativamente menor.

Un enfoque semejante al solapamiento es el *intercalado*. Un bloque de generaciones (*bloque de intercalado*) se envía símbolo a símbolo con multiplexación temporal. Esto es, en primer lugar se envía el primer símbolo de cada generación del bloque, luego el segundo, y así sucesivamente hasta enviar el bloque de intercalado completamente, incluyendo los símbolos codificados. La ventaja del intercalado es su robustez a pérdidas de símbolos a ráfagas. La multiplexación temporal introduce un retardo que se puede evitar según se plantea en [8], utilizando la técnica denominada *Interleaving with On-the-fly Coding* (IOC). En esta técnica los símbolos fuente son asignados a cada generación del bloque al estilo Round-Robin, lo que permite enviarlos en su orden natural. Su principal desventaja es que puede requerir una memoria elevada, tanto para la codificación como para la decodicación.

El estudio de la viabilidad de IOC, o la de su combinación con solapamiento, en entornos IoT no se ha evaluado hasta la fecha. A pesar de que puede resultar eficiente, podría requerir una complejidad computacional elevada.

Teniendo en cuenta todo lo anterior, se puede decir que el esquema que mejor parece adecuarse al uso en IoT es SysNC con solapamiento. En [6], el protocolo CRLNC utiliza sólo un símbolo codificado por generación. Por su parte, los autores del trabajo original presentan resultados que muestran que para generaciones mayores la probabilidad de pérdida de símbolos se reduce, y que el retardo extremo a extremo era menor para las generaciones más pequeñas. Sin embargo, no se analiza el posible impacto sobre la eficiencia causado por la introducción de más de un símbolo redundante por generación. En este trabajo se diseña e implementa una solución SysNC con solapamiento, para ejecutar una campaña de simulaciones, a fin de analizar de manera detallada su comportamiento.

III. IMPLEMENTACIÓN DE NC SISTEMÁTICO

Dependiendo de la implementación concreta de los algoritmos de NC, se pueden incluir múltiples símbolos en un mismo paquete codificado. Sin embargo, la pérdida de un único paquete puede tener como consecuencia la pérdida de una ráfaga de símbolos. A primera vista, podría parecer por tanto que enviar varios símbolos en un paquete no es apropiado. En este estudio se consideran las transmisiones de un símbolo por paquete. Así, desde este momento se utilizarán indistintamente los términos *paquete* y *símbolo*.

A. Esquema de codificación

Se ha llevado a cabo una implementación completa del esquema SysNC con solapamiento en Python. Para ello, se hizo uso del esquema RLNC para generar los símbolos

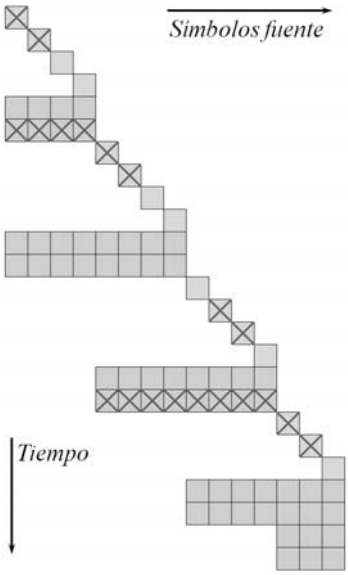


Fig. 1. Ejemplo de la implementación del esquema de codificación. Los cuadrados amarillos son los símbolos fuente, los cuadrados verdes, los símbolos codificados. Las aspas rojas marcan los símbolos perdidos.

redundantes, para lo que se combinan aleatoriamente paquetes de información, utilizando coeficientes aleatorios del campo de Galois $GF(2^8)$.

Se define una *generación* como un conjunto de símbolos originales que, al igual que en [6], se considera constante. Cada generación se protege por r paquetes codificados, y cada paquete se construye combinando símbolos de información que pertenecen a φ generaciones, lo que se define como *solapamiento*. Por otro lado, se considera un *bloque* como el conjunto de símbolos de información que se utilizan para construir un paquete codificado, es decir, compuesto por el solapamiento de φ generaciones. Los tamaños de generación y de bloque se designan con g y k, respectivamente.

Con esta definición se puede ver el esquema RLNC por bloques como un caso particular de la versión con solapamiento con $\varphi = 1$. Esta definición implica la creación de generaciones parciales adicionales para proteger los últimos símbolos fuente de la comunicación. A modo ilustrativo se presenta la Figura 1, donde los cuadrados amarillos son los símbolos fuente y los cuadrados verdes, los símbolos codificados. En este ejemplo de tan sólo 15 símbolos fuente y $r = \varphi = 2$, se puede ver claramente que sin los últimos $r = 2$ símbolos codificados, el último bloque no cumpliría con la definición del solapamiento, perteneciendo únicamente a $1 < \varphi = 2$ generación.

En el ejemplo presentado en la Figura 1, se considera una comunicación con pérdida de paquetes. Los símbolos perdidos se marcan con un aspa roja. En este ejemplo es imposible recuperar los símbolos fuente de los primeros 3 bloques hasta recibir el último símbolo codificado. Este ejemplo puede expandirse a una cantidad de símbolos fuente mucho mayor con el mismo problema: los primeros símbolos perdidos pueden todavía recuperarse, pero al final de la comunicación. Dependiendo del protocolo que haga uso de este esquema de codificación, gran parte

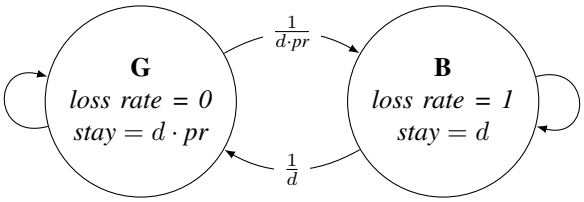


Fig. 2. Implementación del modelo Gillbert-Elliot.

de los símbolos perdidos seguramente acabarán siendo reenviados por el transmisor mucho antes de que puedan ser recuperados. Aun así, mantener todos los símbolos recibidos en memoria permitirá en algún momento recuperar los símbolos perdidos. En otras palabras, *la probabilidad de recuperación depende de la cantidad de memoria disponible en el receptor*. De aquí se llega a la conclusión que la ventana de decodificación debería ser lo más grande posible. Dado que en este estudio se pretende explorar las oportunidades que ofrece el esquema SysNC, en el receptor se define suficiente memoria para almacenar todos los paquetes fuente y codificados.

Cada vez que se recibe un símbolo de información se lanza el proceso de recuperación. Dado que el canal (descrito en la sección B) no reordena los símbolos, esta política de recuperación equivale a intentar recuperar los símbolos perdidos después de cada símbolo recibido.

B. Canal

Se considera un canal inalámbrico entre el transmisor y el receptor, en el que se pueden producir pérdidas. Además, no se reordena los paquetes que viajan por el enlace.

En trabajos previos es bastante común usar una distribución uniforme para modelar la pérdida de los símbolos (canal sin memoria), como en [5]. Sin embargo, en las redes reales los errores a ráfagas pueden ser más comunes. Este tipo de comportamiento se puede capturar con el modelo de Gillbert-Elliot, como en [6] y [8]. Con el fin de utilizar las mismas herramientas de estudios previos, en este se contemplan los dos tipos de comportamientos. A continuación se detalla la implementación que se ha llevado a cabo para el canal a ráfagas.

En el estado de la ráfaga (B, del inglés *burst* o *bad*) la tasa de pérdida es del 100%. La tasa de pérdida en el estado "tranquilo" (G, del inglés *good*) es 0%. Cada ráfaga en media tiene una duración de d paquetes, y el estado *'bueno'* tiene una duración media pr veces mayor que la de la ráfaga, $pr \cdot d$. La figura 2 muestra un esquema del modelo que se ha utilizado. En el presente trabajo se considera la proporción entre las duraciones de los estados de $pr = 10$.

C. Parámetros del modelo

Algunos de los parámetros de entrada del modelo de comunicación del presente estudio ya han sido mencionados: tamaño de generación, símbolos redundantes por generación, solapamiento, tasa de pérdidas, duración de la ráfaga de errores, y número de paquetes a almacenar en el receptor. En este punto únicamente queda un

parámetro por definir: la cantidad de símbolos a transmitir. Un planteamiento que se puede tomar en este tipo de simulaciones es enviar los símbolos fuente con los codificados hasta que el receptor consiga recibir correctamente n símbolos fuente. En el presente trabajo se sigue otro planteamiento: se envían exactamente n símbolos fuente con los correspondientes símbolos codificados al receptor.

Para ver si el es esquema de codificación funciona, es necesario conocer si el receptor ha sido capaz de recibir correctamente todos los paquetes fuente que le fueron enviados, y en caso contrario, cuántos ha recibido. Por tanto, el principal parámetro que se utilizará para evaluar el comportamiento de la solución propuesta será el número de símbolos fuente recibidos, ya sea directamente, o a través de los algoritmos de recuperación. Otra medida interesante podría ser la relación entre los símbolos recuperados y los perdidos. En este trabajo se evalúa la probabilidad de recibir (recepción normal + recuperación) todos los símbolos fuente. Dicha probabilidad se obtiene dividiendo el número total de eventos en los que se reciben o recuperan todos los paquetes fuente entre el número de simulaciones.

Para evaluar el retardo se asumirá un canal ranurado. Para una transmisión de datos correcta y fiable, es de gran importancia la entrega a la capa de aplicación de los paquetes en orden, por lo que se estudiará el retardo extremo a extremo. Se ignora además el retardo de los paquetes no recuperados, teniéndose en cuenta únicamente los casos en los que la probabilidad de recibir (o recuperar) todos los símbolos fuente sea 1.

Como se ha mencionado anteriormente, se pretende evaluar el impacto de variar las redundancias por generación en el rendimiento de SysNC con solapamiento. Por lo tanto, los dos parámetros de entrada más importantes son el solapamiento y las redundancias por generación. Con el fin de simplificar el análisis de los resultados, las dos variables se combinan en una, la *sobrecarga*, que se define como la relación entre el número total de paquetes codificados generados, y el número total de paquetes (fuente y codificados) que envía el transmisor. Como se ha podido observar en el ejemplo de la Figura 1, al final de la comunicación se enviarán $(\varphi - 1) \cdot r$ símbolos codificados adicionales. El número total de paquetes codificados, C, que protegen S paquetes fuente se puede calcular como indica la Ec. 1.

$$C = \left(\left\lceil \frac{S}{k} \right\rceil + \varphi - 1 \right) \cdot r = \left(\left\lceil \frac{S}{\lfloor g/\varphi \rfloor} \right\rceil + \varphi - 1 \right) \cdot r \quad (1)$$

A partir de C (Ec. 1) se puede calcular la sobrecarga $\hat{O}$, como se muestra en la Ec. 2.

$$\hat{O} = \frac{C}{C+S} = \frac{\left(\left\lceil \frac{S}{\lfloor g/\varphi \rfloor} \right\rceil + \varphi - 1 \right) \cdot r}{\left(\left\lceil \frac{S}{\lfloor g/\varphi \rfloor} \right\rceil + \varphi - 1 \right) \cdot r + S} \quad (2)$$

IV. RESULTADOS

Para ver el efecto de la redundancia por cada generación, con diferentes grados de solapamiento, se llevó a cabo una campaña de simulaciones con los siguientes parámetros de entrada: generaciones de $g = 64, 256$ símbolos; solapamientos de $\varphi = 1, 2, 4, 8$ generaciones; 1000 paquetes fuente; tasas de pérdidas de 1%, 5% y 10% con distribución de pérdidas uniforme y una tasa de 9% con ráfagas de pérdidas de 5 paquetes en media y 50 paquetes entre ráfagas. La memoria del receptor ilimitada, pudiéndose almacenar todos los paquetes recibidos y recuperados. En total se ejecutaron 1000 simulaciones independientes por cada configuración, para asegurar la validez estadística de los resultados.

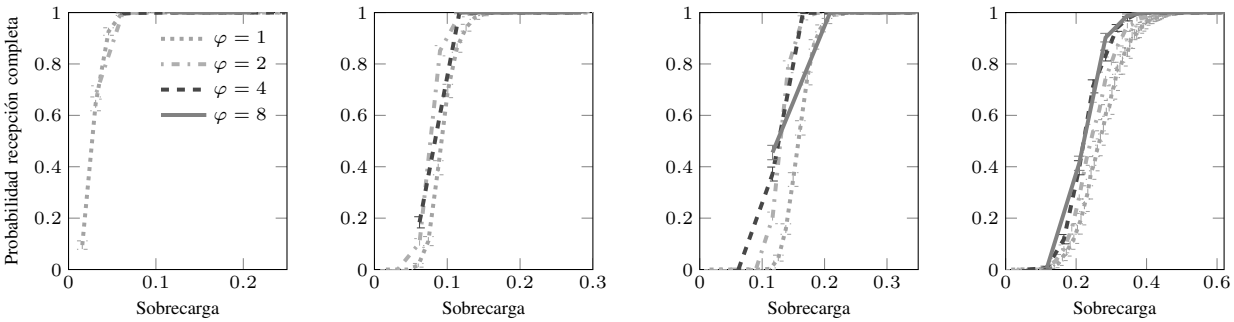
En las Figuras 3 y 4 se representan la probabilidad de recepción para diferentes solapamientos (fila de arriba) y las correspondientes latencias extremo a extremo (fila de abajo).

Dada la definición de la sobrecarga en este trabajo (Ec. 2), sus valores serán más o menos distanciados, dependiendo de φ y r.

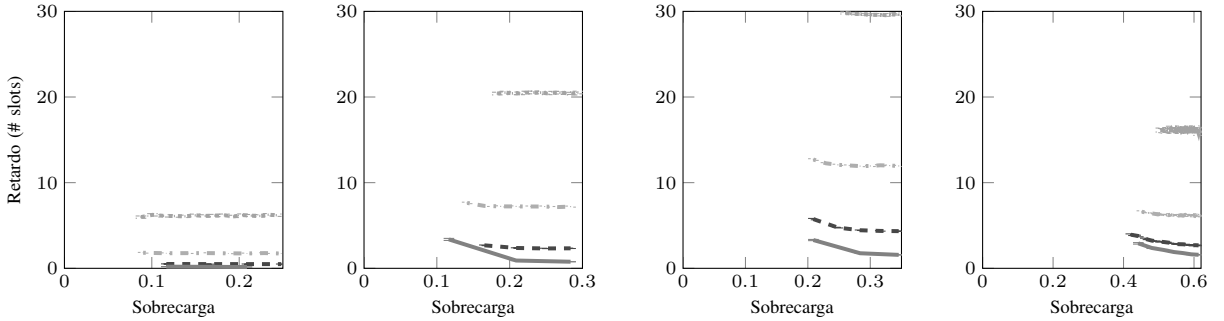
Como se puede ver en la figura 3-c, el solapamiento de 8 generaciones aparentemente no siempre ofrece la mejor probabilidad de recepción. En realidad, entre la primera y la segunda muestra hay mucha distancia, en la que los otros solapamientos tienen más muestras definidas. Para una comparación justa es necesario fijarse en los puntos de la sobrecarga para los que las muestras de las curvas de interés están definidas. Es importante tener en cuenta que *a mayor solapamiento, mayores son las sobrecargas, al aumentar las redundancias por generación*.

En las figuras 3 y 4 se puede observar que al aumentar el solapamiento, crece la probabilidad de recepción y baja el retardo. Como se puede ver, la principal ventaja de SysNC con solapamiento es la reducción de la latencia que se consigue que, según los resultados obtenidos, se puede decir que tiende a ser proporcional a φ, lo cual es de esperar, ya que los bloques (según se definen en la sección III.A) son precisamente φ veces menores con respecto a SysNC sin solapamiento.

Un detalle interesante es la sobrecarga en la que la probabilidad de recepción alcanza el valor 1. Dicha sobrecarga se puede denominar *sobrecarga de saturación*. Este parámetro resulta de especial interés a la hora de implementar NC en una red con un ancho de banda restringido. Dado que el retardo se han representado sólo para aquellos casos en los que se recibe toda la información, la sobrecarga de saturación queda marcada por la primera muestra de cada curva de latencia. En la figura 5 se representan los valores de las sobrecargas de saturación para cada caso estudiado. Como se puede ver, a mayor solapamiento, no siempre se consigue una probabilidad de saturación más baja. Esta depende tanto de la capacidad del esquema de codificación de recuperar los paquetes perdidos, como de la propia sobrecarga que introduce, por lo que no se puede establecer una conclusión clara respecto a este parámetro.

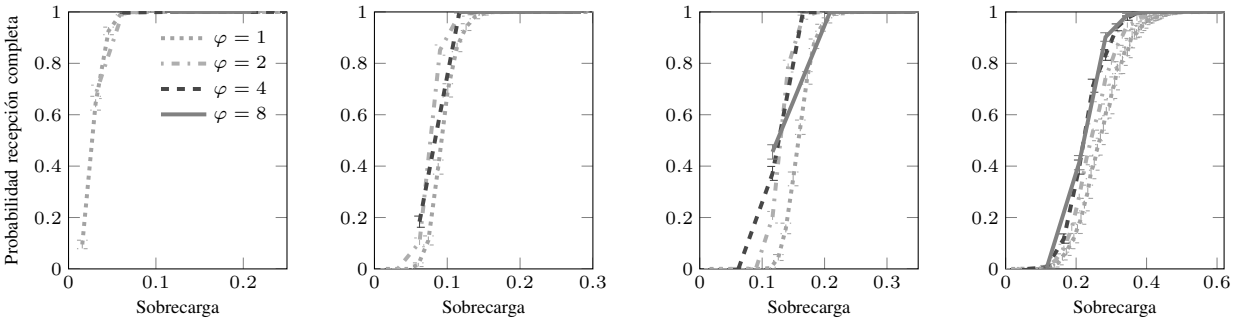


(a) 1% pérdidas uniformes. (b) 5% pérdidas uniformes. (c) 10% pérdidas uniformes. (d) 10% pérdidas a ráfagas.

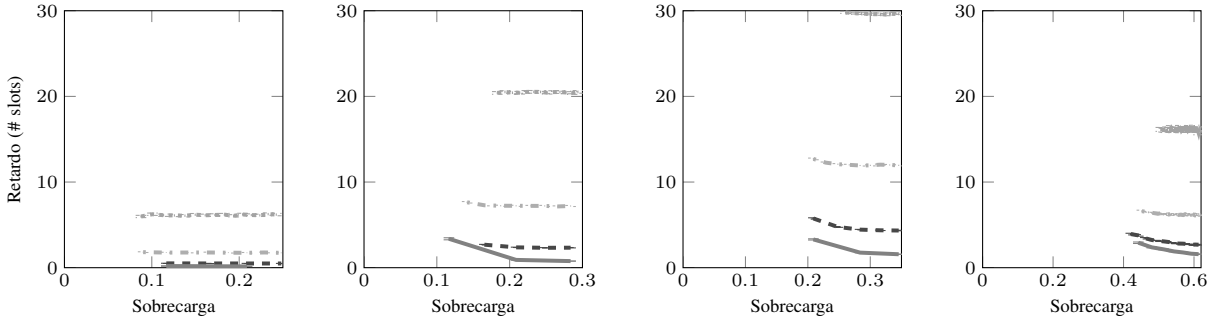


(e) 1% pérdidas uniformes. (f) 5% pérdidas uniformes. (g) 10% pérdidas uniformes. (h) 10% pérdidas a ráfagas.

Fig. 3. Probabilidad de recibir y recuperar todos los paquetes enviados (a–d) y los retardos extremo a extremo (e–h) para un tamaño de generación de 64 símbolos. Las gráficas corresponden a pérdidas uniformes de 1% (a, e), 5% (b, f) y 10% (c, g), y pérdidas en ráfagas (d, h).

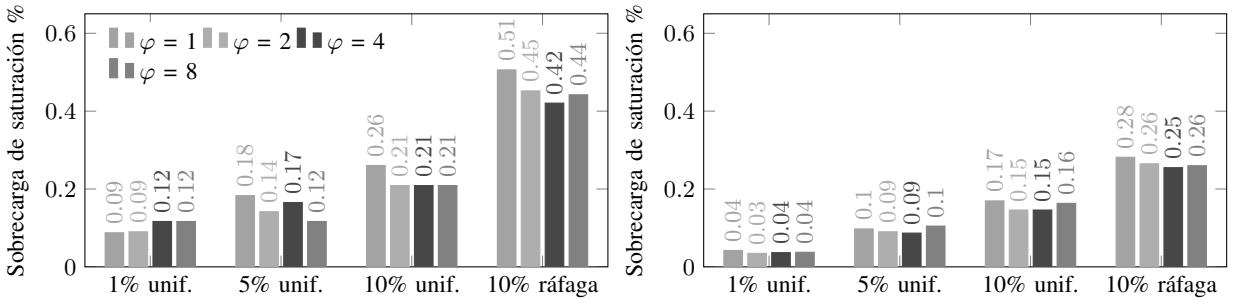


(a) 1% pérdidas uniformes. (b) 5% pérdidas uniformes. (c) 10% pérdidas uniformes. (d) 10% pérdidas a ráfagas.



(e) 1% pérdidas uniformes. (f) 5% pérdidas uniformes. (g) 10% pérdidas uniformes. (h) 10% pérdidas a ráfagas.

Fig. 4. Probabilidad de recibir y recuperar todos los paquetes enviados (a–d) y los retardos extremo a extremo (e–h) para un tamaño de generación de 256 símbolos. Las gráficas corresponden a pérdidas uniformes de 1% (a, e), 5% (b, f) y 10% (c, g), y pérdidas en ráfagas (d, h).



(a) g = 64 (b) g = 256

Fig. 5. Valores de sobrecargas de saturación para los diferentes casos de estudio contemplados en este trabajo.

V. CONCLUSIONES

En este trabajo se han revisado las técnicas de codificación de redes con especial interés en identificar aquellas que son adecuadas para las comunicaciones IoT. Una de las que más interés ha suscitado es el RLNC sistemático con solapamiento, presentada y analizada en [6]. En dicho estudio se utiliza solamente 1 símbolo codificado (redundante) por cada generación de paquetes, sin que quede claro el impacto en la eficiencia del esquema de codificación al emplear varios símbolos codificados. Para evaluar este aspecto y llevar a cabo una evaluación más exhaustiva de esta técnica se ha implementado el esquema descrito en [6], y se ha extendido su operación inicial, dando la posibilidad de modificar el número de redundancias por generación. Se ha utilizado dicha implementación para llevar a cabo una extensa campaña de simulación.

Los resultados muestran que en los puntos en los que la sobrecarga de la comunicación generada por los símbolos codificados es comparable, los esquemas con mayores grados de solapamiento presentan una mayor probabilidad de recepción de todos los símbolos generados por el transmisor (mediante recepción directa o a través de los paquetes codificados). También se ha comprobado que a mayor grado de solapamiento, menor es la latencia. La conclusión de estas observaciones es que dentro de una sobrecarga determinada, *es más eficiente aumentar el solapamiento en lugar de los símbolos redundantes por generación*.

De cara a las posibles implementaciones de este esquema en el ámbito IoT, cabe destacar que si, por motivos de limitaciones de los dispositivos o requisitos de implementación, no es posible aumentar el solapamiento, es importante tener en cuenta que cuanto mayor sea el grado del solapamiento a utilizar, mayor será la sobrecarga al aumentar las redundancias por generación.

A la hora de elegir el solapamiento y las redundancias por generación, es importante identificar la sobrecarga de saturación más baja posible. En otras palabras, encontrar aquella configuración que permita asegurar la recuperación de la mayoría de los paquetes perdidos con la menor sobrecarga posible. Como se ha visto, al aumentar el solapamiento, la sobrecarga de saturación puede aumentar en lugar de bajar. Al optimizar la sobrecarga de saturación, es posible que la latencia no sea la óptima. Por tanto, es fundamental priorizar entre la eficiencia del canal (*throughput*) y el retardo.

En el futuro se implementará el esquema sobre dispositivos reales, y se analizará el impacto de las limitaciones de éstos en la configuración del esquema de codificación. Se estudiará el efecto de considerar redes con más de un salto inalámbrico y cuál es la influencia de no disponer de un ancho de banda ilimitado. También se analizará su impacto sobre el consumo energético, que también es un parámetro fundamental en los escenarios IIoT.

AGRADECIMIENTOS

Los autoress agradecen la financiación del Programa de Doctorados Industriales de la Universidad de Cantabria (Convocatoria 2018). Los autores agradecen asimismo la financiación por parte del Gobierno de País Vasco (programa Elkartek) a través del proyecto DIGITAL (KK-2019/00095) y la financiación por parte del Gobierno de España (MINECO, MCIU, AEI, FEDER) a través de los proyectos ADVICE (TEC2015-71329-C2-1-R) y FIERCE (RTI2018-093475-A-100).

REFERENCIAS

[1] Dina Katabi, Sachin Katti, and Hariharan Rahul. *Chapter 2 - Harnessing Network Coding in Wireless Systems*, pages 39–60. Academic Press, Boston, 2012.

[2] T. Ho, R. Koetter, M. Medard, D. R. Karger, and M. Effros. The Benefits of Coding over Routing in a Randomized Setting. In *IEEE International Symposium on Information Theory, 2003. Proceedings.*, page 442, 2003.

[3] T. Ho, M. Medard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong. A Random Linear Network Coding Approach to Multicast. *IEEE Transactions on Information Theory*, 52(10):4413–4430, 2006.

[4] M. Wang and B. Li. How Practical is Network Coding? In *2006 14th IEEE International Workshop on Quality of Service*, pages 274–278, 2006.

[5] J. Heide, M. V. Pedersen, F. H. P. Fitzek, and T. Larsen. Network Coding for Mobile Devices - Systematic Binary Random Rateless Codes. In *2009 IEEE International Conference on Communications Workshops*, pages 1–6, 2009.

[6] S. Wunderlich, F. Gabriel, S. Pandi, F. H. P. Fitzek, and M. Reisslein. Caterpillar rlnc (crlnc): A Practical Finite Sliding Window RLNC approach. *IEEE Access*, 5:20183–20197, 2017.

[7] F. Gabriel, S. Wunderlich, S. Pandi, F. H. P. Fitzek, and M. Reisslein. Caterpillar RLNC With Feedback (crlnc-fb): Reducing Delay in Selective Repeat ARQ Through Coding. *IEEE Access*, 6:44787–44802, 2018.

[8] D. Stolpmann, C. Petersen, V. Eichhorn, and A. Timm-Giel. Extending On-the-fly Network Coding by Interleaving for Avionic satellite links. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–5, 2018.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Modelado basado en Ontologías para Redes de Transporte en Carreteras

Susel Fernandez, Luis Cruz-Piris, Ivan Marsa-Maestre
Departamento de Automática
Universidad de Alcalá
Escuela Politécnica Superior. Campus Universitario, Ctra. Madrid-Barcelona km. 33, 600. 28805. Alcalá de Henares. Madrid
susel.fernandez@uah.es, luis.cruz@uah.es, ivan.marsa@uah.es

***Resumen*- Los sistemas inteligentes de transporte son un conjunto de soluciones tecnológicas que se utilizan para mejorar el rendimiento y la seguridad del transporte por carretera. Un elemento crucial para el éxito de estos sistemas es que los vehículos puedan intercambiar información no solo entre ellos, sino también con otros elementos de la infraestructura vial a través de diferentes aplicaciones. Para el éxito de este intercambio de información, se necesita un marco común de conocimiento que permita la interoperabilidad. En este trabajo se propone un sistema basado en ontologías para proporcionar asistencia en la carretera, que facilite a los conductores la toma de decisiones en diferentes situaciones, teniendo en cuenta la información sobre diferentes elementos relacionados con el tráfico, como pueden ser las rutas, señales y reglas de tráfico y elementos meteorológicos.**

***Palabras Clave*- sistemas inteligentes de transporte, ontologías, redes de tráfico.**

I. Introducción y antecedentes

La continua evolución de los sistemas de transporte inteligentes ha dado paso a una nueva era de sistemas inteligentes interconectados, que ha significado un salto cuantitativo en la seguridad del transporte por carretera. Estos sistemas permiten el intercambio de información entre diferentes aplicaciones, y el análisis posterior de esta información para contribuir a mejorar la seguridad y la comodidad de los conductores en los viajes por carretera.

Debido a su alto grado de expresividad, el uso de ontologías es crucial para garantizar una mayor interoperabilidad entre los agentes de software y las diferentes aplicaciones involucradas en los sistemas de transporte inteligentes. Las ontologías proporcionan un vocabulario común en un dominio determinado y permiten definir, con diferentes niveles de formalidad, el significado de los términos y las relaciones entre ellos [1]. Las ontologías facilitan el diseño de esquemas conceptuales exhaustivos y rigurosos para permitir la comunicación y el intercambio de información entre diferentes sistemas e instituciones.

Hay algunos trabajos previos enfocados en ontologías para sistemas de transporte por carretera. En [2] se presenta una ontología para representar el tráfico en carreteras. Su objetivo fue la construcción de un sistema de información de tráfico fiable que brindara información sobre las carreteras, el tráfico y los escenarios relacionados con los vehículos en las carreteras. También proporciona formas para analizar qué tan crítica es una situación específica. Por ejemplo, una ambulancia puede necesitar conocer el estado de congestión de una zona de peaje. Solicitar esta información es crítico si la ambulancia se está dirigiendo a la escena de un accidente. En cambio, en el caso de un vehículo común que circule sin prisas por una carretera, esta información no sería crítica.

En [3] se propone una representación de alto nivel para los vehículos autónomos y su entorno. El sistema sirve de ayuda a los conductores para tomar decisiones "ilegales" pero prácticas en determinadas circunstancias (por ejemplo, cuando un automóvil dañado no permite la circulación, tomar la decisión de moverse a otro carril cruzando una línea continua para adelantar al vehículo detenido, siempre que el otro carril esté despejado). Esta representación incluye conocimiento topológico y reglas de inferencia para calcular el siguiente movimiento que un vehículo autónomo debería tomar, como asistencia al conductor.

El trabajo propuesto en [4] es un enfoque para crear una descripción genérica de la situación para sistemas avanzados de asistencia al conductor utilizando un razonamiento lógico sobre una base de conocimiento de la situación del tráfico. Contiene múltiples objetos de diferentes tipos, como vehículos y elementos de infraestructura como carreteras, carriles, intersecciones, señales de tráfico, semáforos y relaciones entre ellos. El proceso de inferencia lógica se realiza para verificar e interpretar la situación razonando sobre las reglas de tráfico.

En el trabajo en [5] se propone una ontología para la gestión del tráfico, que agrega ciertos conceptos de tráfico a la ontología general de sensores A3ME [6]. Los conceptos agregados son especializaciones de posición, distancia y clases de sensores de aceleración, y las diferentes acciones que tienen lugar en los movimientos del vehículo.

En [7] se introdujo una base de conocimientos basada en ontologías, con mapas y reglas de tráfico. Se pueden detectar las situaciones de exceso de velocidad y tomar decisiones en las intersecciones para cumplir con las reglas de tráfico. En este trabajo pero no se consideran elementos importantes como las señales de tráfico y las condiciones climatológicas.

La mayoría de los trabajos encontrados en la literatura se centran en describir situaciones de tráfico muy específicas, tales como encontrar estacionamiento, acciones de vehículos de emergencia y situaciones de intersección [8][9], comportamiento del conductor [10]. Pero ninguno de ellos es lo suficientemente general y expresivo como para abarcar una amplia variedad de situaciones de tráfico. Por lo tanto, es necesario desarrollar ontologías en el dominio del tráfico vial lo suficientemente expresivas como para describir cualquier situación de tráfico.

Este trabajo presenta un sistema basado en ontologías para la gestión del transporte por carretera, con el objetivo de proporcionar asistencia al conductor en diferentes situaciones de tráfico. La ontología desarrollada gestiona el conocimiento relacionados con los vehículos y los elementos del entorno que pueden influir en el tráfico vial, como por ejemplo los elementos de la infraestructura, las condiciones climáticas y las reglas de tráfico.

El documento está organizado de la siguiente manera. La sección II presenta la arquitectura del sistema. En la sección III se explican casos de estudio con diferentes escenarios de tráfico. Finalmente, las conclusiones y líneas de trabajo futuro se resumen en la sección IV.

II. ARQUITECTURA DEL SISTEMA

En la Figura 1. se muestra el sistema propuesto para los servicios de asistencia al conductor. En la base de la arquitectura está la ontología [11], desarrollada para el dominio específico del tráfico en carretaras. Para desarrollar el proceso de razonamiento se definen los mecanismos de inferencia lógica, utilizando el razonador *Pellet*. En el nivel superior se encuentran las distintas aplicaciones que acceden a la Información de la ontología a través de consultas SPARQL.

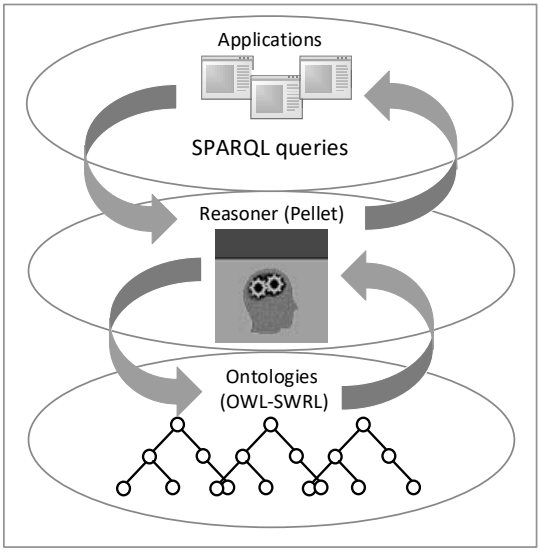


Fig. 1. Arquitectura del sistema.

A. *Descripción de la ontología*

La ontología desarrollada en el sistema permite modelar y relacionar las diferentes entidades de tráfico vial identificadas. La implementación se desarrolló en el lenguaje OWL-RDF [12] utilizando la herramienta *Protégé* [13].

Para una mejor comprensión, presentamos modelo del conocimiento de la ontología dividido en dos grupos de conceptos interrelacionados. El primer grupo contiene los elementos referentes a los vehículos, y el segundo grupo los elementos referentes a la infraestructura vial. El grupo principal está relacionado con los vehículos. Los conceptos de este grupo se muestran en la Figura 2.

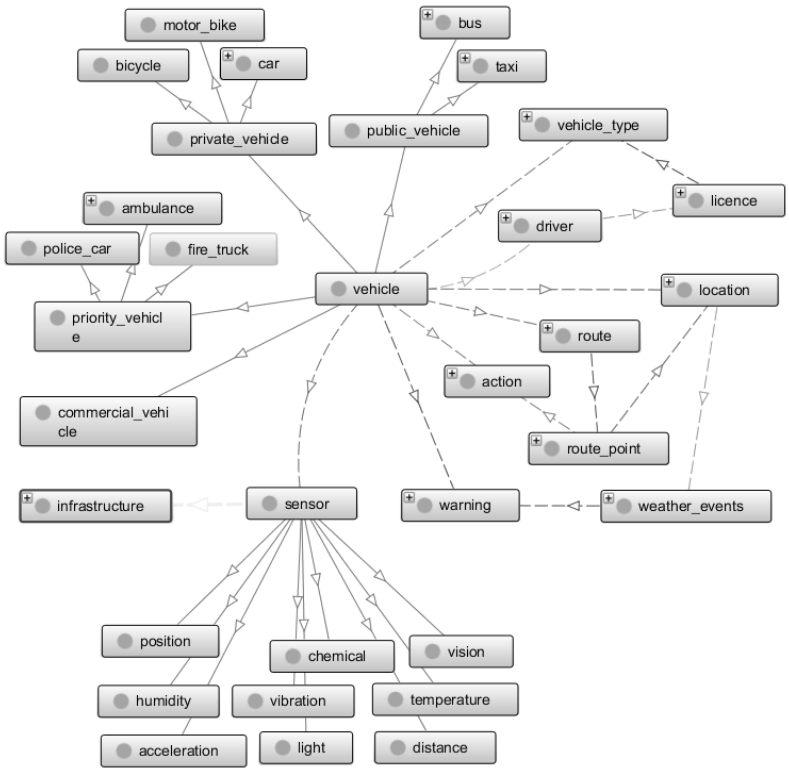


Fig. 2. Conceptos relacionados con vehículos.

La figura muestra la taxonomía de los vehículos, que se pueden clasificar en: vehículos comerciales, vehículos públicos (autobuses y taxis), vehículos privados (automóviles, bicicletas y motocicletas) y vehículos prioritarios (ambulancias, camiones de bomberos y coches de policía). Las diferentes relaciones entre los vehículos y otras entidades se definen también en este grupo. Algunas de estas entidades son: ubicación, que muestra la ubicación exacta (latitud y longitud) de un vehículo, punto de ruta o elemento de infraestructura; información sobre los conductores y los tipos de vehículos que pueden conducir según su tipo de permiso de conducción.

Una de las características fundamentales de este grupo es que cada vehículo tiene asociado un conjunto de acciones a realizar, que pueden variar según la ruta y las señales de tráfico encontradas, así como un conjunto de advertencias según la situación meteorológica en el área.

Con respecto a los sensores, estos pueden ubicarse no solo en los vehículos sino también en diferentes partes de la infraestructura, tales como puentes, carreteras, señales, etc. En la ontología se han definido varios tipos de sensores como: vibración, aceleración, humedad, temperatura, etc. La Figura 3 muestra el segundo grupo, que organiza los elementos relacionados con la infraestructura vial. En este grupo, el concepto más importante representa las carreteras.

Para una mejor gestión de las situaciones de tráfico, dividimos las carreteras en segmentos, conectados a través de intersecciones. Cada segmento contiene carriles, y en cada carril hay señales diferentes, como señales de alto o control de velocidad, semáforos o señales viales. Cada señal tiene una acción asociada a las normas de tráfico correspondientes.

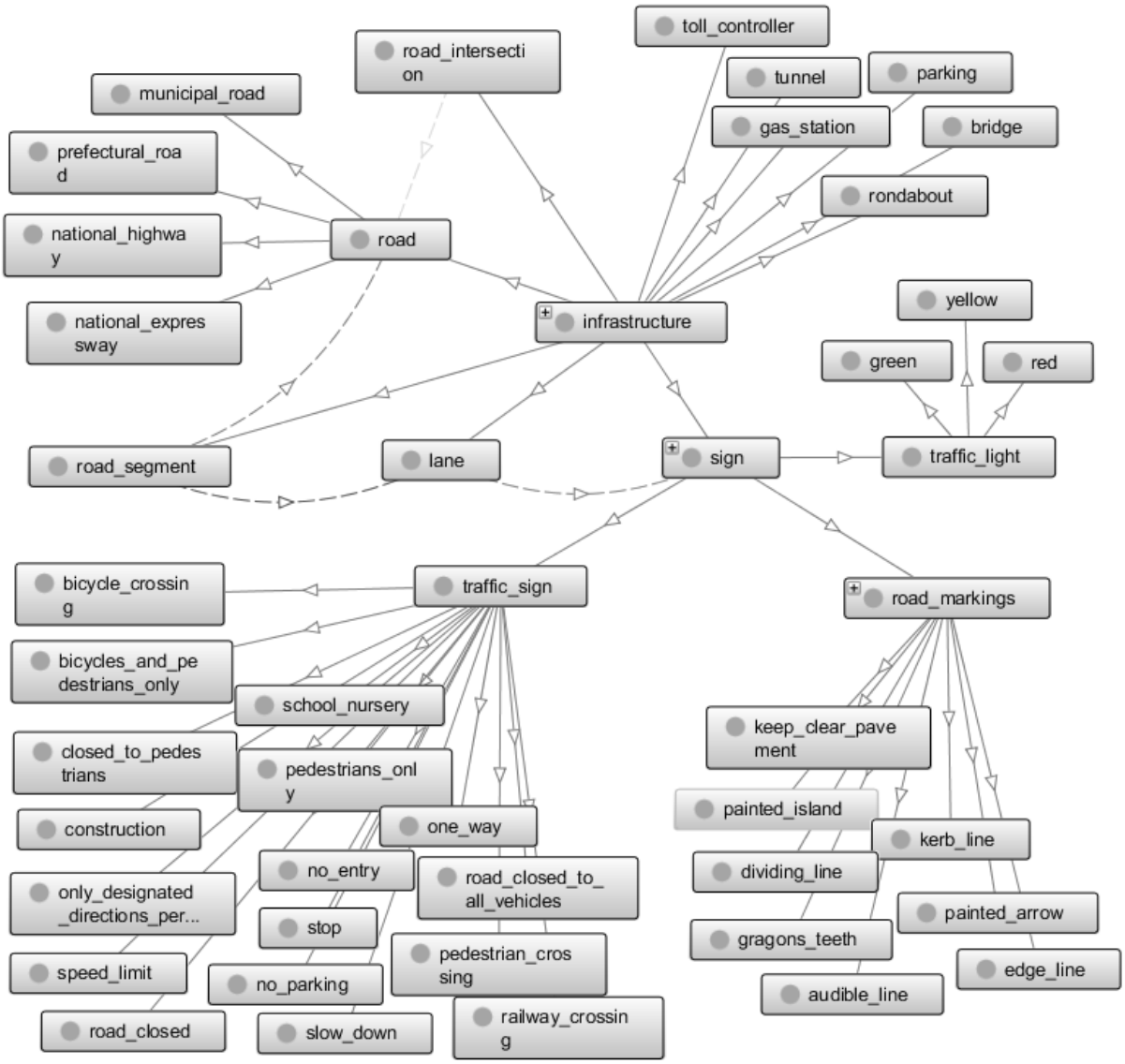


Fig. 3. Conceptos relacionados con la infraestructura de la vía.

B. Mecanismo de Razonamiento

Un aspecto crucial cuando se trabaja con ontologías es el mecanismo de razonamiento, que en la Inteligencia Artificial es simplemente la capacidad de obtener nuevo conocimiento a partir del conocimiento ya disponible mediante estrategias de inferencia. Para razonar con ontologías, se utilizan principalmente tres técnicas: razonamiento con lógica de primer orden, razonamiento con lógica de descripción y razonamiento con reglas.

En este trabajo utilizamos el razonador *Pellet* [14], que es una herramienta para razonar con ontologías, que admite los tres tipos de razonamiento. *Pellet* se implementa en *Java*; está disponible de forma gratuita y permite verificar la consistencia de la ontología.

Las reglas de razonamiento en la ontología del tráfico se han desarrollado utilizando el Lenguaje de Reglas de la Web Semántica (*SWRL*) [15]. En esta ontología, las reglas *SWRL* se utilizan para definir diferentes regulaciones de tráfico y las diferentes acciones que un conductor puede tomar, de acuerdo con la situación actual de la carretera. Entre las reglas definidas en la ontología se encuentran, por ejemplo, aquellas que permiten al razonador inferir la transitividad con respecto a la ubicación de los elementos de tráfico. Esto significa que si un elemento de tráfico (e.g un vehículo o señal de tráfico) está ubicado en un carril, y ese carril está ubicado en un segmento de la carretera, entonces el elemento de tráfico también se encuentra en esa carretera. Otros conjuntos de reglas definidas están dirigidas a determinar la acción que el conductor de un vehículo debería ejecutar dependiendo de determinadas circunstancias, por ejemplo, cuando se circula en el mismo carril que un vehículo prioritario en situación de emergencia o cuando nos encontramos en el mismo segmento que una señal de tráfico concreta.

C. Consultas a la ontología

En el nivel superior del sistema, las diferentes aplicaciones consultan la información almacenada en la ontología para llevar a cabo su ejecución. Como lenguaje de consulta ontológica se ha utilizado *SPARQL* [16].

Un ejemplo simple de consulta *SPARQL* con la ontología sería obtener la lista de vehículos que se encuentran en una ruta determinada.

Otro ejemplo de consulta devolvería todos los puntos asociados con una ruta de un vehículo y la acción que debe realizarse para ir de un punto a otro, considerando su ubicación en el mapa. Esta consulta en concreto resulta muy simple, teniendo en cuenta que por dieño, en la ontología cada punto de ruta está relacionado con el siguiente a través de una acción específica (girar a la izquierda, girar a la derecha o seguir recto), y cada acción depende del tipo de relación (*isAtNorthOf*, *isAtSouthOf*, *isAtWestOf*, *isAtEastOf*) que conecta los segmentos en los que se encuentran los puntos de la ruta. Se ha definido una regla *SWRL* en la ontología que asocia una acción o movimiento para trasladarse de un punto al siguiente punto de la ruta según la relación entre los segmentos en los que se encuentra cada punto de ruta.

III. EXPERIMENTOS

Hemos realizado pruebas de la expresividad de la ontología, realizando consultas para situaciones de tráfico simuladas. En esta sección presentamos parte de los experimentos realizados en diversas situaciones de tráfico simples. El escenario de tráfico definido para los experimentos consta de varias carreteras y sus intersecciones. Cada carretera se divide en varios segmentos, con dos carriles cada uno.

De cada vehículo se conoce su ubicación, velocidad y la ruta que desea seguir. Cada ruta tiene un conjunto de puntos y cada punto de ruta tiene una ubicación (latitud y longitud), así como la información sobre el siguiente punto de la ruta. Se definen previamente una serie de situaciones climatológicas específicas en diferentes puntos del mapa. Consultando con la ontología podemos saber la próxima acción que debe tomar el conductor, dada la posición del vehículo, la ruta elegida y las señales de tráfico ubicadas a lo largo de la ruta. También podemos recibir recomendaciones con respecto a la situación del clima a lo largo de la ruta.

En este trabajo, hemos llamado *Desired_Action* a la acción que el conductor desea llevar a cabo para moverse de un punto al siguiente a lo largo de la ruta, independientemente de las señales de tráfico; *Next_Action* es la acción que el conductor realmente debería tomar en cada punto considerando únicamente las señales de tráfico correspondientes. Hemos definido una serie de advertencias para diferentes situaciones climatológicas que se pueden encontrar en la ruta, como lluvia, nieve, niebla, viento, etc. Cada una de estas advertencias está asociada con una serie de recomendaciones para facilitar la circulación en estas condiciones.

A partir de la posición del vehículo en cada punto, la ubicación de las señales de tráfico y la ruta, las acciones se deducen mediante el razonamiento aplicando diferentes reglas de la ontología, en cada uno de los pasos que se describen a continuación:

1. Localizar en qué segmento de la carretera está ubicado el vehículo y cuál es el siguiente punto de la ruta. Esto se hace considerando la posición (latitud y longitud) y las coordenadas de los puntos de inicio y final de cada segmento.
2. Elegir la acción deseada para ir de un punto a otro dependiendo del tipo de conexión entre los segmentos en los que se encuentran los puntos. Por ejemplo, si el vehículo está en segmento1 y el siguiente punto de la ruta está en segmento2; el segmento2 está ubicado al este del segmento1, entonces la acción que debe tomar el vehículo para ir del punto1 al punto2 es girar a la derecha.
3. Elegir la siguiente acción a ejecutar por el vehículo, teniendo en cuenta únicamente las señales de tráfico. Esta es la misma acción asociada con la siguiente señal de tráfico ubicada en el segmento donde se encuentra el vehículo. Por ejemplo, si el vehículo está en un segmento con una señal de *Stop*, la acción que debe tomar el conductor es detenerse.
4. Si hay alguna condición climática especial en el siguiente segmento de la ruta, entonces se le asigna al vehículo la advertencia correspondiente a esa condición climática.

Los experimentos se realizaron en simulación con 50 vehículos y 20 rutas. Para cada ruta se definieron distintos escenarios de tráfico específicos variando diversos factores como el nivel de congestión, el estado de los semáforos y las condiciones climatológicas en distintos puntos.

Para cada vehículo se definieron a priori las distintas acciones a tomar a lo largo de los diferentes puntos de la ruta y luego se compararon estos resultados con los obtenidos por el sistema para evaluar la expresividad de la ontología. Las medidas utilizadas para evaluar la expresividad de la ontología fueron la *Precisión* y el *Recall*, que representan el nivel de exactitud y completitud de los resutados respectivamente.

Dado un conjunto de referencia *R* y un conjunto resultante *A*, la *Precisión* es un indicador de la exactitud y se define como la razón entre el número de instancias correctas y aquellas que el algoritmo considera que pertenecen al conjunto de instancias correctas (ecuación 1).

$$Precision\ (A, R) = \frac{|R\ \cap A|}{|A|} \quad (1)$$

El *Recall* describe la completitud y se define como la razón entre el número de instancias correctas y todas las instancias que realmente pertenecen an conjunto de instancias correctas (ecuación 2).

$$Recall\ (A, R) = \frac{|R\ \cap A|}{|R|} \quad (2)$$

La Tabla 1 presenta los resultados de los experimentos realizados sobre la expresividad de la ontología en términos de *Precisión* y *Recall*. La tabla muestra que para un total de 50 vehículos y 20 rutas, la media de los valores de *Precisión* obtenidos fue de 0,95 mientras que la media del *Recall* fue de 0,98, lo que demuestra que la ontología es válida para proporcionar la información necesaria para la toma de decisiones en los distintos escenarios de tráfico evaluados.

Tabla 1
RESULTADOS DE LA EXPRESIVIDAD DE LA ONTOLOGÍA EN TERMINOS DE PRECISION Y RECALL

Nº vehículos	***Nº Rutas***	***Media Precisión***	***Media Recall***
50	20	0,95	0,98

En general, los resultados muestran que la ontología es suficientemente expresiva en términos de señales de tráfico, rutas y reglas de tráfico. La ontología permite inferir el conocimiento relacionado con el clima a partir de datos de sensores, sin embargo hay sensores de infraestructura que miden otros datos útiles, como el flujo de multitudes y el flujo de tráfico, que aún no se han tenido en cuenta en la ontología. El procesamiento de los datos de esos sensores mejoraría el trabajo en la optimización de la ruta. Consideramos también que para obtener mejores resultados de cara a la mejora de la conducción se necesita ampliar la ontología incorporando una serie de conceptos y relaciones que permitan tener en cuenta otros factores importantes como el comportamiento de los conductores.

IV. CONCLUSIONES

En este documento se presenta un sistema basado en ontologías para la gestión del transporte por carretera. El objetivo principal de este trabajo es proporcionar asistencia al conductor en diferentes situaciones de tráfico, teniendo en cuenta la ruta, el clima y las reglas de tráfico.

La expresividad de la ontología ha sido probada a través de consultas en diferentes situaciones de tráfico que involucran varias señales y las reglas de tráfico. Los resultados de los escenarios probados han sido satisfactorios, pero aún es necesario enriquecer la ontología para abarcar y relacionar más conocimiento. Como trabajo futuro tenemos la intención de continuar mejorando la expresividad de la ontología, con el procesamiento de datos de más sensores ubicados en la infraestructura, por ejemplo, en puentes, carreteras, ríos, túneles. Esos sensores podrían medir el flujo de multitudes, el flujo de tráfico y muchos otros parámetros que son importantes en la optimización del tráfico. También pretendemos mejorar la expresividad de la ontología, agregando información sobre el comportamiento de los conductores, debido a su importancia en todo el proceso de conducción en carretera. Finalmente, planeamos agregar reglas SWRL que describan múltiples mecanismos de negociación automática entre agentes en diferentes escenarios de tráfico.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto MOON-MOdelado basado en ONtologías para redes complejas. CCG2018-EXP-041, de la Universidad de Alcalá.

REFERENCIAS

[1] Studer, R; Benjamins, R.; Fensel, D. Knowledge Engineering: Principles and Methods. In: Data and Knowledge Engineering, 1998, v.25, n.1-2, pp.161-197

[2] Sérgio Gorender, Ícaro Silva. AN ONTOLOGY FOR A FAULT TOLERANT TRAFFIC INFORMATION SYSTEM. 22nd International Congress of Mechanical Engineering (COBEM 2013). November 3-7, 2013, Ribeirão Preto, SP, Brazil

[3] Evangeline Pollard, Philippe Morignot, Fawzi Nashashibi. An ontology-based model to determine the automation level of an automated vehicle for co-driving. FUSION 2013: 596-603

[4] Michael Hülsen, J. Marius Zöllner, Christian Weiss. Traffic Intersection Situation Description Ontology for Advanced Driver Assistance. In 2011 IEEE Intelligent Vehicles Symposium (IV) Baden-Baden, Germany, June 5-9, 2011

[5] A.J. Bermejo, J. Villadangos, J. J. Astrain, A. Cordoba. Ontology Based Road Traffic Management. Intelligent Distributed Computing VI, G. Fortino et al. eds., SCI 446, pp. 103-108.

[6] Herzog, A.; Jacobi, D.; Buchmann, A. A3ME-An Agent-Based Middleware Approach for Mixed Mode Environments. In Proceeding of Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2008), Valencia, Spain, 29 September-4 October 2008; pp. 191-196.

[7] Zhao, L., Ichise, R., Mita, S., & Sasaki, Y. Ontologies for Advanced Driver Assistance Systems.

[8] Fernandez, S., & Ito, T. (2016, September). Using SSN ontology for automatic traffic light settings on inteligent transportation systems. In 2016 IEEE International Conference on Agents (ICA) (pp. 106-107). IEEE.

[9] Cruz-Piris, L., Rivera, D., Fernandez, S., & Marsa-Maestre, I. (2018). Optimized sensor network and multi-agent decision support for smart traffic light management. Sensors, 18(2), 435.

[10] Fernandez, S., & Ito, T. (2015, October). Driver behavior model based on ontology for intelligent transportation systems. In 2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA) (pp. 227-231). IEEE.

[11] Fernandez, S., Ito, T., & Hadfi, R. (2016). Architecture for intelligent transportation system based in a general traffic ontology. In Computer and Information Science 2015 (pp. 43-55). Springer, Cham.

[12] M. Dean, and G. Schreiber, OWL Web Ontology Language Reference. http: //www.w3.org/TR/2004/REC-owl-ref-20040210/; 2004.

[13] Protégé: http://protege.stanford.edu/

[14] Pellet http://clarkparsia.com/pellet/
[15] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, Submission to W3C, May 2004 http://www.w3.org/Submission/SWRL/
[16] SPARQL http://sparql.org/

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Nintendo DS para tareas de gestión de red en campo

Andrés Martinez, Antonio Estepa, Rafael Estepa, Vicente Mayor.
Departamento Ingeniería Telemática,
Universidad Sevilla
C/ Avenida de los descubrimientos s/n. 41092 Sevilla.
andresmargar98@gmail.com, {aestepa,rafa,vmayor}@trajano.us.es.

***Resumen*- Los dispositivos móviles que se utilizan en tareas de gestión suelen utilizar sistemas operativos de uso común y propósito general (p.ej. Windows, Android) y por ello son susceptibles de portar malware y transmitirlo a los elementos gestionados. En este artículo proponemos usar la consola de juegos Nintendo DS como consola de gestión de red.**

Hemos desarrollado librerías para poder ejecutar aplicaciones de gestión basadas en el protocolo SNMP. Utilizando estas librerías, hemos desarrollado una aplicación de demostración que permite realizar tareas simples de gestión a través de esta consola de juegos. Este trabajo es aún preliminar y nuestro objetivo es añadir nuevas funcionalidades para conseguir una consola gestión segura de propósito específico usable en la industria.

***Palabras Clave*- gestión, SNMP, IoT**

I. INTRODUCCIÓN

La gestión de redes de datos extensas, tanto en el mundo industrial (Industrial Control Systems o ICS) como en el mundo corporativo (LAN tradicionales) suele realizarse de forma centralizada desde un centro de operación de red o NOC provisto de múltiples herramientas software que son utilizadas por los gestores para distintas tareas de gestión. No obstante, en ocasiones es necesario el desplazamiento físico de técnicos a la red de campo para realizar labores de configuración inicial y puesta en marcha de nuevos nodos, auditorías de inventariado o en la resolución de incidentes relacionados con la seguridad o con fallos.

En estos casos, por simple comodidad o por falta de alternativas, el operador suele llevar un equipo portátil con un sistema operativo de propósito general (p.ej. laptop con Windows o Linux) con el que se conecta a los nodos de red de forma cableada o inalámbrica. Además de la conexión vía CLI con el nodo gestionado, existen infinidad de herramientas de utilidad en las diferentes tareas de gestión [1], e incluso suites integradas de gestión (p.ej. Solarwinds) que los operarios pueden instalar en sus terminales portátiles. Sin embargo, cada vez es más común el uso de programas o utilidades de gestión en dispositivos móviles (smartphones o tablets) que usan iOS y, sobre todo, Android (p.ej. PingTools, Network Monitor, etc.).

Sin embargo, cualquier equipo con un sistema operativo común (p.ej. Windows) está expuesto a innumerables amenazas de ciberseguridad que aprovechan las vulnerabilidades tanto del propio sistema operativo como de las aplicaciones que éste ejecuta (p.ej. Office, Adobe, navegador, etc..)[2]. Los ataques han crecido exponencialmente en la última década en todos los sistemas operativos y en todos los ámbitos de aplicación (corporativo [3], ICS[4], IoT[5]) y los dispositivos móviles (iOS y, sobre todo Android) no son ninguna excepción [6]. Según [7] el malware en Android ha crecido de forma exponencial en los últimos años y ya representa el 6% del total del malware (donde Windows representa más del 75%)

Una forma pasiva de protección es que el hardware y software utilice diseños propietarios y cerrados. Esta propia singularidad reduce la exposición a ataques y originalmente se daba (cada vez menos) en los entornos industriales. Además, las buenas prácticas en seguridad recomiendan la segregación tanto software como hardware de todos aquellos sistemas de gestión que estén relacionados con las tareas de monitorización de la seguridad.

En este artículo presentamos una idea novedosa: el uso de la videoconsola Nintendo DS como dispositivo para realizar, en campo, tareas gestión de red. Este dispositivo presenta varias ventajas sobre otros dispositivos móviles tradicionales: poco peso y tamaño reducido, menor exposición a ataques por tratarse de un entorno cerrado y propietario (aunque se ha reportado algún caso de malware [11], su incidencia es despreciable en comparación con los sistemas operativos de uso común), alta durabilidad, una batería con gran duración y un precio reducido. Como principal inconveniente, el desarrollo de aplicaciones en este tipo de entorno tiene un

mayor coste ya que apenas cuenta con librerías o APIs públicas que puedan ser reutilizadas. Otro potencial inconveniente menor es que el interfaz hombre-máquina se encuentra limitado por el interfaz de usuario de la propia consola, aunque esto no ha representado ningún inconveniente hasta la fecha e incluso resulta atractivo para aquellas personas acostumbradas a su uso. En este trabajo preliminar hemos programado librerías en C que implementan el protocolo de gestión SNMP y una aplicación sencilla de demostración, pero este trabajo sigue en curso y en el futuro pensamos aumentar su funcionalidad y el número de protocolos soportados, convirtiendo este dispositivo móvil en una consola de gestión usable en la industria.

II. STANDARDS EN SISTEMAS DE GESTIÓN

La RFC 6632 ofrece una panorámica sobre los diferentes standards utilizados en la gestión de red y su encuadre en las áreas de gestión FCAPS (fallos, configuración, contabilidad, prestaciones y seguridad). Esto se refleja en la tabla I.

Tabla I
STANDARDS CLAVE EN LA GESTIÓN SEGÚN IETF RFC662

Protocolo	Utilidad principal (en áreas FCAPS)	Grado adopción	Consumo recursos
SNMP *(RFC1157)*	Uso genérico (cualquier área FCAPS).	Alto	Bajo
SYSLOG *(RFC5424)*	Distribución de eventos (logs) (FAS)	Alto	Bajo
NETCONF *(RFC6241)*	Monitorización y control de configuraciones (C)	Bajo	Alto
IPFIX *(RFC5101)*	Monitorización de flujos de tráfico (PS)	Medio/bajo	Alto

Además, existen otros standards usados en funciones específicas (p.ej. protocolos AAA, DHCP, etc..) o en dominios de aplicación (p.ej. LwM2M o CoMI en IoT). Debido a que SNMP (Simple Network Management Protocol) es un protocolo muy maduro y ampliamente adoptado por todos los fabricantes de electrónica de red, y que éste puede ser utilizado para la realización de tareas de gestión en cualquier ámbito de FCAPS, hemos seleccionado éste como pilar básico del sistema de gestión implementado, dejando para futuras ampliaciones la incorporación de nuevos protocolos.

A. Arquitectura de SNMP

SNMP es un protocolo sencillo utilizado para enviar comandos a un agente (i.e. nodo gestionado) y recibir sus respuestas, así como para recibir notificaciones (TRAP) generadas por el agente. La figura 2 ilustra las interacciones del protocolo, que en su versión más básica cuenta sólo con 5 A-PDUs que son transportadas por el protocolo UDP. Las peticiones y respuestas tienen la misma sintaxis y llevan una secuencia de tuplas de tipo (identificador de objeto, valor).

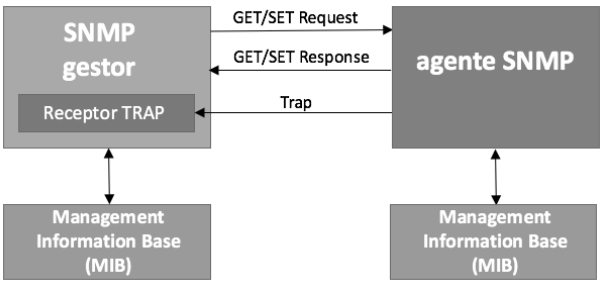


Fig. 1. Elementos en el esquema de Gestión SNMP.

SNMP se apoya en un modelo de datos cuyo marco normativo esta estandarizado (Structure of Management Information o SMI) en la RFC1155. El modelado de objetos de gestión se reduce a objetos de tipo escalar o tablas de tipos de datos predefinidos. Siguiendo esta norma, los fabricantes definen los catálogos de objetos de gestión (Management Information Base o MIB) que han implementado en sus dispositivos. Las tareas de gestión se realizan a través de la lectura o modificación de estos objetos. Podemos ver entonces una MIB como una especie de diccionario que nos permite identificar, conocer y organizar a los objetos de gestión que ofrecería un agente que implementase dicha MIB. Existen miles de catálogos (MIBs) con objetos standard definidos por el ietf o por los propios fabricantes (ver http://www.mibdepot.com).

A diferencia de otros protocolos de aplicación (p.ej. HTTP), SNMP no es un protocolo textual y utiliza una sintaxis de transferencia binaria basada en ASN.1. En particular la codificación Basic Encoding Rules (BER). Por ello, en la programación de aplicaciones es necesario el uso de funciones o métodos de codificación / decodificación ASN.1 BER que procesen adecuadamente las A-PDUs recibidas o a enviar a través de los sockets udp.

III. NINTENDO DS

La consola Nintendo DS es un dispositivo de propósito específico (entretenimiento) con recursos algo inferiores a la mayoría de ordenadores o smartphones actuales pero superiores a la mayoría de dispositivos IoT. Según [8] la Nintendo DS posee el siguiente hardware:

- Memorias: una memoria principal RAM de 4MB y una memoria destinada a gráficos (VRAM) de unos 656KB. Como memorias secundarias, dispone de una ROM de 256KB que almacena el firmware de la consola, una RAM de 8KB para el encolado de las tramas enviadas y recibidas por WiFi, y una RAM de 248KB para la representación de gráficos en 3D.
- Pantallas: posee 2 pantallas, una superior y otra inferior de 256x192 píxeles. Cada pantalla tiene su propio controlador que lee la memoria de video y los registros hardware para dibujar gráficos.
- Sonido: posee un micrófono capaz de grabar en mono a 8/16 bits, y 2 altavoces con 16 canales, audio estéreo.
- Controles: botones A, B, X, Y, Start, Select, cruceta direccional, gatillos L y R, y botón de encendido/apagado. Además, la pantalla inferior posee una pantalla táctil capaz de detectar 1 pulsación a la vez.

- Tarjeta de red WiFi: con un transmisor y receptor que implementa el estándar IEEE 802.11b en el firmware que permite el envío y recepción de tramas.
- Slots: son las ranuras donde se introducen los cartuchos con los programas. Tenemos el Slot-1 donde introducimos una FlashCard que contiene los programas que queremos ejecutar. El Slot-2 se destina a ejecutar juegos de su consola antecesora y la conexión de periféricos especiales: por ejemplo, existe uno que permite que la consola vibre.
- Procesadores. La Nintendo DS tiene 2 procesadores de 32 bits: un ARMv7 a 33 MHz y un ARMv9 a 66 MHz que funcionan de forma conjunta. Ambos procesadores tienen acceso a la memoria RAM, y presentan una cola de mensajes FIFO para facilitar la comunicación entre ellos, especialmente en el envío y recepción de señales. Ambos procesadores se destinan a funciones diferentes para repartir la carga de procesamiento de los periféricos. El ARMv7 se encarga del procesado del sonido, leer el estado de los botones, la pantalla táctil, lectura de grabaciones realizadas en el micrófono, control del reloj en tiempo real, y del envío y recepción de tramas usando la tarjeta WiFi implementada en la consola. Por otro lado, el ARMv9 se encarga de ejecutar la lógica principal de los programas y la configuración de los registros hardware para la visualización de gráficos.

La programación en la consola Nintendo DS requiere el uso de un kit de desarrollo. Existen dos opciones en la actualidad: uno (el oficial) caro y dirigido a la industria del sector de los videojuegos (https://developer.nintendo.com/tools), y otro (no oficial) desarrollado por particulares y gratuito llamado **devkitpro** (https://devkitpro.org/) que ha sido el utilizado en este proyecto. Este kit puede instalarse en los sistemas operativos más conocidos: Windows, la mayoría de distribuciones de Linux y Mac. El kit incluye:

- Compilador C/C++ para arquitectura ARM de 32bits y depurador GDB.
- Librerías:
  - *Libnds*: contiene funciones que modifican los registros hardware y nos permite programar la consola con más facilidad. Nos facilita la visualización de gráficos, la comunicación con los periféricos y la gestión del ARMv7. Contiene un programa por defecto para el ARMv7, de forma que nosotros como programadores nos preocuparemos de desarrollar solamente la parte del ARMv9, y la librería se encargará de comunicar ambos procesadores.
  - *DsWifi*: nos proporciona una interfaz para la conexión con puntos de acceso WiFi y una implementación de la librería socket de C estándar, para mandar datagramas hasta nivel de transporte (TCP/UDP).
  - *Filesystem* y FAT: nos permite acceder a la tarjeta microSD de la FlashCard para leer y escribir archivos en ella.
  - *Maxmod*: librería de reproducción de audio.
- Emuladores de la consola para la realización de pruebas.

Como se puede apreciar, no existen librerías SNMP ni de codificación BER. Por lo tanto, es necesario la programación de estas librerías, lo cual constituye la principal aportación de este trabajo. Se puede encontrar más información y ejemplos de programación en [9]y [10].

IV. APLICACIÓN DE GESTIÓN DESARROLLADA

Se ha programado en C una librería de codificación /decodificación binaria BER así como las funciones de creación y recepción de las cinco APDUS del protocolo. Esto ha producido 18 funciones, y 4.000 líneas de código disponibles en github (https://github.com/amg98/SNMPDS/).

Usando las librerías desarrolladas, se ha realizado una aplicación de gestión sencilla con las siguientes funcionalidades.

- Lectura/escritura de objetos escalares de MIB-II
- Lectura de tablas y escritura de objetos columna en la tabla Interfaces (MIB-II).
- Programación de alarmas por umbrales de tráfico o cambio de estado en puertos usando la MIB de RMON.
- Recepción y procesado de notificaciones (TRAP) generadas por los agentes.
- Generación de alarma con vibración si se recibe una tormenta de traps.
- Envío de los traps recibidos y procesados en formato html, junto con una nota de voz a un servidor mediante tftp.

V. PRUEBAS

La aplicación desarrollada se ha probado en el laboratorio del Departamento de Ingeniería Telemática de la Universidad de Sevilla. El nodo gestionado en las pruebas ha sido un conmutador HP Procurve 2650 que implementa la MIB-II y RMON entre otras. La comunicación entre la consola y el nodo gestionado se ha realizado a través de un punto de acceso tal y como se ilustra en la figura 2.

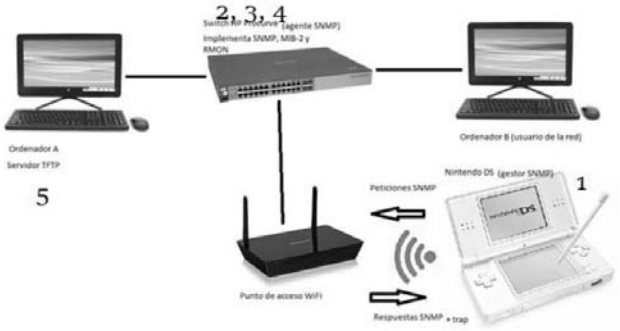


Fig. 2. Esquema de uso en el laboratorio.

En las siguientes figuras se muestra el contenido de las pantallas superior e inferior de la consola durante diferentes pruebas de funcionamiento.

En primer lugar, es necesario configurar inicialmente la consola, estableciendo la IP del agente SNMP con el que se comunicará y del servidor TFTP donde se mandarán los TRAPs. Esto se realiza mediante el interfaz de configuración mostrado en la figura 3 (izquierda).

Fig.3. Pantalla de configuración inicial (izquierda) y de información del agente (derecha)

Una vez configurada la consola, puede comenzar a realizar tareas de gestión. En la figura 3 (derecha) vemos una pantalla donde se muestran las características principales del agente gestionado mediante consultas SNMP a los objetos del grupo system de la MIB-II. Nótese que la aplicación muestra el símbolo de un lápiz en aquellos objetos que sean editables (i.e. se puede escribir un nuevo valor).

En la figura 4 se muestra la lectura de un objeto de tipo tabla. En particular la tabla una fila de la tabla de Interfaces del agente (también definida en la MIB-II). Por ejemplo, si cambiamos el estado operativo de un puerto (objeto *ifOperStatus*), el puerto se apagará.

Fig.4. Fila de la tabla Interfaces del agente.

El resto de características de la aplicación desarrollada (p.ej. la programación de alarmas, la recepción de traps, etc..) pueden ser vistas a través del siguiente vídeo de demostración: https://youtu.be/O8U2l7aX62c

VI. TRABAJOS EN CURSO

El trabajo presentado es un primer paso en un proyecto que está aún en curso. Una de las primeras ampliaciones a realizar es la incorporación de versiones de SNMP más seguras (p.ej. SNMPv3). En el futuro, queremos avanzar añadiendo nuevas funcionalidades de gestión y nuevos standards de forma que sea también utilizable en entornos industriales y en entornos IoT a través de las siguientes líneas de continuación.

Tabla I
LÍNEAS DE CONTINUACIÓN

Tipo	Descripción
Ampliación de funcionalidades	• Autodescubrimiento de agentes • Carga y navegación de MIBs • Creación de interface CLI • Auto-inventariado • Vibración con la recepción de traps
Nuevos protocolos	• Syslog para recepción de logs • Versión 2 y 3 de SNMP • Protocolo netconf y repositorios YANG • Lado recolector de IPFIX / netflow • Protocolos REST y coap • Protocolos LwM2M y CoMi

REFERENCIAS

[1] url: http://www.slac.stanford.edu/xorg/nmtf/nmtf-tools.html

[2] url: https://www.microsoft.com/en-us/wdsi/threats

[3] CCN-CERT IA-09/18. “Ciberamenazas y Tendencias. Edición 2018”. Centro Criptológico Nacional.

[4] Wolfgang Schwab. “The State of Industrial Cybersecurity 2018”. Kaspersky.

[5] Cisco, "Annual Security Report, 2018", url: https://www.cisco.com/c/dam/m/hu_hu/campaigns/security-hub/pdf/acr-2018.pdf .

[6] CCN-CERT IA-04/19. “Informe Anual 2018 Dispositivos y comunicaciones móviles”. Enero 2018. Centro Criptológico Nacional

[7] AV-TEST Security report. Disponible en: https://www.av-test.org/fileadmin/pdf/security_report/AV-TEST_Security_Report_2016-2017.pdf

[8] url. https://problemkaputt.de/gbatek.htm

[9] url. https://problemkaputt.de/gbatek.htm

[10] url. https://problemkaputt.de/gbatek.htm

[11] url: https://venturebeat.com/2010/07/31/live-demos-of-hacking-the-nintendo-ds-and-the-wii-to-spread-malware/


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Hands-on Data Transfer Nodes: implementation and performance evaluation

Jorge Sasiain, Eduardo Jacob, Jasone Astorga, Juanjo Unzilla
Department of Communications Engineering,
University of the Basque Country UPV/EHU.
48013 Bilbao, Spain.
jorge.sasiain@ehu.eus, eduardo.jacob@ehu.eus, jasone.astorga@ehu.eus, juanjo.unzilla@ehu.eus

***Abstract*—A Data Transfer Node (DTN) is a Linux server comprised of high quality hardware and software components specifically tuned for maximum performance on high-speed data transfers. In this paper, we first explore the benefits and use cases of DTNs in the current world. Then, we analyze the requirements of a DTN when it comes to hardware components and proper tuning of the different subsystems involved, in order to achieve the optimal level of performance. Finally, we introduce the testbed scenario that we have used to experiment with DTNs, and present the obtained results and conclusions in regards to the impact of the different configurations and tuning possibilities studied.**

***Keywords*—Dedicated networks, Applications and services, DTN, Data Transfer Node, Data transfer, Performance, Research networks**

I. INTRODUCTION

A Data Transfer Node (DTN) is, according to ESnet [1], a Linux server that includes high quality hardware components and that is carefully configured and tuned in order to maximize performance when it comes to reading, writing, and transferring high amounts of data across the Internet. A typical DTN workload implies read and write operations of very large files, and their transmission across a high-speed WAN network, with all these tasks having critical performance requirements.

As a result, for a DTN to perform optimally, several components are required. A DTN should have access to high-speed storage, be it a local disk or a connection to a local storage infrastructure. It should also feature high speed network interfaces, typically between 10 Gbps and 100 Gbps depending on the specific implementation, and should be complemented by a network infrastructure that is able to support such bandwidth capabilities.

General-purpose tasks, such as web browsers and email clients, are excluded from a DTN in order to mitigate security risks and avoid the inclusion of security components that could negatively impact the DTN's performance. On the other hand, a DTN could include or be complemented by tools specifically designed to support the execution of its usual tasks, like high-speed data transfer tools used for remote transfer and network performance monitoring tools, such as a perfSONAR deployment [2].

II. BACKGROUND

The concept of DTN is first introduced by ESnet circa 2010, in the context of a network model known as Science DMZ. The Science DMZ term itself comes from the DMZ networks (demilitarized zone), which refers to an organization's subnetwork, located in the network's perimeter, in which the services that need to be accesible from the outside are exposed. The goal of a DMZ is so that its security configuration does not compromise the internal part the network.

The Science DMZ adapts this concept of DMZ to a subnetwork that, also being isolated from the general-purpose network, is specifically focused on supporting scientific applications with high performance requirements that demand the movement of very large amounts of scientific data across separate scientific networks. Support for these kind of applications over a general-purpose network would lead to very poor performance, as they would clash with the requirements of the organization's usual network traffic, and would be further degraded by required security mechanisms such as firewall processing as well as network equipment with potentially insufficient capacity to handle the high-speed bursts of a DTN.

Examples of scientific disciplines that can benefit from a Science DMZ network model and the high performance offered by DTNs are geophysical sciences, genomics, bioinformatics, and precision medicine [3]. Nowadays, several R&E (Research and Education) network providers around the world, such as ESnet, Geant and Internet2 provide these kinds of high-performance dedicated networks in order to satisfy the highly intensive data requirements of the scientific research scope [4].

In the context of a Science DMZ, a DTN is the component that, connected through a high-performance

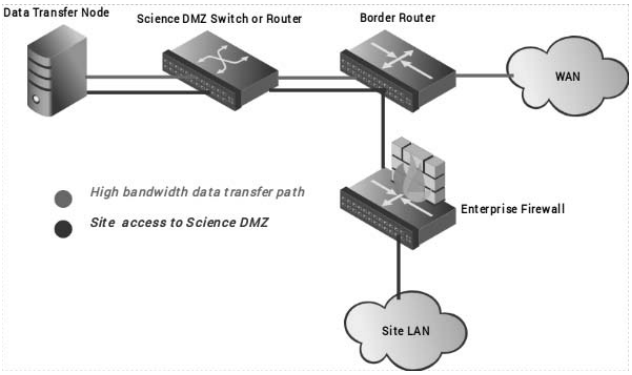


Fig. 1. Basic network diagram of a Science DMZ.

switch or router, is responsible of moving the scientific data from and to other scientific networks. The data transfers of a DTN only traverse the high-speed switch or router of the Science DMZ and the organization's border router, both of which are prepared to support said traffic. It is, however, completely isolated from the rest of the network and its delimiting firewall or router. The security mechanisms of a DTN are implemented with ACLs rather than with a dedicated firewall that could negatively impact performance. Figure 1 shows the basic network diagram of a Science DMZ.

Other than their home in Science DMZ networks built for scientific applications, DTNs also have a place in any other kind of scenario where an efficient movement and management of high amounts of data is required, be it external traffic against a remote server or internal traffic. Examples of other disciplines where DTNs are being used are supercomputation, Big Data, AI, Deep Learning, software-defined networking (SDN) and network functions virtualization (NFV). In regards to these examples, Hong, Wontaek, et al. [4] propose the use of DTNs to enhance the transfer of virual machine images in a cloud environment, while Basu, Kashinath, et al. [5] explore how the use of SDN could impact performance in a Science DMZ. Liu, Zhengchun, et al. [7] explore the usage of machine learning as a tool rather than an application in order to advance towards a smart DTN able to self-optimize and adapt to the properties of a specific scientific workflow. The highly optimized hardware and software characteristics of a DTN also makes possible its usage within the context of network performance characterization and problem resolution in high-speed networks [8].

An alternative to Data Transfer Nodes exist when it comes to high-density data movement. Instead of a DTN with a local storage system, the data files may be located in a centralized infrastructure running a dedicated high-performance data transfer application. Such is the case of IBM Aspera, where users can obtain a license for the software and use it to upload and transfer their files. A DTN may also run one of these file transfer applications, such as the commercial Aspera FASP protocol or the open-source GridFTP, for better performance than the classic SCP or SFTP [9].

III. HARDWARE SELECTION

A DTN is comprised of various subsystems, and each of them needs to be optimized separately. Hardware selection and optimization in a DTN is focused on maximizing speed and latency of sequential read and write operations over a storage medium, as well as achieving optimal performance in data transfer through the network interfaces. With this in mind, there are three main subsystems to consider: storage subsystem, network subsystem, and motherboard and chassis.

A. Storage subsystem

Internal and external storage architectures are both possible. The simplest approach is using an internal storage medium, but other designs are possible, such as a connection to a Storage Area Network (SAN), or to a distributed file system. The latter options provide additional flexibility in that the DTN and the storage servers are separate systems, whereas a local storage is usually cheaper and easier to deploy.

The performace obtained from the storage subsystem depends on several factors. In a local storage scenario, the main elements to consider are the storage medium (disk), the RAM memory, the file system, and the RAID controller, if used. In an external storage scenario, the connection between the DTN and the storage is another factor to consider.

As far as storage medium selection is concerned, it comes down to a trade-of between performance, capacity and cost in the choice between using HDD drives (hard disk drive) or SSD drives (solid state drive). SSD drives have been traditionally more expensive than HDD drives, but offer much greater speeds and are therefore the best choices for a DTN. The best performance will come from the use of NVMe drives, which use PCIe (PCI Express) interfaces for memory access rather than traditional SATA or SAS interfaces [6]. The NVMe standard is specifically designed to leverage the low latency and parallelism characteristics of SSD drives. For example, a NVMe drive with a four-lane PCie Gen 3 interface, of which various models are available in the market, can achieve speeds of around 3GB/s for sequential reading and 2GB/s for sequential writing.

The RAID technology makes possible to virtualize multiple physical drives into one or more logical units for the purpose of improving performance and/or data redundancy. When it comes to choosing the proper configuration for a RAID controller, known as RAID level, there is a trade-of bewteen performance, capacity, and data redundancy to consider. In general, a RAID1 configuration is undesired in a DTN, as it offers maximum data redundancy at the cost of performance. The RAID0 configuration achieves the opposite goal, but without any kind of data replication, it is the most vulnerable configuration. There are, however, several other RAID levels with a more balanced compromise between data security and performance and capacity, such as RAID5, RAID6, and RAID10. Lastly, for the RAID controller itself not to

negatively impact performance when used alongside high-performance NVMe drives, at least 1GB of on-board cache and PCIe-3 support are required [10].

B. Network subsystem

The quality and speed of the network interface cards (NIC) used in a DTN is of great importance. Typical network interface speeds for a DTN are 10Gb/s, 40Gb/s, and even 100 Gb/s. The actual choice of network interface speed will depend on the specific use case and requirements of the DTN, but should also match the bandwidth limitations of the network where it is going to be deployed. In addition to speed, there are other features to consider about a NIC. Support or lack of support of these features will also affect in different ways a NIC's performance, and, by extension, the network subsystem's performance. The most important features to look for in a network interface card are described below.

- Interrupt coalescing support: interrupt coalescing is a technique that allows delaying the generation of hardware interrupts so that a single interrupt occurs for multiple events. Allowing the arrival of more packets before an interrupt is generated by the NIC results on a much lower CPU utilization.
- TCP Offload Engine (TOE): TOE is an optimization that allows offloading processing related to the TCP/IP stack from the system's CPU, by having dedicated NIC hardware take care of such tasks. These methods used to reduce CPU overhead can range from the offloading of the TCP/IP checksum calculation to the offloading of the lifecycle management of a TCP connection, or even of the entire TCP/IP stack. TCP Segmentation offload, described hereafter, is another example of a TCP offloading technique.
- TCP Segmentation Offload (TSO) and UDP Fragmentation Offload (UFO): the use of TSO makes possible the creation of TCP/IP packets bigger than the NIC's MTU, which are then broken into smaller packets by the NIC itself before being sent. This reduces the work that needs to be done by the CPU in order to construct the TCP/IP packets. UFO is the equivalent mechanism for the UDP protocol.
- MSI-X support: MSI is a method for signaling interrupts with special in-band messages rather than using out-of-band interrupt lines. With MSI, the device can signal an interrupt by writing in a specific memory address. MSI-X, defined in PCI 3.0, permits a device to allocate up to 2048 interrupts.
- Zero-copy protocol support: the use of remote direct memory access (RDMA) protocols enables direct access from the memory of a source host to the memory of a remote host, without the involvement of intermediate buffers in the operating system or of the CPU of either host.
- Receive Side Scaling (RSS): with RSS, the interrupts generated due to the arrival of packets are split across different CPUs. This helps parallelize the associated processing instead of having a single CPU do all the work.

C. Motherboard and chassis

The motherboard provides all the buses that interconnect the CPUs, memory, and controllers, so it is also a critical part that needs to be optimally designed in a DTN. The main aspects to consider include the CPU and CPU architecture, support for PCI Express and available PCIe slots, and RAM memory. A minimum of 32GB of RAM is recommended for a DTN [11]. Other important considerations are having adequate cooling and adequate power supply.

In regards to the CPU, the clock rate stands out as a key aspect of a DTN, often being even more relevant than the core count of the system [11]. The optimal architecture for a DTN is NUMA (non-uniform memory access), a multiprocessor architecture in which each processor has local access to an individual memory, and remote access to the memories allocated to the other processors. This way, different memories may be accessed by different processors simultaneously. Communication between processors involves latency, however, and the speed of the QPI (Intel QuickPath Interconnect) or HT (HyperTransport) buses determines it.

It is important to ensure that the PCIe slots available in the system can satisfy the requirements of the tasks carried out by the DTN. PCI slots, and, by extension, PCIe slots, are defined by their form factor, which refers to the number of lanes supported by the PCIe slot. The speed of each lane depends on the PCIe generation, with PCIe-2 offering 500 MB/s per lane and PCIe-3 doubling it to 1 GB/s. The NIC or NICs of the DTN, the drives (e.g. SSD NVMe), and the RAID controller will each have their own requirements when it comes to PCIe slots. For example, a 8-lane PCIe-2 slot would suffice for a 10G NIC, whereas a 100G NIC requires a 16-lane PCIe-3 slot [12].

IV. DTN TUNING

The default configuration of a Linux system is not optimal to obtain the best performance from the DTN. Tuning several elements of the DTN can improve its I/O throughput considerably, resulting in much better performance [13]. Tuning a DTN, however, varies with the hardware it is running. In general, among the elements that need proper tuning, there are network system, BIOS and CPU configuration, operating system, file system, memory, and I/O. Tuning of the network system is covered in section V.

A. BIOS and CPU configuration

Some suggested changes to be applied in the BIOS configuration are the following.

- Disable Hyper-Threading: Hyper-Threading is an Intel processor feature to split a CPU into two logical CPUs, so that the second uses a small percentage of the real CPU's processing capabilities. While this offers certain parallelism to help I/O-bound processes,

CPU-bound processes could be negatively impacted instead.

- Enable Turbo-Boost: Turbo-Boost is an Intel technology that allows an increase in the clock rate of the processor above its nominal frequency for as long as its design limits of energy, current and temperature have not been exceeded.
- Disable CPU frequency scaling mechanisms: CPU frequency scaling is used to dynamically adjust the maximum operating frequency of the CPU to save energy. In a DTN, it's generally desired to work at maximum frequency. Alternatively, the scaling governor mode of each CPU can be set to *performace* instead of to other modes like *powersave*.

B. NUMA CPU pinning

Proper core assignment to processes and to interrupts (IRQs) is important for performance reasons in NUMA architectures. For example, when it comes to network performance, one should make sure that the core or cores assigned to the NIC IRQs and to the related application processes belong to the same socket as the PCI slot of the NIC. This is also known as interrupt or process binding, and helps prevent situations where data would need to cross an interconnect bus between two NUMA nodes in order to access remote memory.

In order to manually assign cores to interrupts and processes, the *irqbalance* service must first be stopped. Then, IRQs can be bound to one or more specific cores by writing to their *smp_affinity* property, while core assignment to processes can be done through the *numactl* command.

C. I/O scheduler

The default I/O scheduler in some Linux versions is *fair*. It's recommended to use the *deadline* scheduler instead, which imposes a time limit on every I/O transactions so the bulkier ones don't slow down smaller ones. The I/O scheduler can be configured in the GRUB configuration file. This is not necessary for SSDs using the NVMe protocol, as they bypass the traditional I/O scheduler [14].

D. Virtual memory subsystem

Virtual memory tuning can improve write performance. This mainly involves tuning the values of the sysctl files *vm.dirty_background_bytes* and *vm.dirty_bytes*. The former determines the number of bytes in memory before the system begins to write data to disk, while the latter determines the maximum number of bytes in memory before blocking the write processes. Decreasing the value of the former and increasing the value of the latter can result in an improvement of write performance. Alternatively, *vm.dirty_background_ratio* and *vm.dirty_ratio* fulfill the same purposes, except they expect a value in percentages.

E. File system

The first parameter that can be configured to improve file system performance is the *readahead*. When a process reads a file sequentially, the kernel may prefetch some of its contents into RAM memory, so that it results in a lower latency than reading the contents from disk. The file system's readahead determines the maximum amount of 512-byte blocks that may be prefetched with this purpose. In a DTN, increasing the typical default value of 128 or 256 blocks by some orders of magnitude can result in much better sequential read performance. This can be achieved using the *blockdev* utility.

Other optimizations to consider in regards to the file system and focusing on a EXT4 file system in particular are the following. Firstly, disabling journaling can greatly improve write performance at the expense of reliability and risk of data corruption. If disabling journaling is not feasible, it can be configured in *writeback* mode so that only metadata is logged. If using RAID to form a multiple disk array, the *stride* and *stripe_width* shall be adjusted to match the design of the RAID array. The optimal stride size is calculated as the file system's block size multiplied by the RAID's chunk size, while the stripe width should be equal to the stride size multiplied by the equivalent number of drives in the RAID array providing capacity (depending on the RAID level, a different amount of stroage is used to provide redundancy). Finally, other configuration changes to consider when mounting an EXT4 file system in order to improve performance are increasing the value of the *commit* parameter, disabling write barriers (*barrier*), increasing *inode_readahead_blks*, and disabling *noatime* and *nodiratime*.

F. RAID controller

For a DTN, it is preferable to use an smaller chunk size as to optimize throughput involving operations over large files. This allows these large files to be divided across multiple drives, which can then be accessed in parallel. Other than that, different RAID controllers provide different configuration options, so it is generally beneficial for a DTN to tune them to favor performance with workflows involving large files.

G. SSD drive

TRIM support an usage is critical for a SSD drive in order to increase durability and performance. The TRIM command is used to tell the operating system which blocks of the SSD do not contain valid data and can be erased, thus considerably lowering the number of operations required on the SSD to write in one of these empty blocks. The *fstrim* command can be used to execute the TRIM operation on a mounted file system.

V. NETWORK TUNING

Network tuning is perhaps the most critical part of a DTN, as, ultimately, its job is to transmit data to another DTN over a network. Network tuning involves the optimal configuration of the DTN's NIC, and mainly of the behavior of the TCP and UDP protocolos.

A. Network interface card

The most important aspects to configure in a NIC are the size, in number of descriptors, of transmission (TX) and reception (RX) buffers. The TX and RX descriptors are typically arranged as a FIFO queue, and instead of containing the actual data to transmit, they point to other structures containing the data packets stored in the system [15]. Therefore, the number of descriptors determines the length of the queue; a higher number of descriptors is better at ensuring that the queue is not empty due to the speed at which IP packets are generated by the system, potentially improving throughput. However, a high number of descriptors could also negatively impact latency as a packet may need to wait for more packets already in the queue before it can be transmitted.

As a result, the optimal number of TX and RX descriptor comes down to a compromise between throughput and latency. As latency is inversely proportional to the NIC's speed, it's generally better to use larger queue lengths in a DTN. It should however be noted that, when using a mechanism such as TSO which allows packets much bigger than the MTU to arrive at the NIC, the latency per packet would be the equivalent to the latency of all the packets resulting from the single big packet after its split. Setting the number of RX and TX descriptors can be done by using the *ethtool* utility. Some NIC drivers may default to 256 descriptors, even if the hardware supports more, such as 4096 [16].

B. TCP

TCP is the primary transport protocol used to transmit large files between DTNs. Tuninig the behavior of TCP is critical to achieve flows of tens of Gbps with the DTN. It's important to take into account that, in the context of network performance, it's not only one's own DTN what comes into play, but also the bandwidth and latency characteristics of the network and intermediate nodes, as well as the remote DTN. Therefore, the bottleneck of the network system could be a different one in different situations or scenarios. The goal of TCP tuning in a DTN is so that the DTN itself is not the bottleneck in a high-speed TCP transmission, either as the transmitting or receiving node. TCP tuning focuses on optimizing the capacity of the buffers used in TCP connections, on selecting the optimal congestion control algorithm, and on fine tuning several additional parameters to help improve overall TCP performance.

The theoretical highest throughput achievable over a network is determined by a parameter known as the *Bandwidth Delay Product (BDP)*. BDP can be obtained as the product of available bandwidth and the *Round-Trip Time (RTT)* between source and destination, as shown in Ecuation 1. RTT depends itself on the distance of the link as well as the latency introduced by intermediate nodes.

$$BDP(bytes) = available\_bandwidth(kB/s) * RTT(ms) \quad (1)$$

BDP is an indicator of ideal throughput, and network tuning in a DTN aims to achieve a throughput that approaches the BDP as much as possible. Several tunable factors have an impact on the actually achieved throughput, but the TCP window size is by far the most relevant one. The window size during a TCP connection is dynamically updated following the flow control and congestion control mechanisms of TCP.

For a given connection, the TCP window size determines the number of packets (or, more specifically, bytes), that may be transmitted simultaneously without having received a confirmation (ACK) from the receiving side. The TCP window will thus be the limiting factor in terms of throughput if it's of smaller size that the link's or network's capacity. The maximum window size is directly limited by the size of the buffer used in the TCP connection, for both sender and receiver. In TCP, the source node must wait to receive the ACK confirming a packet before being able to discard it from the buffer, in case of retransmission being necessary; a full buffer will therefore cause the sender to be unable to send new packets until ACKs of some of the already transmitted packets are recevied. A similar effect can occur with the receiving side, depending on the speed at which it's able to process the received packets.

The minimum, maximum and default sizes of the transmit and receive buffers can be set via sysctl by writing to the appropriate files in the *proc/sys/net/ipv4* and *proc/sys/net/core* directories, for TCP connections and for any type of connection, respectively. Increasing the maximum values allows any application requiring it to reserve that much memory for a TCP connection. The necessary buffer sizes scale up with the available bandwidth and RTT between source and destination, so they should be tuned according to these variables.

Congestion control is also an important factor impacting the throughput and overall performance of TCP. Different congestion control algorithms react to different congestion events, such as duplicate ACKs or timeouts, differently. Typical algorithms respond to duplicate ACKs by halving the current congestion control window (CWND) value, interpreting that the event has been a consequence of congestion. This is not necessarily always the case, however, as buffering in intermediate nodes, for example, can cause a very small percentage of random packet loss over time. Similarly, traversing a device that may parallelize the processing of a TCP stream or split traffic between links can cause out of order arrivals on the receiver. At very high speeds, even a very small percentage of packet loss or out of order arrivals can lead to a very noticeable throughput degradation. The approach of the relatively new BBR algorithm is precisely to avoid misinterpreting every packet loss event as a signal of congestion and to avoid shrinking the CWND as a consequence [17]. BBR comes as an alternative to other existing algorithms such as CUBIC and HTCP.

Other aspects and parameters that are worth looking at in order to improve the DTN's TCP performance are

described as follows.

- Fair Queuing (FQ): a mismatch between the DTN's network interface speed and the capacity of the netowrk or of the receiver can be a source of poor performance, as it would lead to traffic being sent in bursts and could cause buffering issues in intermediate network devices. This mismatch between sender and receiver can be dealt with by using a packet pacing technique known as Fair Queuing [18]. The FQ scheduler can be enabled with the *tc* utility, which allows to set the maximum rate at which to send data. TSO should be disabled in order to use Fair Queuing.
- TCP Segmentation Offload (TSO): if using a NIC that supports TSO, it can be activated with the *ethtool* utility.
- Initial CWND and RWND values: the initial values of the congestion and receive windows for a given route can be modified in the routing table through the *initcwnd* and *initrwnd* parameters. Increasing them above their default values could help speed up the slow start period of a TCP connection.
- Sysctl TCP parameters: it is important to make sure that *tcp_window_scaling* and *tcp_sack* are both enabled, and fine tuning several other parameters could also be helpful.

C. UDP

UDP performance is generally limited by CPU. UDP throughput usually benefits the most from modifications focused on alleviating the processing requirements of the DTN when it comes to generating and transmitting UDP packets at high speeds. Two examples of these adjustments are increasing the size of the UDP packets sent over the network and assigning optimal NUMA cores to the NIC interrupts and UDP processes.

Using *jumbo frames* instead of sending packets with the usual MTU of 1500 bytes can result in a very substantial performance improvement. Jumbo frames have a MTU of 9000. As they can fit more bytes of data in the same frame, they require less CPU resources and processing than standard frames. This could also prove beneficial to TCP performance, as long as the TCP throughput is limited by CPU rather than any of the other factors previously discussed. However, in order to be able to send TCP or UDP packets with higher than usual MTU, the whole path between source and destination must support said MTU.

In the context of a NUMA achitecture, optimal core assignment to the NIC interrupts and to the application processes generating or processing the UDP traffic can also lead to a small but relevant performance improvement. Lastly, even though the concept of transmission window does not exist in UDP, it's important to ensure that there is enough buffer capacity for the UDP transmissions.

VI. TESTBED AND RESULTS

We ran various tests over our own testbed in order to analyze the implications of the different tunable and configurable parameters in the overall performance of a DTN and the supporiting network. The types of tests ran can be grouped into four main categories. These categories are disk read performance, disk write performance, TCP performance, and UDP performance. To run the tests, we used the following tools. *dd* was used to measure disk performance, *iperf3* to generate constant TCP or UDP traffic to study network performance between two DTNs, *htop* was used for process and CPU usage monitoring, and *netem* was useful for simulating delay and packet loss in the network. Some preliminary results were presented in the 2019 RedIRIS Technical Conference [19].

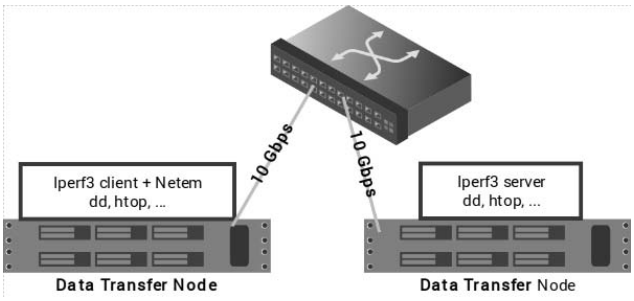


Fig. 2. Network topology of the experiment testbed.

In our testbed, represented in Figure 2, we used the following equipment. For the DTNs, we used two servers with an Intel® Xeon® Processor E5-2699 v4 processor with 2.20 GHz clock rate (3.60 GHz with Turbo-Boost) and 55 MB cache, and featuring a NUMA architecture of 2 NUMA nodes, 22 cores per node, and two 9.6 GT/s QPI buses. As for memory specifications, the DTNs had 64 GB of DDR4 memory working at 2133 MT/s. We used a Intel® Ethernet Controller X540-AT2 network interface card working at 10GbE and connected to a PCIe v2.1 x8 slot. Each DTN had a Intel® DC P3700 SSDPE2MD400G4 SSD NVMe drive (a single drive without RAID configuration was used), connected to a PCIe NVMe 3.0 x4, and featuring sequential read and sequential write speeds of up to 2700 MB/s and 1080 MB/s respectively, according to specifications. Finally, we used a DELL EMC S4048-ON 10 Gbps switch to interconnect the two DTNs through a 10 Gbps link of approximately 0.17 ms RTT.

For all the tests, we made sure to preset some configration parameters such as the following. Hyper-Threading c-states and irqbalance were disabled, while Turbo-Boost was enabled. The scaling governor was set to *performance*, and the I/O scheduler to *deadline*. The NIC IRQs were bound to cores in the local NUMA node. We used an EXT4 file system.

A. Disk read performance tests

For disk read performance testing, we first tried to identify the optimal range of block sizes in our system, and then compared different values of *readahead* against different values of file size and block size. Results are shown in Table I and Table II. Each test was run thrice and the results were averaged. Cache memory was cleared

Table I
SEQUENTIAL READ THROUGHPUT (GB/S) OVER A 32 GIB FILE FOR DIFFERENT BLOCK SIZES AND READAHEAD VALUES.

	Block size					
Readahead (blks)	4kB	16kB	64kB	256kB	16MB	1GB
256	0.67	0.71	1.37	1.36	1.35	1.33
4096	–	–	–	2.18	–	–
8192	–	–	–	2.62	–	–
16384	2.40	2.62	2.62	2.62	2.62	2.30
262144	2.41	2.62	2.63	2.63	2.62	2.40
524288	–	–	–	2.29	–	–

Table II
SEQUENTIAL READ THROUGHPUT (GB/S) OVER FILES OF DIFFERENT SIZE FOR DIFFERENT READAHEAD VALUES USING 256 KB BLOCK SIZE.

	File size (GiB)		
Readahead (blocks)	8	32	128
256	1.45	1.36	1.41
8192	2.80	2.62	2.53
16384	2.79	2.62	2.55
262144	2.79	2.63	2.41
524288	2.41	2.29	2.16

between each test to ensure the data comes from disk.

B. Disk write performance tests

For disk write tests, we settled to a block size of 256 kB and to writes of 32 GiB per test. The disk write tests focused on determining the benefits of tuning the virtual memory subsystem and the impact of using a file system with journaling. Write performance was tested in two operation modes. The *fdatasync* mode ensures that all data has been persisted into disk, while in the *default* mode the write process may end with some of the data still on memory. Results are shown in Table III and Table IV. Each test was run thrice and the results were averaged.

C. TCP performance tests

As far as TCP performance tests are concerned, we first tested the impact of using buffers that are not large enough for the available bandwidth and a given RTT. Then, we analyzed the performance of different congestion control algorithms against different packet loss rates. Finally, we compared the throughput obtained when lowering the MTU size, with and without TSO. All the tests were ran

Table III
SEQUENTIAL WRITE THROUGHPUT (GB/S) FOR DIFFERENT VALUES OF DIRTY_RATIO AND DIRTY_BACKGROUND_RATIO.

		Write mode	
dirty_background_ratio (%)	dirty_ratio (%)	default	fdatasync
10	20	1.41	1.02
20	20	1.42	0.93
10	40	1.72	1.02
20	40	1.71	0.92

Table IV
SEQUENTIAL WRITE THROUGHPUT (GB/S) IN FDATASYNC MODE USING A EXT4 FILE SYSTEM WITH DIFFERENT JOURNALING OPTIONS.

data=writeback mode	data=ordered mode	data=journal mode
1.02	1.02	0.36

Table V
TCP THROUGHPUT (GB/S) FOR DIFFERENT BUFFER SIZES AND RTTS

		RTT (ms)		
TX mem (bits)	RX mem (bits)	25	50	100
33554432	33554432	5.18	2.61	1.27
67108864	67108864	8.93	5.15	2.50
16777216	67108864	3.65	1.80	0.89
67108864	16777216	2.61	1.31	0.65

Table VI
TCP THROUGHPUT (GB/S) FOR DIFFERENT PACKET LOSS RATES USING DIFFERENT CONGESTION CONTROL ALGORITHMS OVER A 25MS RTT

	Packet loss (%)		
Algorithm	0.001	0.01	0.1
HTCP	8.18	1.67	0.21
Cubic	7.80	2.79	0.38
BBR	8.87	8.29	5.68

using *iperf3*, and *netem* was used to introduce packet loss and/or to introduce delays to simulate different RTTs. We experienced a misbehavior of the *netem* tool when used with TSO enabled, so TSO was disabled in all the tests that required the use of *netem*. Results are shown in Table V, Table VI, and Table VII. Each test was run thrice and the results were averaged.

D. UDP performance tests

For UDP tests, we focused on comparing the performance of a 1500-byte MTU versus a 9000-byte MTU corresponding to *jumbo frames*. We also did so by comparing the effects of binding the *iperf3* process to a core in the same NUMA node as the NIC versus a core in a different node. Results are shown in Table VIII. Each test was run thrice and the results were averaged.

VII. DISCUSSION

From the results in Table I we can conclude optimal block sizes range between approximately 16kB and 16MB. For these block sizes, the best throughput was achieved with a *readahead* of at least 8192 blocks, but not higher than 262144 blocks. We checked that CPU resources were never the bottleneck of the tests, as CPU usage went up to around 75% at most. Indeed, the obtained throughput values of around 2.62 GB/s come very close to matching the sequential read performance of the SSD specifications,

Table VII
TCP THROUGHPUT (GB/S) FOR DIFFERENT MTU SIZES

	MTU (bytes)			
TSO	1040	540	340	140
enabled	9.03	8.29	6.36	1.44
disabled	8.81	4.72	1.49	0.38

Table VIII
UDP THROUGHPUT (GB/S) FOR DIFFERENT MTU SIZES

	MTU (bytes)	
Process binding	1500	9000
unoptimal	2.31	7.10
optimal	2.52	7.92

which is 2.70 GB/s. From the results of Table II, we can notice a small decrease of throughput with the file size, perhaps related to the effectiveness of the readahead algorithm. In fact, when the file size is small enough, throughput actually surpasses the 2.70 GB/s mark, as readahead can help fetching a percentage of the data early.

Table III reflects the effects of tuning the *vm.dirty_background_ratio* and *vm.dirty_ratio* parameters. Decreasing the former and increasing the latter results in an increase of throughput, as this translates to less blocking of the write process and more priority writing to disk, respectively. The second column of the table suggests that a *vm.dirty_background_ratio* of 10% comes close to achieving the maximum sequential write performance of the SSD drive. On the other hand, decreasing the *vm.dirty_ratio* improves the throughput without *fdatasync*. As for the journaling results in Table IV, we can conclude that neither using a journal in *data=writeback* nor in *data=ordered* mode affected throughput negatively. On the other hand, *data=journal* mode led to a throughput degradation of almost 300%. This is because, in this mode, all data must be first written to the journal.

Table V suggests that the impact of not using large enough buffers in a TCP connection is very significant. If the buffer size is just optimal enough for a given RTT, doubling the RTT of the connection or halving the size of the buffer results in the throughput being approximately halved. This is a direct consequence of the congestion control window not being able to grow past a certain size, due to not having available memory to store more packet data. From the last two rows of Table V, we can also notice that, when the buffer size bottleneck is in the receiver side, throughput is worse than when the equivalent bottleneck is in the transmitter side instead.

The congestion control algorithm comparsion of Table VI reflects that BBR offers a throughput one order of magnitude above that of HTCP and Cubic in a 0.1% packet loss situation. Cubic performed slightly better than HTCP, but both collapse with a loss rate of 0.01% or higher, leading to throughputs of less than a third of the link's 10 Gbps bandwidth. Finally, Table VII shows how throughput degrades as MTU decreases and the TCP performance begins to be limited by CPU resources. Enabling TSO helps alleviate this degradation.

The UDP performance results in Table VIII reflect a throughput increase of over three times by using jumbo frames rather than typical 1500 MTU UDP frames. In addition, proper core assignment to the transmit and receive processes led to a throughput increase of around 10% regardless of the MTU size. All UDP tests ran were CPU-bound; the exact reason why CPU usage turned out to be a bigger limiting factor with UDP than with TCP wasn't clear, but it's assumed to be related to the software tool used for the tests (iperf3).

VIII. CONCLUSIONS

In this article, we introduced the DTN technology and the hardware and software characteristics and optimizations that a DTN should have in order to obtain the best performance. We introduced the results and conclusions of our experiments with DTNs in our testbed in regards to disk read and write performance and data transport using TCP and UDP protocols, which reflect the importance of optimally tuning several software components and system parameters.

Disk performance tests suggested the importance of choosing the right block size for maximum sequential write and read throughput. In addition, read tests benefited from *readahead* tuning, whereas write tests improved with adjustments to some system memory parameters. As for network performance, the results revealed how critical can be to have enough buffer size in both ends of a connection, for TCP in particular. The BBR congestion control algorithm performed much better than the other tested algorithms at overcoming random packet loss events. Finally, we showed the benefits of using larger MTU sizes, as well as optimal process binding in a NUMA architecture, to help performance of CPU-bound applications. This was particularly useful for improving UDP throughput.

Overall, the results obtained with the described hardware and software configuration suggest that, under the optimal tuning configurations, read and write accesses are faster than the achieved network speeds on a 10G link. On a higher speed link, such as 40G or 100G, however, assuming the maximum available bandwidth is achiveable through parallel connections or any other means, the bottleneck would begin to appear in disk performance if all the load is put in the same SSD disk.

ACKNOWLEDGEMENTS

This work was supported in part by the Spanish Ministry of Economy, Industry and Competitiveness through the State Secretariat for Research, Development and Innovation under the "Adaptive Management of 5G Services to Support Critical Events in Cities (5G-City)" project TEC2016-76795-C6-5-R.

REFERENCES

[1] ESnet, fasterdata.es.net/science-dmz/DTN/. Accessed 3 September 2019.
[2] PerfSONAR, www.perfsonar.net/. Accessed 3 September 2019.
[3] "Next Generation Data Transfer Nodes (DTNs) For Global Science: A." TNC18, tnc18.geant.org/core/presentation/166. Accessed 3 September 2019.
[4] Hong, Wontaek, et al. "Enhancing Data Transfer Performance Utilizing a DTN between Cloud Service Providers." Symmetry, vol. 10, no. 4, 2018, p. 110., doi:10.3390/sym10040110.
[5] Basu, Kashinath, et al. "Performance Comparison of a SDN Network between Cloud-Based and Locally Hosted SDN Controllers." 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), 2018
[6] Mughal, A, and H Newman. "Data Transfer Nodes and Demonstration of 100-400 Gbps Wide Area Throughput Using the Caltech SDN Testbed." Journal of Physics: Conference Series, vol. 898, 2017, p. 082052., doi:10.1088/1742-6596/898/8/082052.
[7] Liu, Zhengchun, et al. "Toward a Smart Data Transfer Node." Future Generation Computer Systems, vol. 89, 2018, pp. 10?18.
[8] "FAIL-Transfer: Removing the Mystery of Network Performance from Scientific Data Movement." PDF, docplayer.net/25726808-Fail-transfer-removing-the-mystery-of-network-performance-from-scientific-data-movement.html. Accessed 3 September 2019.
[9] ESnet, fasterdata.es.net/data-transfer-tools/. Accessed 3 September 2019.

[10] ESnet, fasterdata.es.net/science-dmz/DTN/hardware-selection/storage/. Accessed 3 September 2019.
[11] ESnet, fasterdata.es.net/science-dmz/DTN/hardware-selection/motherboard-and-chassis/. Accessed 3 September 2019.
[12] ESnet, fasterdata.es.net/science-dmz/DTN/100g-dtn/. Accessed 3 September 2019.
[13] ESnet, fasterdata.es.net/science-dmz/DTN/tuning/. Accessed 3 September 2019.
[14] "SSDOptimization." - Debian Wiki, wiki.debian.org/SSDOptimization. Accessed 3 September 2019.
[15] "Queueing in the Linux Network Stack ." Linux Journal, www.linuxjournal.com/content/queueing-linux-network-stack. Accessed 3 September 2019.
[16] ESnet, fasterdata.es.net/host-tuning/nic-tuning/. Accessed 3 September 2019.
[17] Cardwell, Neal, et al. "BBR: Congestion-Based Congestion Control - Google AI." Google AI, ai.google/research/pubs/pub45646.
[18] ESnet, fasterdata.es.net/network-tuning/packet-pacing/. Accessed 3 September 2019.
[19] Jornadas Técnicas de RedIRIS 2019 https://www.rediris.es/jt/jt2019/. Accessed 3 September 2019.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Simulación y análisis de rendimiento en dispositivos LoRa sobre drones

Jose-Manuel Martinez-Caro, Maria-Dolores Cano
Departamento de Tecnologías de la Información y las Comunicaciones
Universidad Politécnica de Cartagena (UPCT)
Plaza del Hospital 1, 30202. Cartagena (Murcia)
josem.martinezcaro@upct.es, mdolores.cano@upct.es.

***Resumen*- La tendencia actual es conectar a la red todo aquello que nos rodea y darle cierta inteligencia bajo el paraguas de Internet of Things (IoT). Las tecnologías Low-Power Wide-Area-Network (LPWAN) son un claro ejemplo de esta revolución tecnológica que ofrecen un gran rango de cobertura, precios bajos y un gran número de dispositivos conectados utilizando el mínimo consumo energético. La desventaja es el bajo Data-Rate (DR) que se da en estas comunicaciones, aunque puede adaptarse a los envíos de información no prioritarios. Long-Range y LoRa Wide-Area-Network (LoRa/LoRaWAN) son un tipo de LPWAN donde los dispositivos LoRa transmiten datos en bruto a los LoRa Gateways, que a su vez reenvían estos datos a servidores (normalmente en la nube) para ser procesados y generar información. Por otro lado, cada vez es más común la incorporación de vehículos aéreos no tripulados (Unmanned Aerial Vehicles, UAV) en las redes de comunicación dada su versatilidad. En este trabajo, se presentan los resultados de análisis de rendimiento de un servicio IoT desplegado sobre LoRa/LoRaWAN, en el que los nodos LoRa están embebidos en diferentes UAV. La evaluación se ha realizado mediante simulaciones extensivas, tras añadir nuevas funcionalidades al simulador FLoRa (Framework for LoRa), usando como métrica de evaluación diferentes componentes de calidad y bajo diferentes patrones de movilidad (Circular, Random Waypoint, Random Direction, Gauss-Markov y Tractor). Los resultados indican que, para el servicio testeado se alcanzan mayores niveles de calidad con Random Direction.**

***Palabras Clave*- IoT, LPWAN, LoRa, UAV, drones, QoS, QoD, QoI, QoE, QC**

I. INTRODUCCIÓN

La Internet-of-Things (IoT) [1] ha creado un ecosistema dinámico de gran escala [2] que está revolucionando y simplificando las comunicaciones entre dispositivos y/o sistemas dando lugar a redes más heterogéneas [3]. A partir de los datos obtenidos en los dispositivos IoT se obtiene información útil para la toma de decisiones. Con la tendencia ascendente de IoT, se ha incrementado el número de dispositivos conectados a las redes de comunicación y todo apunta a que continuará hasta llegar a los 75 mil millones de dispositivos conectados a la red en 2025 [4]. La tecnología IoT está en constante desarrollo y presente en múltiples aplicaciones y servicios como Smart-Home, Smart-City, Industria 4.0, Smart-Grid, etc. [4][5].

Las características de IoT permiten englobar bajo su nomenclatura a tecnologías como Low-Power Wireless-Area-Network (LPWAN), que destaca por su alta eficiencia energética, bajo coste económico y baja potencia en la transmisión de mensajes. Pese a estas características, los dispositivos logran cubrir una gran área de cobertura obteniendo un alto rendimiento. La desventaja de estos dispositivos es el bajo Data-Rate (DR) con el que suelen trabajar, debiendo cumplir la restricción del *duty-cycle* (establecido en un 1%), por lo que no es útil para el envío de tráfico demandante en términos temporales. En topologías con alta densidad de dispositivos, éstos pueden conectarse a uno o varios gateways, que dan acceso a la nube. Esta tecnología puede alcanzar distancias de hasta 30 km [7] entre el emisor y receptor en la banda ISM (Industrial, Scientific and Medical), lo que supone una considerable mejora respecto a Wireless Local Area Network (WLAN), Wireless Sensor Networks (WSN) o redes móviles (2G, 3G y 4G) [7], [8]. La frecuencia utilizada en Europa es 868 MHz y 433 MHz, mientras que en USA es 915 MHz y 433 MHz. Algunos ejemplos de tecnologías LPWAN son Long-Range y LoRa Wide-Area-Network (LoRa/LoRaWAN) [9], Weightless [10], NWave [11], Telensa [12], Random Phase Multiple Access (RPMA) [13], Sigfox [14], and Narrow Band-IoT (NB-IoT) [14]. LoRa/LoRaWAN es una de las tecnologías LPWAN más populares por su rendimiento. Del mismo modo, los drones o UAV (Unmanned Aerial Vehicle) se han convertido en tendencia, al igual que IoT, en múltiples sectores y comienzan a utilizarse en gran cantidad de aplicaciones [15]. Aunque existen artículos científicos con propuestas para optimizar el despliegue de drones mejorando el valor de QoE (Quality of user Experience)/QoS (Quality of Service) son escasos los trabajos que hayan utilizado LoRa con drones [16]–[18]. Por

todo ello, el objetivo de este trabajo es doble: por un lado, añadir valor a nuestro simulador de redes LoRa empleando herramientas "open-source" tales como OMNeT++ [19], INET framework [20], FLoRa framework [21] y Crypto++ [22]; y por otro lado, evaluar las prestaciones en términos de calidad (desde distintas perspectivas) de un servicio de monitorización de calidad del aire en el que los dispositivo IoT LoRa se encuentran embebidos en drones. Así, nuestra contribución ha sido:

1. Por defecto, FLoRa framework implementa una versión simplificada del algoritmo Okumura-Hata [23] basada en una regresión lineal para calcular las pérdidas de la señal en espacio libre. Debido a imprecisiones en los cálculos, se redefine este modelo en el simulador obteniendo unas mediciones más reales a partir de [23].
2. Introducción de mecanismos de seguridad en LoRa mediante la incorporación de cifrado/descifrado y firma digital en la comunicación usando los métodos Counter Mode (CRT) y Cipher-based Authentication Code (CMAC), respectivamente. Ambos métodos basados en Advanced Encryption Standard (AES) [24], [25].
3. Los datos que leen los dispositivos LoRa son obtenidos de un dataset real [26]. A partir de él, se simula el sistema donde los dispositivos IoT están embebidos en drones que cubren un área semiurbana de 40x40km a una altura entre 30 y 100 metros. En la zona se ubican 4 LoRa Gateways (LoRaGW) que recogen las tramas enviadas por los dispositivos y las reenvían a través de Internet a un Network Server.
4. De forma periódica, cada periodo de tiempo (T_{eval}) se llevan a cabo tareas de evaluación de prestaciones para cuatro componentes de calidad: Quality of Data (QoD), Quality of Information (QoI) Quality of user Experience (QoE) y Quality Cost (QC) [27]–[32]. Estos componentes se calculan a partir de diferentes métricas obtenidas durante todo el proceso, desde que se obtiene el dato hasta que se procesa en el Network Server y se compara el rendimiento de diferentes patrones de movimiento aéreo.

El resto de documento se ha organizado del siguiente modo. En la sección 2 describimos de forma general el estado-del-arte en protocolos de comunicación LoRa/LoRaWAN, drones y herramientas de simulación de redes LoRa/LoRaWAN. Las herramientas y librerías necesarias para añadir funcionalidades al simulador se detallan en la sección 3. La sección 4 destaca las novedades introducidas y, por último, los resultados se muestran en la sección 5. El documento finaliza con las conclusiones.

II. DESCRIPCIÓN DE TECNOLOGÍAS

A. LoRa/LoRaWAN

Un despliegue típico de redes LoRa/LoRaWAN se puede observar en la Fig. 1, donde nos encontramos principalmente tres tipos de dispositivos: (1) Módulos LoRa embebidos en drones, que simulan la lectura de datos mediante los sensores en la capa de aplicación y que se envían a través de la capa física a uno o varios LoRaGW; (2)LoRaGW que reciben las tramas LoRa, las convierten a *EthernetIIFrame* y las reenvían

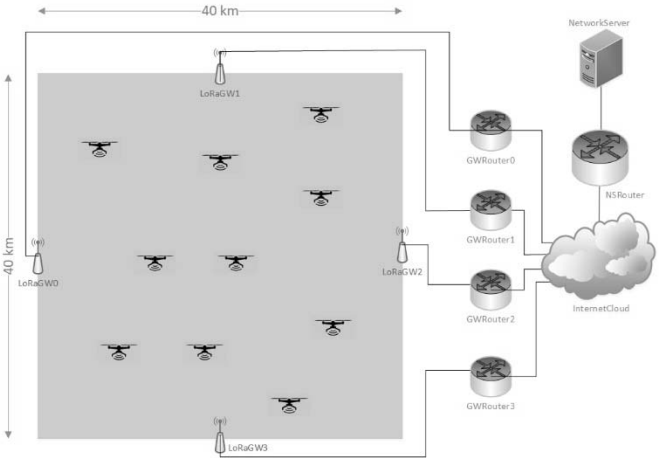


Fig. 1. Despliegue de red (10 dispositivos IoT LoRa en drones y 4 LoRaGW).

por la red de backhaul hasta alcanzar el destinatario del mensaje; (3) NetworkServer, el destinatario de los mensajes donde se procesan los datos recibidos (principalmente métricas) en cada paquete y genera información útil para la toma de decisiones, computando las componentes de calidad de forma periódica cada periodo de evaluación (T_{eval}).

La capa física LoRa utiliza modulación Chirp Spread Spectrum (CSS) en la banda ISM. Para ganar resistencia frente a interferencias y ruido, LoRa ensancha una señal de banda estrecha sobre un canal de comunicaciones con mayor ancho de banda [5], donde la sensibilidad del receptor puede llegar a ser de hasta 19.5dB por debajo de la señal de ruido. Hay varios parámetros que caracterizan la comunicación entre los dispositivos LoRa y LoRaGW: Spreading Factor (SF), Transmission Power (TP), Carrier Frequency (CF), Coding Rate (CR) y Bandwidth (BW) [33].

En primer lugar, SF puede tomar cualquier valor entre 7 y 12 (ambos incluidos) y TP varía teóricamente desde los -4 dBm hasta los 20 dBm, fijando la intensidad que los dispositivos LoRa utilizan para transmitir paquetes a LoRaGW; aunque en la práctica, el rango de TP va desde 2 dBm hasta 17 dBm. Para mayores valores de SF y TP se obtendrán mayores áreas de cobertura. CF es la frecuencia central que utilizan los dispositivos en función de la zona de uso, dividida en pasos de 61 Hz. CR proporciona seguridad frente a interferencias (4/5, 4/6, 4/7 y 4/8) donde los valores mayores proporcionan una mayor protección. Por último, BW es el ancho de frecuencia utilizado en la banda, y cuanto mayor sea, menor será la sensibilidad del LoRaGW [33]. Además, se introducen otros parámetros secundarios a partir de los anteriores. Time-on-Air (ToA), es el tiempo que un paquete está siendo transmitido por el canal, desde un dispositivo LoRa a LoRaGW. Data-Rate (DR) es la cantidad de datos entregados desde un dispositivo LoRa a un LoRaGW en un tiempo dado. Ambos dependen directamente de SF y BW, y son opuestos uno del otro (a mayor DR menor ToA y viceversa). Esta tecnología tiene tres grados de diversidad (tiempo, frecuencia y SF) [5]. Entre los dispositivos LoRa y LoRaGW, el sentido de la comunicación puede ser unidireccional o bidireccional y el direccionamiento puede ser unicast, multicast o broadcast. Algunos autores proponen la implementación de algoritmos como Adaptive-Data-Rate (ADR) [34], Distributed Coordination Functions (DCF) como Carrier Sense Multiple Access (CSMA) [35] y Channel

Activity Detection (CAD) [36], [37] con el objetivo de gestionar los parámetros de los enlaces, optimizar los procesos de la red, proporcionar mecanismos de control de acceso y detectar el preámbulo de los paquetes LoRa que circulan por el canal de la forma más eficiente posible.

Por otro lado, LoRaWAN [9] especifica la arquitectura, capas y protocolos que operan sobre LoRa utilizando dos tipos de topologías: estrella o malla. Según sea la gestión de las ventanas de recepción, se consideran tres clases de dispositivos LoRa: clase A, clase B y clase C. Clase A está siempre durmiendo (sleep mode) a no ser que tenga algo para transmitir. Cuando transmite, es entonces cuando el dispositivo LoRa programa abrir una ventana de recepción pasado un cierto tiempo. Los dispositivos de clase B añaden ventanas de recepción adicionales a las programadas en clase A. Por último, clase C mantiene la ventana de recepción abierta siempre por lo que no requiere de ningún sincronismo como ocurre en las dos primeras clases.

B. Drones

Los drones o UAV (Unmanned Aerial Vehicle) se han convertido en tendencia, al igual que IoT, en múltiples sectores y comienzan a utilizarse en gran cantidad de aplicaciones [15]. Estos drones pueden clasificarse en función de ciertas características como el tamaño, tipo de alas, capacidad de comunicación, tipo de vuelo, etc. Además de poder tener un funcionamiento aislado, los drones también son capaces de crear redes de comunicación en el aire, son las denominadas redes ad hoc voladoras o FANET (Flying Ad-hoc Network) [38]. Se pueden considerar una extensión de las MANET (Mobile Ad-hoc Network) [39] aunque con características únicas en términos de movilidad, topología, propagación de la señal y restricción energética.

En nuestro escenario de simulación, los UAV se conectarán con nodos terrestres (LoRaGWs). Respecto a los patrones de movilidad, existen varios modelos [40]: (1) movilidad aleatoria: son los modelos más sencillos y más comunes. Random Direction (RD) (Fig. 2.a) y Random WayPoint (RWP) (Fig. 2.b) forman parte de esta categoría. En RWP, se definen los parámetros de dirección, velocidad y pausa; donde al alcanzar el destino, se pausará un tiempo dado y buscará una nueva dirección. Para el caso de RD, desde un punto inicial se escogerá una dirección hasta chocar con el borde del escenario donde rebotará con un ángulo α; (2) movilidad temporal: depende de ecuaciones matemáticas que aplican cambios en velocidad y dirección. El modelo Gauss-Markov (Fig. 2.c) es un ejemplo más realista que los modelos RWP o RD; (3) movilidad enrutada: los nodos siguen una trayectoria precalculada sin tomar ninguna dirección aleatoria. En esta categoría se encuentra el modelo Circular (Fig. 3.a), en torno al centro del escenario de simulación y un radio *r* de distancia, y el modelo Tractor [41] mostrado en la Fig. 3.b, cuyo inicio y final es el mismo punto dentro del escenario de simulación; y (4) movilidad en grupo: diferentes nodos de la red se mueven de forma conjunta respecto una referencia. Estos modelos se han dejado para trabajos futuros.

Existen numerosos artículos que tratan de optimizar el despliegue de drones mejorando el valor de QoE/QoS. Uno de ellos [42] trata de maximizar QoE, minimizando la potencia total transmitida por los UAVs. Para ello, se consideran las limitaciones en el canal de comunicaciones entre los UAVs y las estaciones base localizadas en tierra. Las métricas

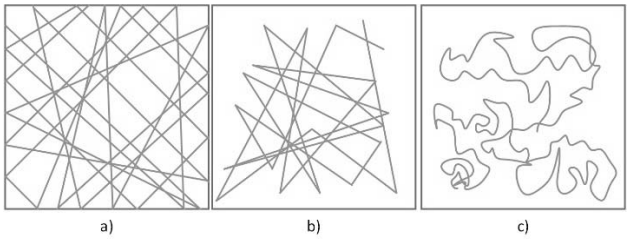


Fig. 2. Movilidad aleatoria y temporal: a) modelo Random Direction, b) modelo Random Waypoint, c) modelo Gauss-Markov.

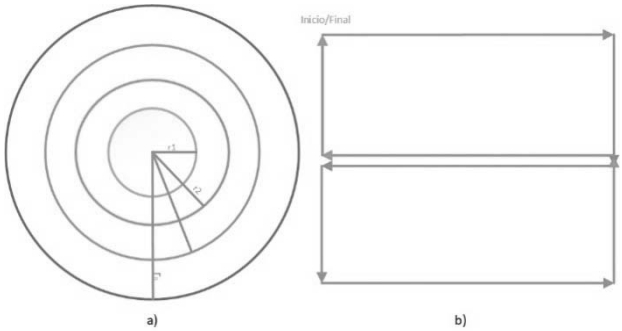


Fig. 3. Movilidad enrutada: a) modelo Circular, b) modelo Tractor.

utilizadas para computar el QoE son: el tipo de dispositivo, el retardo de la comunicación (delay) y el DR. Los autores tratan de encontrar los valores óptimos de DR y TP para que, al llegar un paquete al destinatario, la potencia de la señal recibida sea mayor que la sensibilidad del dispositivo, consiguiendo una correcta recepción y maximizar el valor de QoE. Ang Gao *et al.* [43] introducen un esquema de recursos energéticos eficientes para mejorar la QoE en una red de UAVs con diferentes necesidades y arquitectura heterogénea. El incremento de dispositivos en la red aumenta su capacidad y su complejidad [15]. Una de las mayores preocupaciones en este tipo de redes es la seguridad en las comunicaciones, ya que pueden ser un blanco fácil [44]. Sahil *et al.* [15] se suma a los aportes anteriores siguiendo una estrategia y manejo de los recursos eficiente debido a la limitación de los mismos manteniendo un equilibrio con la QoS. El fallo de un componente puede desencadenar el fallo de todo el sistema y ser una amenaza tanto para humanos como para infraestructuras [45]. Los trabajos [42], [46], [47] plantean un problema de optimización con el que minimizar el número de UAVs maximizando el área de cobertura satisfaciendo los requisitos de QoS. Bouachir *et al.* [48] ratifica a Zhu *et al.* [46] y añade que la movilidad de los dispositivos afectará a métricas de QoS como el *end-to-end delay* y el Packet Delivery Ratio (PDR).

III. SIMULADOR

Las herramientas utilizadas para el desarrollo del simulador son *open-source* y permiten adaptarse a cualquier escenario obteniendo una exhaustiva evaluación de prestaciones dada una topología. Algunos simuladores de redes conocidos son OMNeT++, NS3 [49], NetSim [50], SimPy [51], Cooja [52] u OPNET [53] entre otros. Además, existen un gran número de librerías *open-source* disponibles para ser importadas y añadir características extra a los simuladores.

Las librerías utilizadas en nuestro caso son: OMNeT++, INET framework, FLoRa framework y Crypto++. Con las

modificaciones introducidas al simulador FLoRa se pretende disponer de un entorno de simulación completo donde poder replicar el funcionamiento de una red LoRa y estimar el rendimiento de forma previa a pruebas experimentales. Permite además simular redes con cientos o miles de dispositivos que sería muy costoso económicamente y en tiempo [52].

A. OMNeT++

OMNeT++ [19] utiliza el entorno de desarrollo integrado (IDE, Integrated Development Environment) de Eclipse [54] como la principal plataforma de desarrollo y la mejora con nuevas funciones. Permite a los usuarios crear y/o reconfigurar modelos a partir de los ficheros NED e *ini*, ejecutar simulaciones y evaluar el rendimiento a partir de los resultados obtenidos en la simulación. OMNeT++ utiliza el lenguaje de programación C++ para definir componentes simples con los que definir el funcionamiento de diferentes módulos disponibles en la red e integración con GIT. Los ficheros NED establecen un modelo (simple o compuesto) y puede definirse mediante programación o emplazando y uniendo módulos compuestos, canales, características y otros tipos de componentes mediante interfaz gráfica. Por otro lado, los ficheros *ini* proporcionan los parámetros para adaptar y configurar el entorno de simulación deseado, reconociendo todos los componentes NED (desde el nivel superior hasta el último nivel heredado) e inicializando parámetros con diferentes valores a los establecidos por defecto. Los ficheros *ini* pueden editarse de la misma forma que los ficheros NED, pudiendo definir diferentes escenarios utilizando secciones que pueden heredar unas de otras [55].

Más de un proceso puede llevarse a cabo a la vez, así que mediante el uso de diferentes *cores* del equipo se puede acelerar la simulación. Mientras que la simulación corre, el programador puede continuar su desarrollo trabajando de forma paralela permitiendo una mayor eficiencia. Una vez finaliza la simulación, los resultados son almacenados en: vectores, donde se almacenan los resultados intermedios de la simulación; y escalares como valores tomados al final de la simulación, medias o desviaciones típicas de los vectores. Estos resultados pueden ser analizados con las herramientas ofrecidas por el IDE o exportadas para ser procesadas con otras herramientas externas como Python [56].

B. INET Framework

Este framework aporta nuevas capacidades a OMNeT++ proveyendo agentes, protocolos y modelos a nivel físico, de enlace, red, transporte y aplicación para usarlos y redefinirlos para diferentes tipos de redes de comunicación (cableada, inalámbrica, ad-hoc, WSN, etc.). Además, INET [20] basa sus operaciones en el intercambio de mensajes entre módulos permitiendo la certificación y validación de nuevos protocolos y escenarios creados.

C. FLoRa Framework

FLoRa [21] es una librería que simula y evalúa redes LoRa/LoRaWAN desde el nivel físico hasta el nivel de aplicación. Se basa en el emplazamiento de múltiples dispositivos LoRa en el escenario de simulación y generan paquetes que se destinan al dispositivo NetworkServer. Las redes LoRa/LoRaWAN deben albergar al menos un LoRaGW que será el encargado de recibir las tramas LoRa, encapsularlas en un *EthernetIIFrame* y reenviarlas por la red *backhaul* utilizando el protocolo UDP (User Datagram Protocol). Este protocolo sigue una comunicación simple sin conexión, sin confirmación ni control de flujo hasta el NetworkServer. La comunicación será bidireccional y se llevará a cabo una estimación del consumo energético de cada dispositivo LoRa que participe en el intercambio de mensajes en función del estado en el que se encuentre el *transceiver* (transmitiendo, recibiendo, durmiendo y apagado), el tiempo en cada estado y el valor de TP fijado. El consumo de cada estado será dado por el data-sheet de Semtech SX1272/73 [57] para una alimentación de 3.3V. FLoRa fija los principales parámetros de los dispositivos LoRa (SF, CF, BW, CR y TP). Estos parámetros influirán en el rango de cobertura y la probabilidad de colisión de paquetes. Como LoRa utiliza interfaz inalámbrica para la transmisión y recepción de mensajes, una trama será recibida correctamente si la señal llega con una potencia superior a la sensibilidad de LoRaGW (que depende de SF y TP).

Además de los dispositivos LoRa, detrás de los LoRaGW incluimos una red backhaul por donde encaminar los paquetes hasta llegar al NetworkServer. Esta red estará compuesta por 4 GWRouters que conectarán con cada LoRaGW de la red, un elemento que simula Internet, el NSRouter que encaminará el paquete hasta alcanzar el NetworkServer descartando los paquetes recibidos de forma duplicada. Para la simulación de esta parte se utilizan módulos y componentes desarrollados en INET framework [20].

D. Crypto++

Como las anteriores, Crypto++ [22] es una librería online basada en el lenguaje de programación C++ que incluye algoritmos de cifrado, códigos de autenticación de mensajes, generadores de HASH, sistemas de criptografía de clave pública, etc. Al importar la librería, nos permite utilizar múltiples métodos y esquemas como Diffie-Hellman, Advanced Encryption Standard (AES), RSA, Elliptic Curve Cryptography, Digital Signature Algorithm (DSA), etc.

IV. NOVEDADES

Aunque el montaje de FLoRa framework sobre INET framework y OMNeT++ ya era algo conocido, se han aplicado muchos cambios respecto a la versión original con el objetivo de tener una sencilla herramienta para la evaluación de prestaciones en servicios IoT y redes basadas en LoRa/LoRaWAN.

A. Nueva implementación del modelo Okumura-Hata

Originalmente, FLoRa framework incluye una implementación del modelo Okumura-Hata para estimar las pérdidas por espacio libre. Tras revisar la documentación y los ficheros de FLoRa, esta versión es una aproximación basada en una regresión lineal con tres factores (K1, K2 y la distancia entre el dispositivo LoRa y el LoRaGW). Los factores K1 y K2 toman como valores por defecto 127.5 y 35.2 respectivamente. Consideramos que este método no es preciso para calcular las pérdidas por espacio libre ya que el alcance máximo observado en simulaciones ronda los 6 km. Según la literatura, esta distancia es demasiado pequeña para el alcance máximo de la tecnología LoRa [7], [58]. Además, la versión existente no permitía diferenciar los entornos disponibles en

el método Okumura-Hata original (rural, sub-urbano y urbano).

A partir de las fórmulas Ec. (1)-(4)[23] utilizamos el modelo completo de Okumura-Hata en FLoRa evitando de este modo aproximaciones y obtiendo resultados mucho más precisos mediante la selección del entorno conveniente (rural, sub-urbano y urbano). Los factores a la hora de estimar las pérdidas son: la frecuencia (*f*), la altura de los dispositivos LoRa (h_m), la altura del LoRaGW (h_b) y la distancia (d_m) entre el dispositivo LoRa y el LoRaGW [23]. Las pérdidas serán menores en un entorno rural, ya que las señales electromagnéticas pueden propagarse mejor al darse una menor densidad de edificios; y mayor en un entorno urbano. Se pretende mediante el uso de este modelo de propagación representar el efecto de la capa física proporcionando un entorno lo más real posible descartando las aproximaciones que no son tan precisas

$$a(h_m) = 3.2(log_{10}(11.75 \cdot h_m))^2 - 4.97 \quad (1)$$

$$L_{urban} = 69.55 + 26.16log_{10}f - 13.82log_{10}h_b - a(h_m) + (44.9 - 6.55log_{10}h_b)log_{10}d_m \quad (2)$$

$$L_{sub-urban} = L_{urban} - 2\left(log_{10}\left(\frac{f}{28}\right)\right)^2 - 5.4 \quad (3)$$

$$L_{rural} = L_{urban} - 4.78\left(log_{10}(f)\right)^2 + 18.33log_{10}(f) - 40.94 \quad (4)$$

B. Security

Se ha modificado el simulador para enviar la información segura extremo a extremo en la comunicación y autenticar al emisor de los paquetes recibidos.

Para garantizar que la información viaja segura por el canal hasta su destino se emplea AES Counter Mode (CRT) para cifrar y descifrar la información en el emisor y receptor, respectivamente. El método CRT es un método de cifrado simétrico que usa una clave compartida entre el emisor y el receptor para mantener oculta la información sensible en su paso por la red. La información será cifrada en el dispositivo LoRa y descifrada por aquellos dispositivos que almacenen la clave compartida, en este caso el Network Server.

La autenticación del emisor del mensaje se lleva a cabo con AES Cipher-based Message Authentication Code (CMAC). El emisor utiliza también una clave para firmar el mensaje y así, el receptor verificar el origen y la integridad del mismo. Esta técnica permite detectar ataques de usuarios malintencionados que modifican el contenido de los mensajes como *Man-in-the-Middle* [24]. Este proceso se lleva a cabo gracias a la librería Crypto++ importada en OMNeT++. La longitud de las claves es de 128 bits.

C. Capa de aplicación

La configuración que incluye FLoRa framework en la capa de aplicación tanto para dispositivos LoRa como para NetworkServer es muy básica. Cada dispositivo LoRa genera un paquete de forma aleatoria siguiendo una distribución exponencial con media 100s, que contiene un número aleatorio dentro del payload del mensaje *LoRaAppMessage*. Este mensaje se envía a la capa física de LoRa donde se encapsula dentro de un mensaje *RadioFrame* para ser transmitido por el canal en texto plano. Cuando este paquete llega al nivel de aplicación del NetworkServer se contabiliza una nueva recepción satisfactoria y se descarta el contenido del mensaje. Sin embargo, este funcionamiento es muy básico para conseguir el propósito deseado.

En consecuencia, realizamos las siguientes modificaciones (Fig. 4). Por un lado, la capa de aplicación de los dispositivos LoRa (*SimpleLoRaApp*) es dividida en tres sub-módulos (*Read, CipherData y SimpleLoRaApp*) donde cada uno de ellos tiene una tarea específica. El sub-módulo *Read* permite a cada dispositivo LoRa leer de su propio dataset y una vez adquiere los datos, envía un *ReadDataPacket* al módulo *CipherData*. Este sub-módulo recibe los datos leídos e inicializa el proceso de cifrado simétrico, donde una vez finalizado enviará los datos cifrados y firmados al sub-módulo *SimpleLoRaApp* que reenviará los datos a la capa física. Entre la capa de aplicación y la capa física se han dispuesto dos medidores de *throughput* para conocer el tráfico generado y recibido por cada dispositivo LoRa en bits/s y packets/s.

Para calcular el número de paquetes perdidos y computar el PDR, en LoRa/LoRaWAN modificamos la interfaz de LoRaGW para que cuando reciba un nuevo paquete compruebe dos valores, el número de secuencia del mensaje y el ID del dispositivo LoRa. Si para un ID *i*, el número de secuencia recibido es mayor que el esperado, la diferencia será el número de paquetes perdidos. Los mensajes *LoRaAppMessage* y *LoRaMacFrame* modifican su payload para incluir las métricas calculadas en la red. LoRaGW encapsula los mensajes LoRaMacFrame dentro de un *EthernetIIFrame* y lo reenvía por la red hasta alcanzar NetworkServer usando el protocolo de transporte UDP.

El módulo NetworkServer también es dividido en tres sub-módulos *CommunicationParameters*, *Decrypt* y *Processing*. *CommunicationParameters* calcula las métricas que nos permitirán obtener las componentes de calidad a diferentes niveles de abstracción. Los mensajes recibidos y las métricas obtenidas de cada mensaje recibido son enviadas al módulo *Decrypt* que a partir del texto cifrado recibido obtiene el texto plano utilizando la clave simétrica compartida. En el caso de que la firma no sea correcta, se descarta el mensaje. Por último, el módulo *Processing* lleva a cabo dos funciones bien diferenciadas: para cada paquete recibido en un periodo de evaluación (T_{eval}) se registran las métricas asociadas a dicho paquete y cuando T_{eval} finaliza, el simulador procesa las métricas obtenidas y calcula el valor de las componentes de calidad en ese periodo.

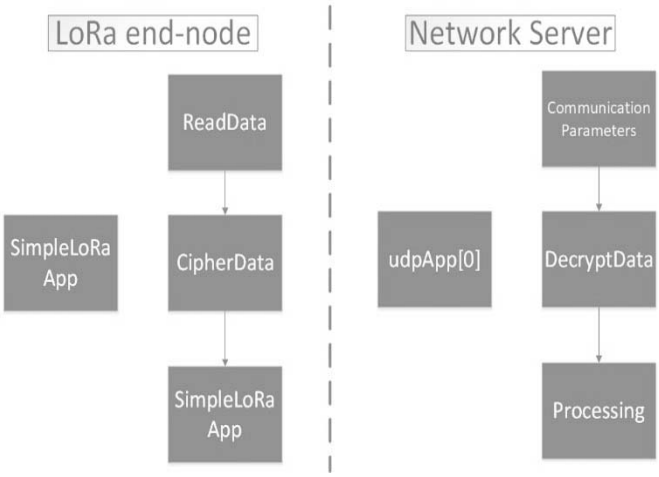


Fig. 4. Evolución de la capa de aplicación

D. Ejemplo de uso

El caso de estudio está basado en datos reales de una red de estaciones terrestres que monitorizan la calidad del aire en Euskadi [26]. Estas estaciones toman múltiples medidas reales (humedad, presión, PM2.5, PM10, CO, etc.). Estos datos se usarán como datos generados por los nodos LoRa. El escenario de simulación tiene unas dimensiones de 40x40km y en él, están desplegados 10 dispositivos LoRa moviéndose según diferentes patrones de movilidad: Random Waypoint, Random Direction, Tractor, Circle y Gauss-Markov. Los principales parámetros de los patrones de movilidad se recogen en Fig. 5. Los dispositivos están desplegados en un entorno sub-urbano con 4 LoRaGW ubicados tal y como aparecen en Fig. 1. La topología alberga más de un LoRaGW pudiéndose recibir tramas duplicadas que serán eliminadas en el nivel de aplicación del NetworkServer. Las componentes de calidad son obtenidas a partir de las métricas recibidas en el NetworkServer y son normalizadas para una mejor comparación, donde los valores más cercanos a 1 se interpretan como un mejor rendimiento. El método para obtener las componentes QoD, QoI, QoE y QC está descrito en [32] y está fuera del alcance de este artículo. Cada una de estas componentes mide la calidad a diferentes niveles teniendo en cuenta el raw-data, la información obtenida tras procesar, los parámetros de red y los costes asociados a las mejoras conseguidas.

V. Resultados

En esta sección se presentan y discuten los resultados obtenidos tras la simulación en un entorno sub-urbano para diferentes escenarios de movilidad. Las condiciones en la redde backhaul son idóneas ya que no se añaden retardos ni pérdidas. Las componentes de calidad son calculadas cada T_{eval} a partir de las métricas obtenidas de cada paquete recibido en el NetworkServer. La simulación para cada tipo de movilidad se ha repetido 10 veces y se muestran los valores medios de cada componente.

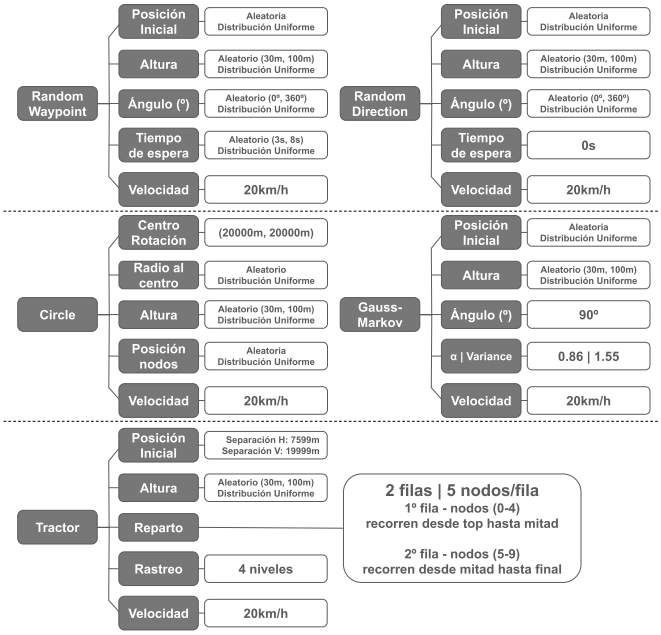


Fig. 5. Parámetros de los patrones de movilidad

Tabla I
TABLA DE EJEMPLO

Acrónimo	Métricas
QoD	Precisión, Veracidad e Integridad
QoI	Cantidad, Precisión, Recuperación, Detalle, Exactitud, Validez y Puntualidad
QoE	Retardo, Jiter, Ratio de paquetes perdidos, Capacidad y Disponibilidad de Gateway
QC	Consumo energético y Ciclo de trabajo (<1%)

Las Fig. 6-10 muestran las componentes de calidad, esto es QoD, QoI, QoE y QC, para cada patrón de movilidad analizado durante 100 horas de funcionamiento del dron. La componente QoD no depende de factores dados en la red de comunicación, sino que depende de lo completa o incompleta que sea la medición de los sensores. Se valora la calidad de los datos en bruto. Un sensor capaz de obtener datos de todas las mediciones que transmite muy frecuentemente, conseguirá incrementar el valor de QoD. En todos los patrones de movilidad estudiados, el valor de QoD se mantiene en torno a 0.8 en la escala normalizada ya que se ha utilizado el mismo dataset en todas las simulaciones. Respecto a la QoI, la información recibida en NetworkServer tras procesar los datos es levemente mejor bajo el patrón Random Direction (Fig. 7). El motivo es que se recibe un mayor número de mensajes (y de forma más precisa) de toda el área de simulación. En el patrón Tractor (Fig. 8) se presentan picos de mejoría en QoI, esto es porque los nodos se distribuyen y mueven por la red de forma simétrica, cubriendo todo el área de cobertura y maximizando QoI por su métrica Recall. Esta métrica penaliza cuando no se recibe información de alguna zona de la topología. La componente QoE engloba las métricas clásicas de QoS además de otros factores como nivel de uso de la red. Nótese que, en este caso las pérdidas y/o retardos sólo se dan en la parte inalámbrica de la comunicación (LoRa). Aunque el valor de QoE es muy similar para todos los patrones de movimiento, en la Fig. 9 vemos una leve mejoría. En el caso de la movilidad circular, los nodos rotan sobre un centro común (centro de la topología) minimizando los parámetros QoS en la comunicación LoRa respecto al LoRaGW más cercano. Por último, QC depende del consumo energético que a su vez depende del número de mensajes transmitidos. Dado que la generación de mensajes sigue una distribución exponencial de media 100s y no hay retransmisiones, se obtiene un valor muy similar para todos los patrones de movilidad estudiados. No sería así de existir bidireccionalidad o control de errores en el sistema, estudio que se deja para trabajos posteriores.

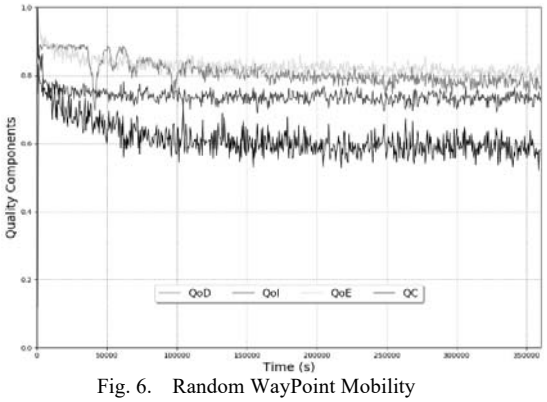


Fig. 6. Random WayPoint Mobility

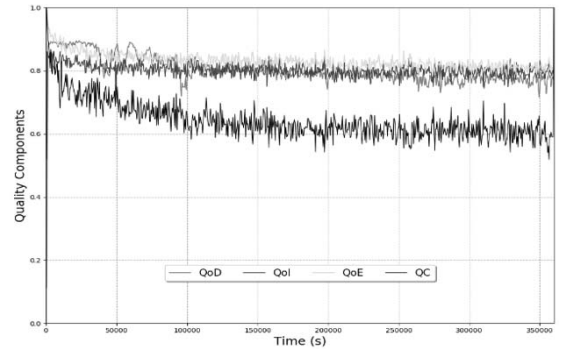


Fig. 7. Random Direction Mobility

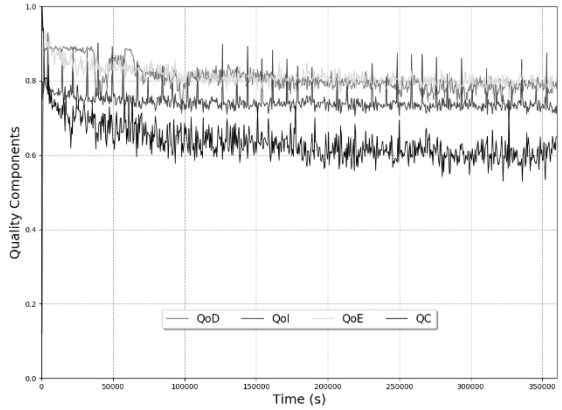


Fig. 8. Tractor Mobility

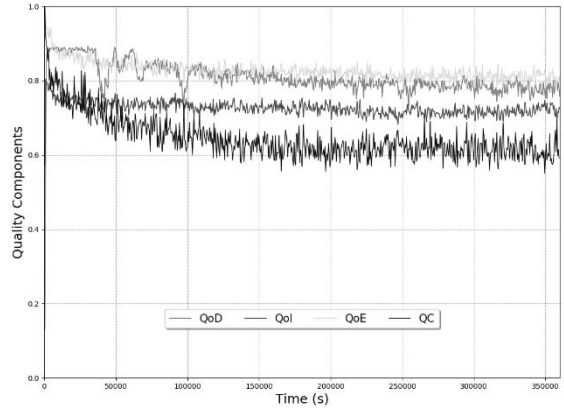


Fig. 9. Circle Mobility

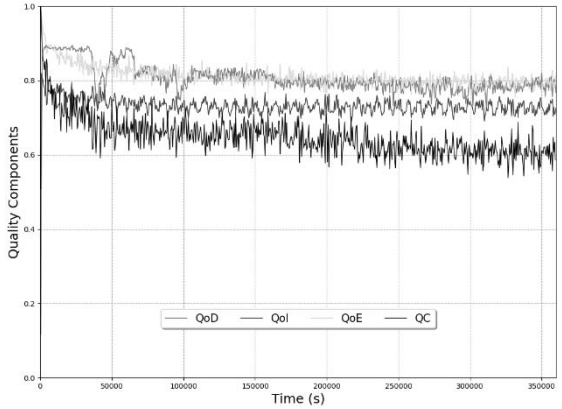


Fig. 10. Gauss-Markov Mobility

VI. CONCLUSIONES

Un correcto despliegue de aplicaciones y servicios sobre IoT requerirá de una correcta planificación previa, como en cualquier sistema de telecomunicación, por lo que es recomendable estudiar y validar el comportamiento de los servicios antes de ser implementados. Presentamos en este documento un simulador de redes LoRa donde analizar el rendimiento de esta tecnología utilizando librerías *open-source*. En este caso concreto, usamos OMNeT++, INET, FLoRa y Crypto++ para introducir nuevas características al entorno que deseamos recrear. Se ha implementado una nueva versión del modelo Okumura-Hata en el simulador para conseguir una mejor estimación de pérdidas por espacio libre y un área cobertura más acorde a la tecnología dada. El nivel de aplicación en dispositivos LoRa y NetworkServer se ha perfeccionado para llevar a cabo operaciones más complejas como cifrar y firmar la información. Además, se han analizado mediante simulación las prestaciones en términos de calidad de una red LoRa cuyos dispositivos están embebidos en drones. El estudio se ha hecho bajo diferentes patrones de movilidad (Random WayPoint Mobility, Random Direction Mobility, Tractor Mobility, Circle Mobility y Gauss-Markov Mobility) para ver cómo afectan éstos al rendimiento. Los resultados, todavía preliminares, indican que para el servicio testeado, el mejor modelo de movilidad es Random Direction Mobility. Como trabajo futuro, profundizaremos en el estudio de prestaciones de este tipo de sistemas.

AGRADECIMIENTOS

This research was supported by the AEI/FEDER, UE project grant TEC2016-76465-C2-1-R (AIM).

REFERENCIAS

[1] P. Suresh, J. V. Daniel, V. Parthasarathy, and R. H. Aswathy, "A state of the art review on the Internet of Things (IoT) history, technology and fields of deployment," *2014 Int. Conf. Sci. Eng. Manag. Res.*, pp. 1–8, 2014.

[2] G. Fortino, C. Savaglio, and M. Zhou, "Toward opportunistic services for the industrial Internet of Things," *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2017-Augus, pp. 825–830, 2018.

[3] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[4] "Internet Of Things (IoT) Connected Devices Installed Base Worldwide From 2015 To 2025," 2015. .

[5] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 2, pp. 855–873, 2017.

[6] R. Casadei, G. Fortino, D. Pianini, W. Russo, C. Savaglio, and M. Viroli, "Modelling and simulation of Opportunistic IoT Services with Aggregate Computing," *Futur. Gener. Comput. Syst.*, vol. 91, pp. 252–262, Feb. 2019.

[7] J. Petäjäjärvi, K. Mikhaylov, A. Roivainen, T. Hänninen, and M. Pettissalo, "On the coverage of LPWANs: Range evaluation and channel attenuation model for LoRa technology," *2015 14th Int. Conf. ITS Telecommun. ITST 2015*, pp. 55–59, 2016.

[8] S. Chieochan, E. Hossain, and J. Diamond, "Channel assignment schemes for infrastructure-based 802.11 WLANs: A survey," *IEEE Commun. Surv. Tutorials*, vol. 12, no. 1, pp. 124–136, 2010.

[9] LoRa Alliance, "LoRaWAN - What is it?. A technical overview of LoRa and LoRaWAN," no. November, pp. 1–20, 2015.

[10] "Weightless.," 2019. [Online]. Available: http://www.weightless.org/.

[11] "NWAVE Technology.," 2019. [Online]. Available: https://www.nwave.io/.

[12] "Telensa.," 2019. [Online]. Available:

https://www.telensa.com/technology.
[13] Ingenu, "RPMA Technology," 2019. [Online]. Available: https://www.ingenu.com/technology/rpma/.
[14] K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, "Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT," *2018 IEEE Int. Conf. Pervasive Comput. Commun. Work. PerCom Work. 2018*, pp. 197–202, 2018.
[15] S. Vashist and S. Jain, "Location-Aware Network of Drones for Consumer Applications: Supporting Efficient Management between Multiple Drones," *IEEE Consum. Electron. Mag.*, vol. 8, no. 3, pp. 68–73, 2019.
[16] Z. Yuan, J. Jin, L. Sun, K. W. Chin, and G. M. Muntean, "Ultra-Reliable IoT Communications with UAVs: A Swarm Use Case," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 90–96, 2018.
[17] A. Rahmadhani, Richard, R. Isswandhana, A. Giovani, and R. A. Syah, "LoRaWAN as secondary telemetry communication system for drone delivery," *Proc. - 2018 IEEE Int. Conf. Internet Things Intell. Syst. IOTAIS 2018*, pp. 116–122, 2019.
[18] L. Y. Chen, H. S. Huang, C. J. Wu, Y. T. Tsai, and Y. S. Chang, "A LoRa-Based Air Quality Monitor on Unmanned Aerial Vehicle for Smart City," *2018 Int. Conf. Syst. Sci. Eng. ICSSE 2018*, pp. 1–5, 2018.
[19] "OMNeT++ Simulator." [Online]. Available: https://omnetpp.org/. [Accessed: 05-Feb-2019].
[20] "INET Framework.," 2019. [Online]. Available: https://inet.omnetpp.org/. [Accessed: 05-Feb-2019].
[21] "FLoRa Framework," 2019. [Online]. Available: https://flora.aalto.fi/. [Accessed: 05-Feb-2019].
[22] "Crypto++ Library 8.0." [Online]. Available: https://www.cryptopp.com/. [Accessed: 05-Feb-2019].
[23] J. D. Parsons, *The Mobile Radio Propagation Channel*. 1992.
[24] J.-M. Martínez-Caro and M.-D. Cano, "Proporcionando seguridad en IoT con AES : un caso práctico de evaluación de consumo energético," in *DESEI+D 2016*, 2016, p. 173.
[25] NIST, "Announcing the ADVANCED ENCRYPTION STANDARD (AES)," *US Dep. Commer. Natl. Inst. Stand. Technol.*, vol. 56, pp. 57–71, 1993.
[26] "Euskadi air quality (2018)," 2018. [Online]. Available: http://opendata.euskadi.eus/catalogo/-/calidad-aire-en-euskadi-2018/. [Accessed: 09-Sep-2019].
[27] Q. Wu *et al.*, "Cognitive internet of things: A new paradigm beyond connection," *IEEE Internet Things J.*, vol. 1, no. 2, pp. 129–143, 2014.
[28] C. H. Liu, J. Fan, J. W. Branch, and K. K. Leung, "Toward QoI and energy-efficiency in internet-of-things sensory environments," *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 4, pp. 473–487, Dec. 2014.
[29] A. Floris and L. Atzori, "Quality of Experience in the Multimedia Internet of Things: Definition and practical use-cases," in *2015 IEEE International Conference on Communication Workshop, ICCW 2015*, 2015, pp. 1747–1752.
[30] P. Pawluk, M. Litoiu, and N. Cercone, "From QoD to QoS: Data quality issues in cloud computing," *CLOSER 2011 - Proc. 1st Int. Conf. Cloud Comput. Serv. Sci.*, no. January, pp. 697–702, 2011.
[31] P. Kasnesis, C. Z. Patrikakis, D. Kogias, L. Toumanidis, and I. S. Venieris, "Cognitive friendship and goal management for the social IoT," *Comput. Electr. Eng.*, vol. 58, pp. 412–428, Feb. 2017.
[32] J.-M. Martinez-Caro and M.-D. Cano, "A holistic approach to evaluate the performance of applications and services in the Internet of Things," *Submitt. to Int. J. Commun. Syst.*, no. Special issue on: Emerging ICT Applications and Service-Big Data, IoT, and Cloud Computing, 2019.
[33] M. Bor and U. Roedig, "LoRa transmission parameter selection," *Proc. - 2017 13th Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2017*, vol. 2018-Janua, pp. 27–34, 2018.
[34] M. Slabicki, G. Premsankar, and M. Di Francesco, "Adaptive configuration of lora networks for dense IoT deployments," *IEEE/IFIP Netw. Oper. Manag. Symp. Cogn. Manag. a Cyber World, NOMS 2018*, pp. 1–9, 2018.
[35] J. R. B. Junior, J. Lau, L. De Oliveira Rech, A. S. Morales, and R. Moraes, "Experimental Evaluation of the Coexistence of IEEE 802.11 EDCA and DCF Mechanisms," *Proc. - IEEE Symp. Comput. Commun.*, vol. 2018-June, pp. 847–852, 2018.
[36] C. Pham, "Investigating and experimenting CSMA channel access mechanisms for LoRa IoT networks," in *IEEE Wireless Communications and Networking Conference, WCNC*, 2018, vol. 2018-April, pp. 1–6.
[37] P. Yuan, X. Wen, H. Lu, and Q. Pan, "Dynamic Backoff Based Access Mechanism for LoRaWAN Class A," in *IEEE International Conference on Energy Internet Dynamic*, 2018, pp. 219–223.
[38] M. A. Khan, A. Safi, I. M. Qureshi, and I. U. Khan, "Flying ad-hoc networks (FANETs): A review of communication architectures, and routing protocols," *2017 1st Int. Conf. Latest Trends Electr. Eng. Comput. Technol. INTELLECT 2017*, vol. 2018-Janua, pp. 1–9, 2018.
[39] C. I. Katsigiannis, D. A. Kateros, E. A. Koutsoloukas, N. D. Tselikas, and I. S. Venieris, "Architecture for reliable service discovery and delivery in manets based on power management employing SIP extensions," *IEEE Wirel. Commun.*, vol. 13, no. 5, pp. 90–95, 2006.
[40] A. Guillen-Perez and M. D. Cano, "Flying ad hoc networks: A new domain for network communications," *Sensors (Switzerland)*, vol. 18, no. 10, 2018.
[41] "Tractor Mobility." [Online]. Available: https://doc.omnetpp.org/inet/api-current/neddoc/inet.mobility.single.TractorMobility.html. [Accessed: 03-Jun-2019].
[42] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the Sky: Proactive Deployment of Cache-Enabled Unmanned Aerial Vehicles for Optimized Quality-of-Experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, 2017.
[43] A. Gao, Y. Hu, W. Liang, Y. Lin, L. Li, and X. Li, "A QoE-Oriented Scheduling Scheme for Energy-Efficient Computation Offloading in UAV Cloud System," *IEEE Access*, vol. 7, no. c, pp. 68656–68668, 2019.
[44] C. Lin, D. He, N. Kumar, K. R. Choo, A. Vinel, and X. Huang, "Security and Privacy for the Internet of Drones : Challenges and Solutions," *IEEE Commun. Mag.*, vol. 56, no. January, pp. 64–69, 2018.
[45] L. Gupta, R. Jain, and G. Vaszkun, "Survey of Important Issues in UAV Communication Networks," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1123–1152, 2016.
[46] Z. Zhu, L. Li, and W. Zhou, "QoS-aware 3D Deployment of UAV Base Stations," *2018 10th Int. Conf. Wirel. Commun. Signal Process.*, pp. 1–6, 2018.
[47] M. Alzenad, A. El-keyi, F. Lagum, and H. Yanikomeroglu, "3D Placement of an Unmanned Aerial Vehicle Base Station (UAV-BS) for Energy-Efficient Maximal Coverage," no. April, 2017.
[48] O. Bouachir, F. Garcia, N. Larrieu, and T. Gayraud, "Ad hoc network QoS architecture for cooperative Unmanned Aerial Vehicles (UAVs)," in *IFIP Wireless Days*, 2013, pp. 1–4.
[49] "Discrete Event Network Simulator - NS3." [Online]. Available: https://www.nsnam.org/. [Accessed: 08-May-2019].
[50] "NetSim - Network Simulator & Emulator." [Online]. Available: https://www.tetcos.com/download.html. [Accessed: 08-May-2019].
[51] "SimPy 3.0.11." [Online]. Available: https://simpy.readthedocs.io/en/latest/. [Accessed: 08-May-2019].
[52] Y. Song, O. Zendra, and O. Zendra, "Using Cooja for WSN Simulations : Some New Uses and Limits To cite this version : Using Cooja for WSN Simulations : Some New Uses and Limits," pp. 319–324, 2016.
[53] "Opnet.com." [Online]. Available: http://www.opnet.com/. [Accessed: 08-May-2019].
[54] "Eclipse." [Online]. Available: https://www.eclipse.org/. [Accessed: 08-May-2019].
[55] A. Varga, "OMNet++ user guide Version 4.2.2," *OpenSim Ltd*, pp. 1–402, 2011.
[56] "Python." [Online]. Available: https://www.python.org/. [Accessed: 29-May-2019].
[57] "SEMTECH SX1272/73 Datasheet," no. January. 2019.
[58] R. Sanchez-Iborra, J. Sanchez-Gomez, J. Ballesta-Viñas, M. D. Cano, and A. F. Skarmeta, "Performance evaluation of lora considering scenario conditions," *Sensors (Switzerland)*, vol. 18, no. 3, 2018.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Detección de ciberataques mediante análisis estadístico con distribuciones α-estables

Luis de Pedro, Eric Crusi, Alberto Ruiz Santos, Jorge E. López de Vergara
Departamento de Tecnología Electrónica y de las Comunicaciones,
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Francisco Tomás y Valiente, 11, 28049 Madrid
luis.depedro@uam.es, eric.crusi@estudiante.uam.es,
alberto.ruizsantos@estudiante.uam.es, jorge.lopez_vergara@uam.es

***Resumen*—El análisis de los estadísticos del tráfico en una red permite caracterizarlo, hasta el punto de poder identificar los ataques utilizando ajustes estadísticos de series temporales con distribuciones α-estables. Otras técnicas se basan en la detección de mesetas, pero si el ataque no las genera, no son eficaces. En el presente artículo se presenta un estudio en el que se tratan diferentes alternativas de análisis estadístico, y se muestra que se puede clasificar el tráfico de ataque en el espacio de fases definido por los parámetros estadísticos de una distribución α-estable.**

***Palabras Clave*—Seguridad en comunicaciones, redes y sistemas. Análisis de datos de redes. Ataque informático, modelado de tráfico, denegación de servicio.**

I. INTRODUCCIÓN

La detección de ciberataques es un tema que está suscitando un extraordinario interés en los últimos años [1]. Diferentes técnicas han sido propuestas para realizar la detección y, óptimamente, la anulación de dichos ataques [1], [2]. Sin embargo, muchos de estos procedimientos adolecen de su fácil evitación, una vez que son de conocimiento público [3]. Por otro lado, el análisis estadístico del tráfico de Internet tiene la ventaja de su muy difícil alteración, al estar basado en propiedades globales del tráfico agregado y no en un aspecto concreto del mismo.

Tradicionalmente, se han utilizado modelos basados en distribuciones normales (gaussianas) para el modelado de diferentes aspectos el tráfico en Internet, tales como el ancho de banda consumido [4]. Para la detección de intrusiones se han utilizados modelos multivariantes que conjugan múltiples características del tráfico de la red [5]. No obstante, estos modelos no son capaces de ajustar distribuciones con cola pesada, por lo que se han propuesto otros modelos más generales. Por ejemplo, en [6] se propone el uso de distribuciones α-estables para la detección automática de anomalías en el tráfico de red, definidas estas como tráfico de probabilidad baja. A partir de dicho trabajo, en este artículo se explora la segmentación del espacio de fases de parámetros de distribuciones α-estables para la detección de ciberataques, incluso cuando los mismos no generan mesetas en las series temporales del tráfico monitorizado.

A continuación, se presenta una introducción a las distribuciones α-estables y se describe la metodología utilizada. Se justifica la utilización de los parámetros estadísticos y se construye el espacio de fases correspondiente. Como conclusión, se presenta un caso en el que es posible identificar un ciberataque de difícil detección, mediante clusterización del espacio de fases de los parámetros estadísticos de la distribución de la tasa de paquetes.

II. DISTRIBUCIONES α-ESTABLES

En los últimos años, se está desarrollando una vía de investigación para el modelado de tráfico basada en la utilización de las distribuciones denominadas α-estables, que se podrían considerar una generalización de las distribuciones gaussianas [7]. En efecto, la propiedad fundamental de las distribuciones α-estables es que lo siguen siendo después de realizar operaciones lineales con ellas. Formalmente, si X_1 y X_2 son dos copias idénticamente distribuidas de una variable aleatoria X y, dados dos números A y B, existen otros dos, C y D, tales que

$$AX_1 + BX_2 = CX + D \tag{1}$$

entonces, X es una variable α-estable. Por tanto, la suma de variables aleatorias α-estables da como resultado otra variable aleatoria α-estable.

A diferencia de otras distribuciones, no existe una expresión cerrada para este tipo de distribuciones. Existen varias parametrizaciones indirectas, como por ejemplo la función característica, indicada en la Ec. 2 y conocida como $S(\alpha, \beta, \gamma, \delta, 0)$ [7].

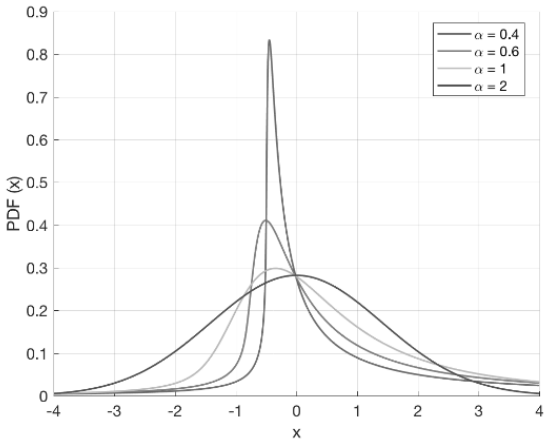


Fig. 1. Funciones de densidad de probabilidad α-estables para diferentes valores de α.

$$E\{\exp(itX)\} = \begin{cases} \exp\{-\gamma^\alpha |t|^\alpha \big[1+ \\ \quad +i\beta \text{tg}\big(\frac{\pi\alpha}{2}\big)\text{sgn}(t)(|\gamma t|^{1-\alpha}-1)\big] + i\delta t\} \text{ si } \alpha \neq 1 \\ \exp\{-\gamma|t|\big[1+i\frac{2\beta}{\pi}\text{sgn}(t)\log(\gamma|t|)\big] + i\delta t\} \text{ si } \alpha = 1 \end{cases} \quad (2)$$

donde α define la estabilidad o sesgo de la distribución, β la simetría, γ la escala y δ la localización de la misma. Para $\alpha = 2$ se tiene la distribución gaussiana como un caso particular, como se puede apreciar en la Fig. 1.

III. MEDIDAS DE TRÁFICO

Se ha elegido para modelar el tráfico el número de paquetes por segundo medidos en un enlace mediante una sonda que realiza una clasificación de los mismos y su agregación en flujos. Como fuente de datos se ha tomado el conjunto UGR'16 [8], que proporciona tanto tráfico normal como tráfico de ataque, seleccionándose ataques de tipo DoS para este estudio.

Un programa "ad hoc" desarrollado en AWK permite seleccionar la información relevante. Los datos procesados en formato CSV incluyen el inicio y duración de cada flujo y el número de paquetes que lo componen. Esta aproximación supone un compromiso entre la exhaustividad de la medida y el volumen de la captura [9]. En efecto, la captura individual de cada paquete supondría una mayor precisión, pero dado el volumen de datos, el almacenamiento y tratamiento de los mismos sería prohibitivo.

Puesto que la información para ajustar el modelo de tráfico está dimensionada en paquetes por segundo, es necesario hacer una estimación (prorrateo) de cómo se distribuyen a lo largo del tiempo los paquetes que constituyen cada flujo. Utilizando las medidas de tiempo inicial y duración de cada flujo se hace una estimación del número de paquetes por segundo en cada segundo de forma proporcional a la duración del flujo [9], [10], tal como se muestra en la Fig. 2. De esta manera, en un caso real se podría monitorizar los paquetes por segundo que

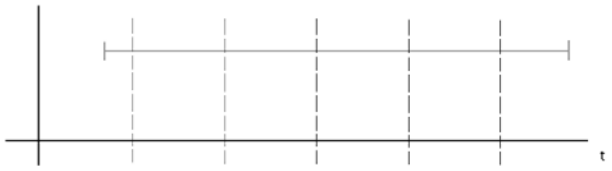


Fig. 2. Estimación del número de paquetes por segundo de un flujo.

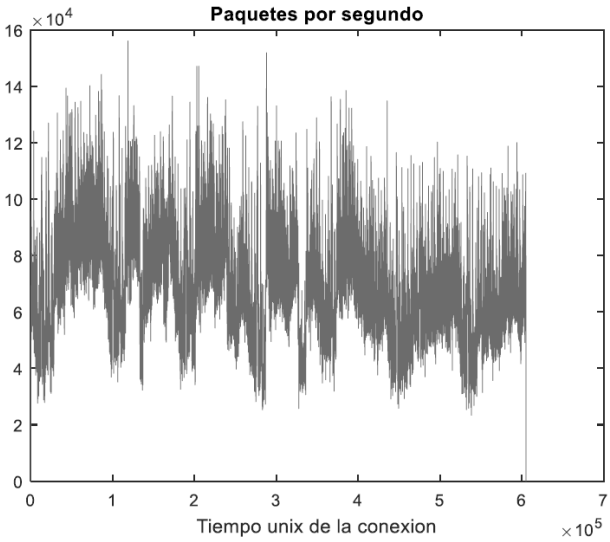


Fig. 3. Serie temporal resultado de la estimación.

atraviesan un enlace, que es menos costoso que obtener los flujos correspondientes.

El reparto del número total de paquetes se realizaría en el caso indicado según la duración del flujo dentro de cada segundo de duración. El resultado se suma al resto de flujos del enlace y se obtiene una serie temporal. Un ejemplo de la serie resultado del proceso de tráfico real se puede ver en la Fig. 3.

IV. ESTIMACIÓN DE PARÁMETROS ESTADÍSTICOS

A. Metodología

Para la estimación de los parámetros, se ha utilizado una ventana deslizante sobre la serie temporal de paquetes por segundo. El ajuste de la distribución α-estable se realiza para los datos incluidos dentro de la ventana, que se va desplazando con una granularidad de un segundo cada vez. Para cada posición de la ventana, se realiza un ajuste de distribución α-estable utilizando Matlab y se recogen los parámetros correspondientes. En la Fig. 4 se puede ver el ajuste realizado sobre tráfico real utilizando una ventana de quince minutos de duración.

Se pueden considerar así series temporales de los cuatro parámetros de ajuste. En paralelo, se calcula la media, mediana y moda de la ventana para poder comparar los resultados. El proceso completo se puede ver en la Fig. 5.

B. Utilidad de parámetros α-estables

Utilizando la metodología descrita en el apartado anterior, se han realizado pruebas con tráfico real y con tráfico de un ataque DoS de 2 minutos, mezclando ambos.

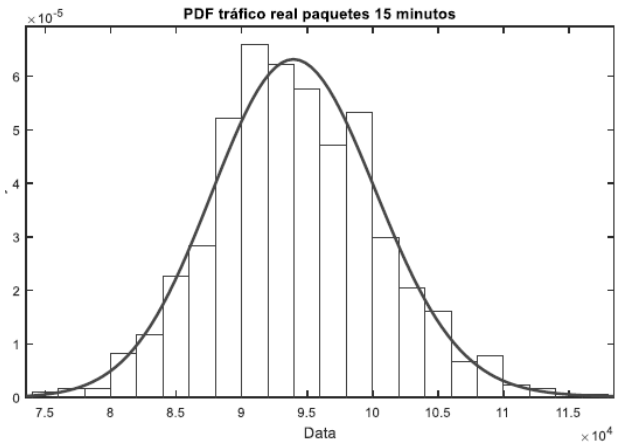


Fig. 4. Ajuste de la distribución del tráfico.

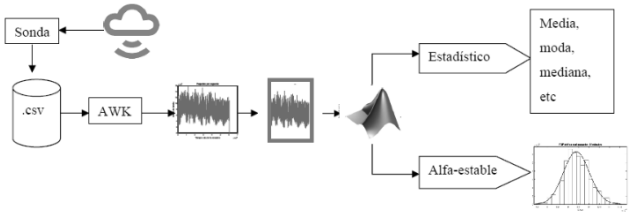


Fig. 5. Proceso de obtención de parámetros estadísticos.

Se ha probado con diferentes longitudes de ventana (dos y quince minutos) y se ha realizado el cálculo de los parámetros de α-estables y de los parámetros estadísticos media, varianza, etc. Las tablas I y II recogen dichos resultados, respectivamente.

Se puede apreciar que las variaciones de la media, mediana, etc. reflejadas en la tabla II no son tan significativas como la variación de algunos de los parámetros de la tabla I. Por ejemplo, el parámetro β presenta una alteración de más del 70% de su valor cuando se añade tráfico malicioso. Por tanto, se puede concluir que, si el objetivo es identificar un ataque de este tipo, los parámetros de la distribución α-estable resultan ser útiles

Tabla I
MEDIDAS DE PARÁMETROS α-ESTABLES.

Parámetro	Ataque 2 min.	Real 2 min.	Mezclado 2 min.	Real 15 min.	Mezclado 15 min.
α	1,36	1,99	1,99	1,91	1,74
β	0,36	1	1	1	0,24
γ	1,84E+02	4,35E+03	4,30E+03	4,39E+03	4,09E+03
δ	2,75E+03	9,24E+04	9,51E+04	9,41E+04	9,46E+04

Tabla II
MEDIDAS DE PARÁMETROS ESTADÍSTICOS (PAQUETES/S).

Estadístico	Ataque 2 min.	Real 2 min.	Mezclado 2 min.	Real 15 min.	Mezclado 15 min.
Media	2,73e+3	9,35e+4	9,63e+4	9,41e+4	9,46e+4
Mediana	2,66e+3	9,24e+4	9,51e+4	9,38e+4	9,43e+4
Moda	1,89e+3	8,37e+4	8,63e+4	9,09e+4	9,09e+4
Varianza	1,36e+5	3,64e+7	3,71e+7	4,11e+7	4,14e+7
Desv. típica	3,69e+2	6,03e+3	6,09e+3	6,41e+3	6,43e+3

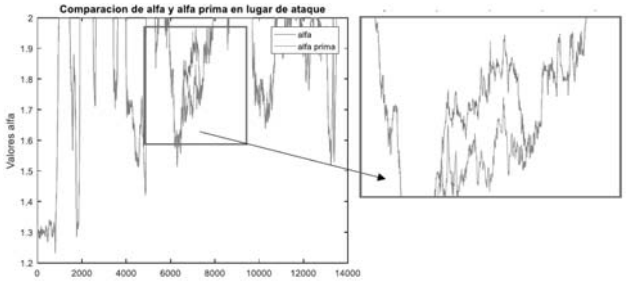


Fig. 6. Variación del parámetro α durante el ataque.

que los estadísticos tipo media o varianza. La Fig. 6 presenta gráficamente la variación del parámetro α en el momento del ataque. Aunque la variación es localmente significativa, se encuentran casos con valores superiores donde no hay ataque, por lo que es necesario combinar varios parámetros para poderlo identificar correctamente, según se muestra en la sección siguiente.

V. ESPACIO DE FASES

Con el resultado indicado en el apartado anterior, y partiendo de [6], se plantea la posibilidad de detección del ataque basándose en la alteración que se produce en los parámetros estimados durante un ataque. Los intentos de identificación usando uno solo de los parámetros de la tabla I no han llevado a una conclusión clara debido a la localidad de los resultados, motivo por el cual se ha procedido a utilizar combinaciones de los cuatro parámetros. En la Fig. 7 se puede ver un ejemplo de espacio de fases utilizando tres parámetros: α, β y δ.

Como se puede apreciar, hay cierta separación espacial entre el tráfico mezclado con ataque (aspas en rojo) y el tráfico sin él (circunferencias en azul). Tras realizar diferentes pruebas, se ha encontrado que un espacio de fases combinación de α y δ permitiría la identificación del tráfico mezclado con el ataque. La proyección de la Fig. 7 sobre el plano α, δ se puede ver en la Fig. 8.

VI. CLUSTERIZACIÓN DEL TRÁFICO

Utilizando el espacio de fases descrito en la Fig. 8 es posible hacer una clusterización del tráfico y clasificarlo

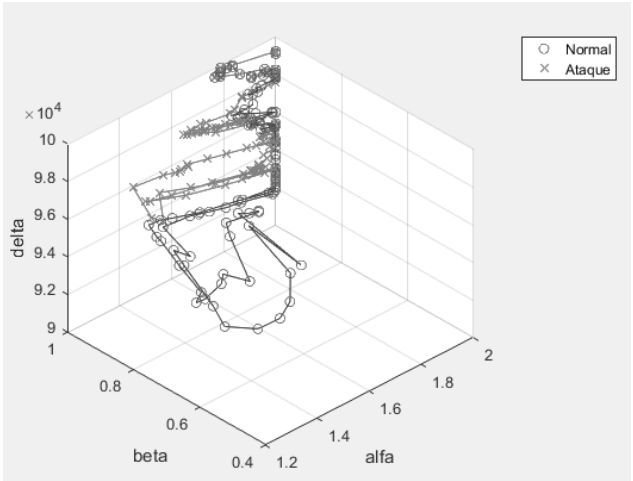


Fig. 7. Espacio de fases de parámetros α, β y δ.

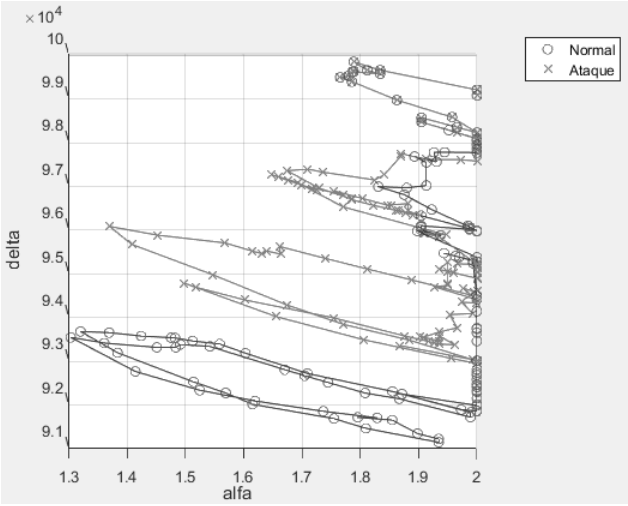


Fig. 8. Espacio de fases α, δ.

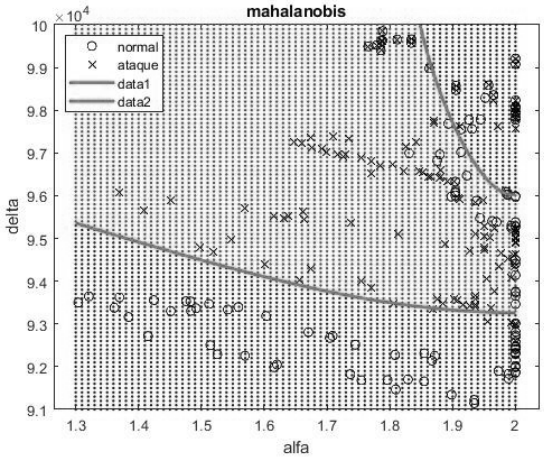


Fig. 9. Clusterización de tráfico sobre el espacio de fases α, δ.

como "normal" y "ataque". Para ello se han etiquetado los puntos correspondientes a las medidas y se han clusterizado utilizando la distancia de Mahalanobis [11]. El resultado puede apreciarse en la Fig. 9. El tráfico normal está representado con circunferencias y el mezclado con tráfico malicioso con aspas.

Se obtienen resultados similares usando algoritmos means y k-means, pero la mejor la matriz de confusión del caso estudiado se obtiene para el caso de Mahalanobis. Los resultados se pueden ver en la Fig. 10.

Se puede observar que, para el caso estudiado, la identificación del tráfico en el momento del ataque es razonablemente eficaz. Las etiquetas "normal1" y "normal2" se corresponden con las dos áreas (color azul) de tráfico sin ataque (superior derecha e inferior) que se pueden ver en la Fig. 9.

VII. CONCLUSIONES

El trabajo realizado ha mostrado un caso concreto en el que la utilización combinada de los parámetros de la distribución α-estable ajustada al tráfico permite identificar ciberataques del tipo DoS que no sean observables como mesetas en las series temporales de tráfico. A diferencia

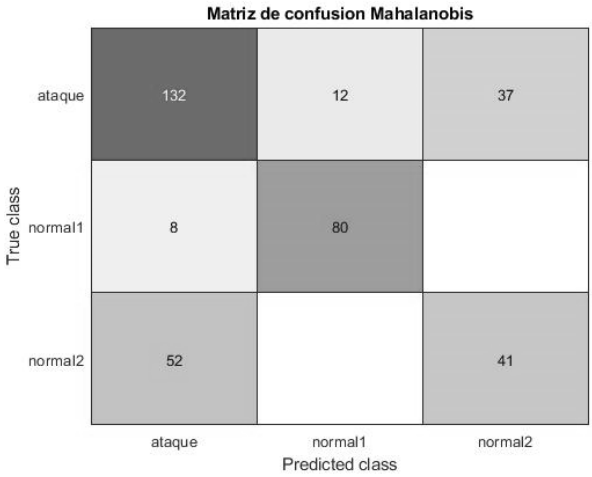


Fig. 10. Matriz de confusión con clusterización Mahalanobis.

de otros métodos de detección, es muy difícil simular los estadísticos de una distribución, por lo que esta vía de investigación se revela prometedora en el futuro próximo. Adicionalmente, el uso de modelos multivariantes, como el test de Hotelling T^2, puede mejorar la fiabilidad del método propuesto y reducir la tasa de falsos positivos [5].

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto TRÁFICA (MINECO/FEDER TEC2015-69417-C2-1-R).

REFERENCIAS

[1] "Prevención de intrusiones". Ayuda de Kaspersky Internet Security. https://help.kaspersky.com/KIS4Mac/16.0/es.lproj/pgs/88075.htm consultado el 23 de mayo de 2019

[2] J. Raiyn, "A survey of Cyber Attack Detection Strategies", *Int. J. Security and Its Applications*, Vol.8, No.1, 2014, pp.247-256

[3] J. Blasco Bermejo, "Ataques DoS en aplicaciones Web", OWA Spain Charter Meeting, viernes 6 de julio de 2007.

[4] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, S. Lopez-Buedo, "Characterization of the busy-hour traffic of IP networks based on their intrinsic features", *Computer Networks*, Vol. 55, No. 9, pp. 2111-2125, junio 2011, Elsevier.

[5] A. A. Sivasamy, B. Sundan, "A Dynamic Intrusion Detection System Based on Multivariate Hotelling's T^2 Statistics Approach for Network Environments". *The Scientific World Journal*, Vol. 2015, 850153, Hindawi.

[6] F. Simmross-Wattenberg, J. I. Asensio-Perez, P. Casaseca-de-la-Higuera, M. Martin-Fernandez, I. A. Dimitriadis and C. Alberola-Lopez, "Anomaly Detection in Network Traffic Based on Statistical Inference and α-Stable Modeling", *IEEE Trans. Dependable and Secure Computing*, Vol. 8, No. 4, pp. 494-509, jul-ago. 2011.

[7] J. P. Nolan. "Stable Distributions" Chapter 1, enero 2018

[8] G. Maciá Fernández, J. Camacho, R. Magán-Carrión, M. Fuentes-García, P. García-Teodoro, "UGR'16: Un nuevo conjunto de datos para la evaluación de IDS de red", Actas de las XIII Jornadas de Ingeniería Telemática - JITEL2017, Valencia, 27-29 sep. 2017.

[9] J. L. García-Dorado, J. E. López de Vergara, J. Aracil, V. López, J. A. Hernández, S. López-Buedo, L. de Pedro, "Utilidad de los flujos NetFlow de RedIRIS para análisis de una red académica", *Boletín de RedIRIS*, número 82-83, abril 2008.

[10] M. Stoppa, J. E. López de Vergara, F. Simmross-Wattenberg, J. L. García-Dorado, "Comparativa entre distribuciones α-estables para modelar tasas de transferencia obtenidas a partir de registros de SNMP y NetFlow", Actas de las XI Jornadas de Ingeniería Telemática, Jitel'2013, Granada, 28-30 oct. 2013.

[11] S. Xiang, F. Nie, C. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification", *Pattern recognition*, Vol. 41, No. 12, pp 3600-3612, 2018, Elsevier.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Escenario virtual inmersivo basado en web para el consumo de contenido omnidireccional

Jair López, Dani Marfil, Fernando Boronat, Javier Pastor
Departamento de Comunicaciones,
Universitat Politècnica de València, Campus de Gandia
C/Paraninf 1, 46730, Grao de Gandia, Valencia (Spain).
{jailogu@epsg, damarre@dcom, fboronat@dcom, fjpastor@dib}.upv.es

***Resumen*- En este artículo se presenta el diseño y desarrollo de un escenario virtual inmersivo basado en web para el consumo de contenido omnidireccional, incluyendo un reproductor capaz de reproducir tanto vídeo 360 como contenido tradicional dentro del mismo. Se trata de un entorno multiplataforma, *responsive* y con una interfaz intuitiva y amigable para el usuario, ya que permite un control sin la necesidad de dispositivos adicionales (como p.ej., un ratón o un joystick). Además, la compatibilidad directa con diferentes tipos de dispositivos (p.ej., portátiles, *tablets*, *smartphones* o *Head Mounted Displays* -HMDs-) ha sido uno de los objetivos principales. El escenario inmersivo es capaz de renderizar vídeo 360 en diferentes tipos de proyecciones (equirectangular o ERP y cúbica o CMP) y soporta streaming adaptativo de contenido basado en HTTP (como p.ej., DASH).**

***Palabras Clave*- contenido omnidireccional, contenido 360, multiplataforma, dash, realidad virtual, escenarios inmersivos, HTML5, Javascript.**

I. INTRODUCCIÓN

Los dispositivos *Head Mounted Displays* (HMD), son dispositivos de visualización que permiten reproducir contenido multimedia sobre una pantalla muy cercana a los ojos del usuario. Ejemplos de este tipo de dispositivos son las Oculus Rift[1], HTC Vive[2] o Samsung Gear VR[3]. Su uso y la disponibilidad de vídeo omnidireccional (también denominado vídeo 360) está actualmente creciendo de manera significativa. Tal y como se expone en [1] (Noviembre, 2018), un 11% de ciudadanos estadounidenses afirman poseer un dispositivo HMD en sus hogares. En agosto de 2017, el dato era del 7%, lo cual supone un crecimiento del 4%. Observando esta tendencia, se puede considerar, que estas nuevas tecnologías y dispositivos están comenzando a adquirir un mayor protagonismo en el mercado y se espera que la demanda de contenidos omnidireccionales y de dispositivos capaces de reproducirlos va a seguir aumentando en los próximos años. Por tanto, es necesario desarrollar aplicaciones que ofrezcan soporte a todos los formatos existentes de contenido audiovisual, con el fin de que sea el usuario quien decida el nivel de inmersión que quiere experimentar durante el consumo de dicho contenido. Por ejemplo, que el usuario tenga la posibilidad de consumir el contenido en un formato omnidireccional o bien que pueda elegir el punto de vista deseado de una producción audiovisual.

Los HMD permiten la visualización de contenido omnidireccional haciendo uso de mecanismos capaces de detectar la rotación del dispositivo (p.ej., acelerómetros y giroscopios). Dichos mecanismos permiten una navegación dinámica alrededor del contenido, en función de los movimientos de la cabeza del usuario. Los usuarios de HMD sólo pueden visualizar en todo momento una pequeña sección respecto al total del contenido 360 de la escena (Fig. 1). Esta sección se conoce como campo de visión o FoV (del inglés, *Field of View*).

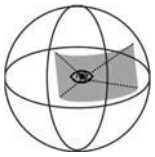


Fig. 1. Campo de visión o FoV.

[1] https://www.oculus.com
[2] https://www.vive.com
[3] https://www.samsung.com/es/wearables/gear-vr/

Por consiguiente, puede deducirse que el consumo de este tipo de contenidos a través de dispositivos HMD dota al usuario de cierto grado de libertad. De hecho, existe una clasificación respecto a los diferentes grados de libertad (*Degrees of Freedom*, DoF), definidos en el estándar MPEG-I [2]: 3DoF, 3DoF+ y 6DoF. Todos ellos se basan en la rotación y desplazamiento sobre los ejes X, Y, Z, tal y como se muestra en la Fig. 2. Como puede observarse en dicha figura, la rotación en los diferentes ejes se denomina *yaw*, *pitch* y *roll.*

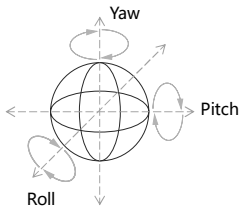


Fig. 2. Representación de los 6 grados de libertad.

El primero (3DoF) se trata del grado de libertad que permite modificar la orientación y campo de visión del contenido, pero desde un punto de vista estático. El segundo (3DoF+), es similar al primero, con la diferencia de que, en este caso, el usuario sí que puede desplazarse sobre los 3 ejes, aunque de una forma muy limitada. Finalmente, el tercero (6DoF) se trata de contenido en el que se puede navegar libremente por el escenario.

En este artículo se presenta un escenario virtual inmersivo basado en web para el consumo de contenido omnidireccional, incluyendo un reproductor de contenido omnidireccional. Además de permitir visualizar el contenido 360 a través de dispositivos HMD o PCs, el escenario incluye la posibilidad de consumir el campo de visión principal del contenido dentro del mismo. El artículo sigue la siguiente estructura: en la Sección II se describen trabajos relacionados y el actual estado del arte en este sector. En la Sección III se presenta el diseño del escenario inmersivo. En la Sección IV se describen las tecnologías utilizadas para su implementación. Finalmente, en la Sección V se exponen las conclusiones y se definen algunos puntos a seguir como trabajo futuro.

II. Estado del arte

En esta Sección se van a exponer los procesos actuales que reciben los vídeos 360 para su codificación y transmisión. Además, se analizarán los reproductores 360 que actualmente están disponibles. Finalmente, en la Tabla I se resumen las principales características que presentan en comparación al que se propone en este trabajo.

Debido a la gran variedad de tipos de contenido multimedia existentes, que han ido evolucionando desde la proyección 2D tradicional hacia otros tipos más inmersivos como el contenido omnidireccional, se requiere de un mayor ancho de banda y de nuevas tecnologías que permitan una codificación de contenido de mayor calidad. Por una parte, a pesar de utilizar comúnmente tecnologías de *streaming* adaptativo (p.ej., MPEG-DASH, Dynamic Adaptive Streaming over HTTP [3]) con este fin, éstas aún cuentan con un margen de mejora respecto a la demanda del ancho de banda. Esto se debe a que, para este tipo de contenidos, el usuario siempre estará consumiendo una pequeña parte del contenido completo (esto es, el FoV). Con el objetivo de especificar los requisitos de este nuevo tipo de contenidos, se ha presentado el estándar MPEG-I [2]. El principal objetivo de dicho estándar es la regulación de contenidos multimedia inmersivos. Existen numerosas características que permiten diferenciar los contenidos tradicionales 2D respecto a los contenidos omnidireccionales, como, por ejemplo, la adopción de nuevas técnicas de codificación para contenido de muy alta resolución (p.ej., *High Efficiency Video Coding* o HEVC [4]) o la utilización de dispositivos de consumo específicos, como los HMDs. De hecho, en [5], se adopta la tecnología HEVC para crear dos versiones "troceadas" (del inglés, *tiled*) de contenido de Realidad Virtual (en adelante, VR). Esto permite que se presente en la calidad más alta únicamente el contenido que está dentro del FoV del usuario y el resto se entregue en una calidad más baja. Así, el ahorro de ancho de banda aumenta entre un 30-40%, tal y como se expone en dicho artículo. Otra tecnología que permite una transmisión más eficiente del contenido 360 es la corrección (del inglés, *amendment*) 23009-1:2014 para la descripción de la representación espacial (*Spatial Representation Description*) DASH-SRD [6]. De hecho, en [7] se hace uso de lo especificado en dicho documento para proporcionar una solución más eficiente a la ocupación del ancho de banda durante la transmisión de contenido 360. En ese artículo, trabajos experimentales llevados a cabo muestran un ahorro de hasta un 72% del ancho de banda respecto a la no adopción de este tipo de mecanismos.

Por otra parte, durante la generación del contenido 360, la manera en la que se proyecta este contenido en un plano 2D cuenta con una complejidad añadida [8]. Durante el proceso de generación, pueden surgir ciertos problemas relacionados con la distorsión de la imagen. Tal y como se expone en [9], existen numerosas técnicas para proyectar el contenido 360 en un plano bidimensional. Estas técnicas de proyección pueden dividirse en dos, según sean independientes del punto de vista del consumidor (p.ej., la proyección equi-rectangular ERP o la proyección cúbica CMP, representadas en la Fig. 3) o no, como la proyección piramidal propuesta por Facebook en [10] (Fig. 4).

Según [10] la proyección CMP reduce un 25% el tamaño del archivo frente al ERP. La proyección piramidal reduce un 80% el tamaño del archivo frente al original, pero su coste de almacenamiento no es aceptable.

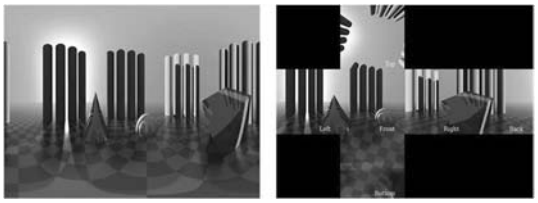


Fig. 3. Ejemplo[4] de proyección ERP (izquierda) y CMP (derecha).

[4] Imágenes obtenidas de https://docs.unity3d.com/Manual/VideoPanoramic.html

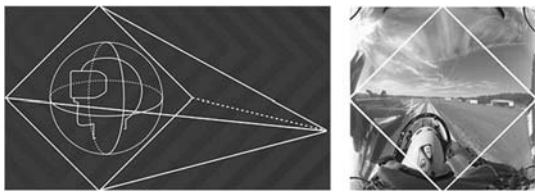


Fig. 4. Proyección piramidal dependiente del FoV [10].

Respecto a sistemas que permiten el consumo de contenido inmersivo y omnidireccional, en [11] se puede encontrar el reproductor *Omnivirt*, en el que se deben almacenar los vídeos que se deseen consumir en su propia nube y cuyo uso gratuito está restringido a ciertas funcionalidades (p.ej., una capacidad de almacenamiento máxima de 2GB o un acceso máximo mensual de 10K visitas). Dicho reproductor emplea la proyección ERP y hace uso de *streaming* adaptativo (HLS y MPEG-DASH). Para su utilización en dispositivos HMD, este reproductor habilita la visión estereoscópica aunque no realiza un desplazamiento de la imagen para cada ojo. Dicho desplazamiento sirve para simular cierta sensación de profundidad. El reproductor *JW Player* [12], también es compatible con el *streaming* adaptativo (HLS y MPEG-DASH), el tipo de proyección compatible es la ERP y ofrece un modo estereoscópico. Además, aloja los contenidos en sus propios servidores, por lo que el servicio tiene coste. El reproductor que ofrece *Bitmovin* [13] permite alojar el contenido en cualquier servidor, aunque se requiere de una licencia para utilizar el reproductor. También permite reproducir el contenido vía *streaming* adaptativo (HLS y MPEG-DASH). Ofrece visión estereoscópica con desplazamiento de la imagen en cada ojo. También acepta la proyección ERP.

Todos los reproductores que se han descrito hasta el momento no requieren de ningún tipo de instalación por parte del usuario, puesto que están basados en tecnología web. Sin embargo, hay otros reproductores que sí requieren de una instalación previa, como, por ejemplo, el *VLC Player* [14], que es gratuito para el usuario y, además, permite la reproducción de contenido del explorador de archivos. Sin embargo, no ofrece un modo estereoscópico. Dicho reproductor, también admite la proyección ERP y ofrece compatibilidad con *streaming* adaptativo (HLS y MPEG-DASH). De forma similar, fabricantes de cámaras para generar contenido omnidireccional, ofrecen sus propios reproductores para el consumo de contenido omnidireccional, tal es el caso de, *GoPro VR* [15]. El reproductor de *GoPro VR* [15] es gratuito. Es compatible con la proyección ERP. Solamente puede reproducir contenidos del explorador de archivos, por lo que no soporta *streaming* adaptativo. Además, no ofrece el modo de visión estereoscópica. Por otra parte, *Youtube* [16] también ofrece de un reproductor gratuito. Es compatible con el *streaming* adaptativo (HLS y MPEG-DASH), aunque se necesita instalar su aplicación móvil para poder consumir este tipo de contenidos a través de un móvil o HMD. Dicho reproductor proporciona visión estereoscópica con desplazamiento de la imagen en cada ojo y admite la subida de video 360 con proyección ERP, aunque internamente lo convierten a CMP.

También, existe una gran variedad de fabricantes de dispositivos HMD que incorporan sus propios escenarios virtuales para consumir vídeo omnidireccional. Algunos ejemplos son los de *Samsung Gear VR* [17] o *HTC Vive* [18]. El reproductor Samsumg VR es gratuito, pero se debe adquirir el HMD Oculus Gear VR y un smartphone de Samsumg compatible, proporciona el modo de visión estereoscópica con desplazamiento de la imagen en cada ojo, pero no implementa streaming adaptativo. La interacción con los elementos del entorno virtual se realiza mediante un mando, propio de la compañía, o mediante el sensor de clics integrado en el propio HMD. La proyección que soporta es la ERP. De la misma manera, el reproductor *de HTC Vive* [18] ofrece un escenario de realidad virtual para el consumo del contenido, pero es necesario el HMD de HTC que debe estar conectado a un ordenador de elevadas prestaciones. Soporta *streaming* adaptativo y es compatible con la proyección ERP.

Cabe resaltar que todos los reproductores y dispositivos descritos en esta Sección son compatibles con contenidos de una resolución de 4K y, cuando se utilizan a través de un HMD o dispositivo móvil, no permite la interacción sin manos (p.ej., utilizando punteros y temporizadores para ejecutar algún evento). A continuación, se muestra la Tabla I, donde se realiza comparativa con el reproductor que se propone en este trabajo.

Tabla I
COMPARATIVA ENTRE LOS REPRODUCTORES 360 EXISTENTES Y EL PROPUESTO

Reproductor 360	Basado en tecn. web	Requiere instalación	Gratuito	4K	Acepta contenido de cualquier origen	Streaming Adaptativo	Modo estereoscópico
Propuesto	✓	✗	✓	✓	✓	✓	✓
Omnivirt	✓	✗	✗	✓	✗	✓	✓
JW player	✓	✗	✗	✓	✗	✓	✓
Bitmovin	✓	✗	✗	✓	✓	✓	✓
GoProVR	✗	✓	✓	✓	✓	✗	✗
VLC player	✗	✓	✓	✓	✓	✓	✗
Youtube	✓	✓	✓	✓	✗	✓	✓
HTC	✗	✓	✗	✓	✗	✓	✓
Samsung Gear VR	✗	✓	✗	✓	✗	✓	✓

Reproductor 360	Admite ERP	Admite CMP	Proporciona escenario VR 360	Navegación con sensores	Interacción con ratón (PC)	Interacción sin manos (HMD o móvil)
Propuesto	✓	✓	✓	✓	✓	✓
Omnivirt	✓	✗	✗	✓	✓	✗
JW player	✓	✗	✗	✓	✓	✗
Bitmovin	✓	✗	✗	✓	✓	✗
GoProVR	✓	✗	✗	✗	✓	✗
VLC player	✓	✗	✗	✗	✓	✗
Youtube	✓	✓	✗	✓	✓	✗
HTC	✓	✗	✓	✓	✓	✗
Samsung Gear VR	✓	✗	✓	✓	✓	✓

III. Escenario inmersivo

En esta Sección se va a describir el escenario inmersivo diseñado. Este escenario pretende emular un salón, donde hay una TV disponible en la que se muestra el punto de vista principal de un determinado contenido. Además, el escenario inmersivo cuenta con una serie de puntos de interacción a partir de los cuales se notifica al usuario la disponibilidad de contenido complementario (p.ej., otros puntos de vista del contenido que se está visualizando, la vista 360 del contenido que se está visualizando o contenido relacionado). La Fig. 5 expone las premisas de diseño del escenario inmersivo, en el que pueden observarse las diferentes funcionalidades de las que se disponen.

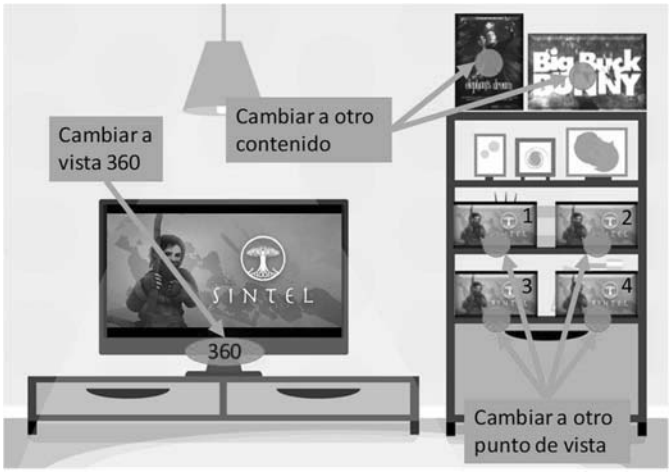


Fig. 5. Funcionalidades diseñadas en el escenario virtual.

A. Diseño y funcionalidades

A continuación, se van a describir en detalle las funcionalidades incluidas en el escenario inmersivo desarrollado en este trabajo. Para el consumo del contenido disponible, se ofrecen al usuario dos posibilidades diferentes.

Por un lado, se puede visualizar el contenido en una TV dispuesta de forma centrada en el propio escenario. En el escenario, se dispone de diferentes iconos gráficos o puntos interactivos, a través de los cuales el usuario tiene la capacidad de cambiar el contenido a visualizar (p.ej., puede cambiar a una de las múltiples vistas disponibles, a contenido omnidireccional, etc.) En caso de seleccionar contenido omnidireccional, dependiendo del dispositivo utilizado, este se podrá consumir en una pantalla plana (p.ej., PC, Tablet o móvil) o bien en un HMD. Para ello, el escenario permite el cambio entre las vistas de salón virtual inmersivo con TV incluida (donde visualizar los vídeos en 2D) y la vista de contenido 360 a pantalla completa o en un HMD, tal y como se muestra en la Fig. 6.

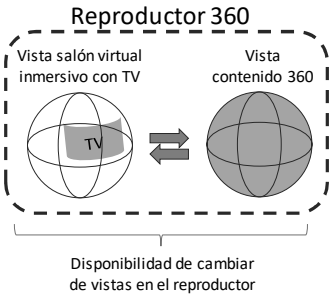


Fig. 6. Vistas disponibles (salón virtual con TV y visión 360).

B. Reproductor multiplataforma

Uno de los objetivos principales del reproductor incluido en el escenario creado es que sea fácilmente desplegable y compatible para el mayor número posible de consumidores. Por este motivo se ha implementado utilizando tecnologías web (que se detallarán en la Sección siguiente). Ello permite aumentar la compatibilidad con gran cantidad de dispositivos con rendimientos heterogéneos, desde ordenadores hasta móviles o HMD. Independientemente, la visualización de contenidos 360 será óptima si se realiza a través de dispositivos HMD. Es a través de estos dispositivos cuando se obtiene una mejor sensación de inmersividad y, por consiguiente, una calidad de experiencia del usuario más alta.

C. Streaming adaptativo de contenido 360

El reproductor integrado en el escenario inmersivo soporta el estándar de *streaming* adaptativo basado en DASH [3]. En definitiva, DASH permite la generación de diferentes representaciones (calidades) del mismo contenido. A su vez, dichas representaciones se trocean en pequeños segmentos

(p.ej., de 1s). Las representaciones y el tamaño de los segmentos, así como la ubicación de los mismos, quedan definidos en un fichero conocido como *Media Presentation Description* (MPD). De esta manera, el reproductor DASH puede conocer las diferentes calidades disponibles y, según el estado de la red, puede seleccionar un segmento de una mayor o menor calidad. Todas las representaciones (y por ende, los segmentos) se almacenan en un servidor HTTP (p.ej., en un servidor Apache[5]), puesto que es a través de este protocolo por el que el cliente DASH solicita el contenido.

Para que un cliente pueda reproducir contenido DASH, el primer paso es solicitar el MPD, ya que en él se encuentra la información necesaria para conocer y acceder al contenido en sus distintas calidades. La Fig. 7 muestra los módulos involucrados en la transmisión y reproducción de contenidos que están basados en la tecnología DASH.

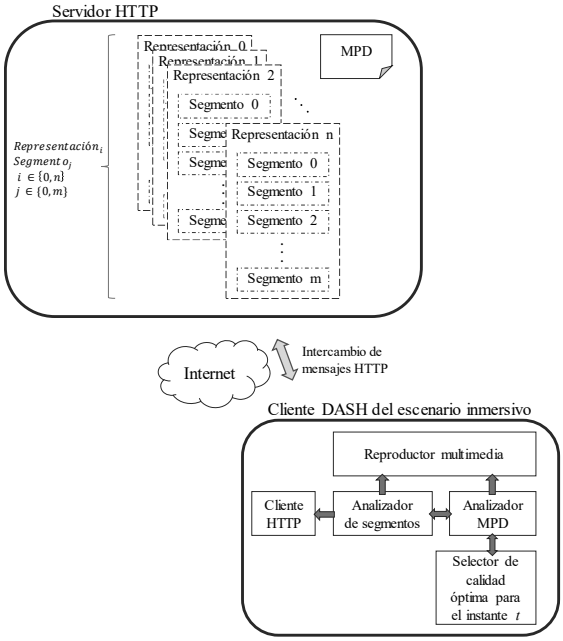


Fig. 7. Diferentes módulos involucrados en la entrega y el consumo de contenido DASH.

D. Proyecciones compatibles

Para ambas vistas (estas son, el salón virtual inmersivo y el contenido 360), el reproductor debe soportar algún tipo de proyección, y de esta manera poder emular un entorno omnidireccional a partir de un contenido en un plano 2D (que es tal y como se reciben y decodifican). Como se ha enunciado anteriormente, se busca compatibilidad con el mayor número de dispositivos posibles. Este objetivo también se puede trasladar a la naturaleza del contenido, puesto que, tal y como se ha observado en la Sección II, no todos los reproductores soportan más de un tipo de proyección. En el reproductor que se presenta, se soportan las proyecciones ERP y CMP (Fig. 3) Se ha decidido implementar ERP porque es la más utilizada según se ha analizado en la Sección II y CMP porque reduce el tamaño del archivo de video, debido a que elimina la distorsión de la imagen que presenta la proyección ERP en la parte de los polos.

E. Salón virtual inmersivo con TV

El salón virtual se ha incluido en el reproductor como una vista para consumir contenido tradicional 2D y, adicionalmente, poder disponer de información respecto a otros contenidos relacionados disponibles. Se trata del escenario principal de este reproductor. A partir de él se puede acceder al contenido omnidireccional o bien seleccionar otro tipo de contenido 2D que estará posicionado de una forma amigable dentro del salón, para que sea visto por el usuario y no resulte intrusivo. Estas opciones estarán resaltadas con una serie de botones (con interacciones asociadas) para poder seleccionarlas, tal y como se ha expuesto anteriormente en la Fig.5.

En el salón virtual se deben incluir una serie de botones o indicadores cuando, para el contenido que se esté consumiendo en la TV, haya disponibles contenidos relacionados adicionales. En el ejemplo de la Fig. 5, se está consumiendo contenido *open-source* (Sintel[6] en la TV), por lo que aparecen disponibles otros contenidos relacionados que también lo son (Elephant's Dream[7] y Big Buck Bunny[8]). Dichos contenidos relacionados se muestran integrados en la estantería virtual, de manera que no resulta intrusivo ni molesto para el usuario. Además, de la misma manera, en dicha figura también aparecen como disponibles otros puntos de vista (asumiendo su existencia) del contenido que está siendo consumido. Por otra parte, en la parte inferior de la TV aparece un botón para cambiar a la vista 360 (asumiendo, de nuevo, que existe este formato para el contenido que aparece en la TV).

En la Fig. 8 puede observarse el escenario inmersivo que ha sido implementado. Muestra el escenario cuando se selecciona un contenido de Fórmula 1 que cuenta con una vista principal (en este caso frontal, visualizada en el TV virtual) y un contenido relacionado omnidireccional. Tal y como se puede observar, en el salón virtual se notifica al usuario de la existencia del contenido omnidireccional mediante un botón naranja con el texto "360".

Fig. 8. Salón virtual implementado visto desde un monitor de PC (imagen superior) o desde un HMD (imagen inferior).

[5] https://httpd.apache.org/
[6] https://durian.blender.org/
[7] https://orange.blender.org/
[8] https://peach.blender.org/

F. Navegación en el escenario. Control del UI.

El escenario virtual de consumo se ha diseñado teniendo en cuenta que el principal dispositivo, con el que se va a consumir el contenido disponible, va a ser el HMD.

Por un lado, se han diseñado mecanismos que permiten la interacción del usuario sin necesidad de utilizar las manos. En el caso de que se esté utilizando un HMD, se activa la visualización de un punto o cursor central, que se desplaza al mover la cabeza (el HMD). La interacción de este punto con los diferentes botones y controles, dispuestos en el escenario inmersivo, accionará eventos de control de la reproducción (p.ej., cambiar de contenido en el escenario inmersivo, pasar a la vista del contenido omnidireccional o volver al escenario).

Por razones de compatibilidad en el uso de otros dispositivos tradicionales, como p.ej. el PC, la interacción con los botones y controles también puede realizarse mediante el uso del ratón y los eventos asociados (esto es, un click). Asimismo, esta funcionalidad también permite que, si el HMD (o móvil) tiene asociado un mando (p.ej., un joystick contectado vía bluetooth al móvil o un controlador del HMD) también está habilitada esta opción, imitando el comportamiento del ratón.

G. Reproducción de contenido 360

El contenido 360 puede consumirse sí, tal y como se ha visto en la anterior subsección, está disponible y se selecciona (o acciona) el cambio a su vista asociada. Para esto, el reproductor debe cambiar el escenario y reproducir únicamente el contenido omnidireccional, ya que con él se ocupan los 360 grados de la escena y no es necesario recrear ningún escenario adicional. Por tanto, en esta vista el usuario verá exclusivamente el contenido omnidireccional seleccionado. La Fig. 9 muestra el visor para dispositivo con pantalla plana, como un PC, tablet o móvil (imagen superior) y para un HMD (imagen inferior). En el primer caso la interacción se realizaría mediante el ratón o equivalente (táctil en el caso de tablets o móviles), mientras que, en el segundo caso, la interación se puede realizar con movimientos de la cabeza (monitorizando los sensores del HMD) o con un mando de control (emulando a un ratón).

Fig. 9. Contenido omnidireccional visualizado en pantalla plana (imagen superior) o en un HMD, en modo esteroscópico (imagen inferior).

IV. Implementación y tecnologías utilizadas

En esta Sección se van a describir las tecnologías que subyacen al reproductor que se propone en este trabajo.

A. Tecnologías multiplataforma: HTML5 y Javascript

Tal y como se ha expuesto en la anterior Sección, el reproductor está basado en tecnologías web con el objetivo de ser multiplataforma y fácilmente accesible para el consumidor. Por consiguiente, se han adoptado los lenguajes HTML5 y Javascript para implementar el reproductor. Concretamente, se ha hecho uso de dos librerías Javascript que han permitido, por un lado, recrear los escenarios omnidireccionales (tanto el salón virtual inmersivo como el reproductor de contenido 360), y, por otro lado, soportar la reproducción de contenido adaptativo. Estas librerías son, respectivamente, *three.js* y *dash.js*. Cabe destacar que se ha implementado un diseño *responsive*. Esto significa que el escenario y, más concretamente, el reproductor integrado en el mismo podrá variar su diseño para ajustarse correctamente al dispositivo utilizado por el usuario. Esto implica que, por ejemplo, se detecte cúando se está consumiendo a través de un HMD o de un PC, para así ofrecer un modo estereoscópico o no, respectivamente (las Fig. 8 y 9 muestran el aspecto del escenario y del reproductor 360 para ambos casos), permitiendo proporcionar una experiencia de usuario satisfactoria. La Fig. 10 muestra las tecnologías adoptadas, que se explican a continuación.

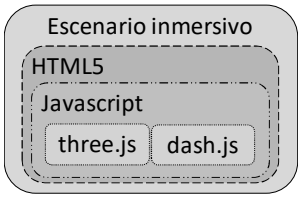


Fig. 10. Tecnologías utilizadas en la implementación del escenario inmersivo.

1. Three.js

Three.js [19], es una librería de uso libre que permite crear y mostrar gráficos animados por ordenador. Además, al estar basada en Javascript puede ser combinada con otros elementos del lenguaje HTML5 (p.ej., con el elemento *canvas*). Concretamente, esta librería se ha empleado para poder implementar el escenario inmersivo correspondiente al salón virtual y, adicionalmente, para proyectar el contenido omnidireccional en la vista 360. Además, esta librería es la encargada de cambiar la presentación del contenido a un modo estereoscópico, desplazando ligeramente una de las imágenes de uno de los ojos para proporcionar sensación de profundidad (si se utiliza un HMD) o presentar el contenido tal y como es (para el resto de los casos/dispositivos).

2. Dash.js

Dash.js [20] es una librería que permite que los navegadores soporten contenido recibido por *streaming* adaptativo basado en DASH [3]. En el reproductor propuesto, se ha adoptado esta tecnología para el contenido adaptativo, con el objetivo de realizar un consumo más eficiente del ancho de banda.

B. Implementación del salón virtual inmersivo con TV

Para la implementación del salón virtual, con *three.js*, por una parte, se ha creado una malla compuesta por una geometría esférica centrada en el origen de coordenadas y por un material cuya textura es la imagen 360 del salón. Para poder visualizarlo se ha creado una cámara en el centro de la esfera y se ha invertido las caras de la geometría. Por otra parte, para la simulación de la TV en el escenario, se ha creado otra malla compuesta por una geometría plana, que posteriormente ha sido curvada, y por un material en cuya textura se renderiza el video 2D inicial o cualquier video 2D seleccionado por el usuario (haciendo uso de la librería *dash.js*). Para la interacción del usuario con el escenario, por un lado, se crean objetos planos que siempre miran hacia la cámara (*sprites*) en donde se plasman materiales cuyas texturas son los iconos de botones o los posters de los videos disponibles. Por otro lado, para accionar dichos controles se hace uso del trazado de rayos para detectar los objetos que atraviesa (*raycaster*), es decir, cuando se hace clic en algún lugar del campo de visión, se crea un rayo en las coordenadas que proporciona el evento clic y si atraviesa algún *sprite* acciona su evento. En el caso del modo estereoscópico, el rayo se crea en el centro del campo de visión de cualquiera de los dos ojos y mediante el evento de giroscopio el usuario debe rotar el escenario acercando los *sprites* a ese punto para accionar su evento.

C. Implementación del reproductor de contenido 360

Para poder reproducir contenido 360, el usuario debe haber seleccionado esta opción en la vista del salón virtual inmersivo. Tras esto, con la librería three.js, se crea en la escena una malla que consta de una geometría cúbica o esférica(según los metadatos del video 360) en donde se renderiza un material cuya textura es el video omnidireccional obtenido vía streaming adaptativo (haciendo uso de la librería *dash.js*). En esta escena también se dispone de un botón que permite volver al salón virtual inmersivo. En la Fig. 9 puede observarse el aspecto de esta vista.

V. CONCLUSIONES Y TRABAJO FUTURO

En este artículo se ha propuesto un escenario virtual inmersivo basado en web y *responsive* para el consumo de contenido omnidireccional en un entorno intuitivo y amigable. Para ello, se ha dotado al escenario de un reproductor de contenido 360 y adaptativo. Se trata de un entorno multiplataforma al estar basado en tecnologías web (adoptando las librerías Javascript *three.js* y *dash.js*). Además, se ha descrito el funcionamiento de los diferentes eventos y puntos de interacción implementados en el escenario, incluyendo los módulos y tecnologías adoptadas para tal fin.

Como trabajo futuro se va a añadir soporte para otras tecnologías adaptativas como HLS [21], y de esta manera aumentar la compatibilidad con un número mayor de navegadores y dispositivos. También se va a implementar la técnica de transmisión *tiled streaming* [22], consistente en dividir el contenido omnidireccional en secciones (del inglés, *tiles*) y transmitir, en una mayor calidad, únicamente las regiones dentro del punto de vista del usuario, para así conseguir mayor eficiencia de consumo de ancho de banda. Además, se pretende integrar este reproductor en plataformas más complejas, como la descrita en [23].

AGRADECIMIENTOS

Este trabajo ha sido financiado, parcialmente, por la Generalitat Valenciana, bajo el programa de Subvenciones para Grupos de Investigación Consolidables, AICO/2017, con referencia AICO/2017/059.

REFERENCIAS

[1] YouGov, "YouGov: VR adoption grows to 11% of US adults – Digital TV Europe." [Online]. Available: https://www.digitaltveurope.com/2019/03/22/yougov-vr-adoption-grows-to-11-of-us-adults/. [Accessed: 01-Apr-2019].

[2] M. Domanski, O. Stankiewicz, K. Wegner, and T. Grajek, "Immersive visual media — MPEG-I: 360 video, virtual navigation and beyond," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2017, pp. 1–9.

[3] International Organization for Standardization, "ISO/IEC 23009-1:2014 Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats," 2014.

[4] International Organization for Standardization, "ISO/IEC 23008-2:2015. High efficiency coding and media delivery in heterogeneous environment - Part 2: High efficiency video coding." 2015.

[5] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "HEVC-compliant Tile-based Streaming of Panoramic Video for Virtual Reality Applications," in *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, 2016, pp. 601–605.

[6] O. A. Niamut, E. Thomas, L. D'Acunto, C. Concolato, F. Denoual, and S. Y. Lim, "MPEG DASH SRD," in *Proceedings of the 7th International Conference on Multimedia Systems - MMSys '16*, 2016, pp. 1–8.

[7] M. Hosseini and V. Swaminathan, "Adaptive 360 VR Video Streaming: Divide and Conquer," in *2016 IEEE International Symposium on Multimedia (ISM)*, 2016, pp. 107–110.

[8] L. Zelnik-Manor, G. Peters, and P. Perona, "Squaring the circle in panoramas," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005, pp. 1292-1299 Vol. 2.

[9] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," *Signal Processing*, vol. 146, pp. 66–78, May 2018.

[10] E. Kuzyakov and D. Pio, "Next-generation video encoding techniques for 360 video and VR," 2016. [Online]. Available: https://code.fb.com/virtual-reality/next-generation-video-encoding-techniques-for-360-video-and-vr/. [Accessed: 27-May-2019].

[11] "OmniVirt." [Online]. Available: https://www.omnivirt.com/. [Accessed: 23-May-2019].

[12] "JW Player." [Online]. Available: https://www.jwplayer.com/. [Accessed: 23-May-2019].

[13] "Bitmovin." [Online]. Available: https://bitmovin.com/. [Accessed: 23-May-2019].

[14] "VLC media player." [Online]. Available: https://www.videolan.org/vlc/index.html. [Accessed: 23-May-2019].

[15] "GoPro VR Player." [Online]. Available: https://gopro.com/news/gopro-vr-player-2-now-available. [Accessed: 23-May-2019].

[16] "YouTube." [Online]. Available: https://www.youtube.com/. [Accessed: 23-May-2019].

[17] "Samsung Gear VR." [Online]. Available: https://www.samsung.com/global/galaxy/gear-vr/. [Accessed: 23-May-2019].

[18] "HTC VIVE." [Online]. Available: https://www.vive.com/eu/. [Accessed: 23-May-2019].

[19] three.js, "three.js - Javascript 3D library." [Online]. Available: https://threejs.org/. [Accessed: 22-Mar-2019].

[20] DASH Industry Forum, "Dash.js." [Online]. Available: https://dashif.org/dash.js/. [Accessed: 24-May-2019].

[21] R. Pantos and E. W. May, "Http Live Streaming. RFC 8216." 2017.

[22] M. Graf, C. Timmerer, and C. Mueller, "Towards Bandwidth Efficient Adaptive Streaming of Omnidirectional Video over HTTP," in *Proceedings of the 8th ACM on Multimedia Systems Conference - MMSys'17*, 2017, pp. 261–271.

[23] D. Marfil, F. Boronat, A. Sapena, and A. Vidal, "Synchronization Mechanisms for Multi-User and Multi-Device Hybrid Broadcast and Broadband Distributed Scenarios," *IEEE Access*, vol. 7, 2019.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Composición de Cadenas de Servicios con un Algoritmo de Optimización Basada en Colonias de Hormigas

Antonio M. Mora, Segundo Moreno
Dto. Teoría de la Señal, Telemática y Comunicaciones
ETSIIT-CITIC, Universidad de Granada, España
amorag@ugr.es, segundomoto@correo.ugr.es

Javier Carmona-Murillo
Dto. Ingeniería de Sistemas Informáticos y Telemáticos
Universidad de Extremadura, España
jcarmur@unex.es

***Resumen*—El crecimiento del tráfico de datos y la demanda de nuevos servicios son dos de los principales retos a tener en cuenta en el diseño de las redes de próxima generación. SFC (*Service Function Chaining*) es una técnica que permite la ejecución de servicios avanzados, dirigiendo el tráfico de red a través de una lista ordenada de funciones virtuales. Este mecanismo está tomando gran importancia gracias al auge de las redes basadas en SDN (*Software-defined Networks*) y NFV (*Network Function Virtualization*), y a las posibilidades que ofrecen en términos de flexibilidad y automatización. Dada la necesidad existente de los operadores de ofrecer servicios de baja latencia en redes 5G, la composición de esta cadena es un proceso crítico que afecta al rendimiento de los propios servicios. Dentro de este contexto, en este trabajo se presenta el diseño e implementación de un algoritmo de Optimización basada en Colonias de Hormigas (OCH) para la minimización del coste de *routing* de cadenas de servicio. OCH es una metaheurística especialmente diseñada para trabajar en grafos con pesos, además teniendo en cuenta restricciones, como es el caso del problema que se aborda. Para probar la valía del algoritmo implementado se han resuelto dos instancias diferentes, una a modo de prueba de concepto (6 nodos), que permite analizar las soluciones obtenidas fácilmente y otra (19 nodos) cuyos resultados reflejan el comportamiento en una red más amplia. Los resultados obtenidos muestran que el algoritmo propuesto puede llevar a soluciones óptimas en muchos casos, además en un tiempo inferior a 0,5 segundos en la instancia mayor, por lo que consideramos dicho método como una solución muy prometedora en este campo.**

***Palabras Clave*—TF, RNG, VRS, SFC, NFV, SDN, Metaheurísticas, ACO, OCH**

I. INTRODUCCIÓN

A lo largo de los últimos años, las redes de comunicaciones están experimentando un cambio en la forma en la que los usuarios accedemos a Internet, pasando de los dispositivos fijos tradicionales a nuevos sistemas inalámbricos y móviles. Además, el enorme crecimiento y popularización de los dispositivos inteligentes, así como las nuevas aplicaciones y servicios disponibles como la realidad virtual, juegos *on-line* en tiempo real o el contenido multimedia UHD (Ultra-High-Definition), requieren de nuevos mecanismos en la red que proporcionen una muy alta velocidad y una baja latencia. En este contexto, las redes de quinta generación (5G) se plantean como una solución que ofrecerá nuevas capacidades a la red permitiendo el despliegue de nuevos servicios.

Para lograrlo, tanto la industria como la academia coinciden en que el motor de los sistemas 5G estará basado en un conjunto de tecnologías activadoras desde distintos niveles. Por una parte, deberán desarrollarse tecnologías de nivel físico que incrementen la capacidad de la red en bits/s/Hz mediante despliegues ultradensos (UDNs, *Ultra-Dense Networks*) en bandas de frecuencias altas. Por otra parte, y en un nivel superior de la pila de protocolos, el funcionamiento de la red recaerá en mecanismos emergentes [1] como las redes definidas por software (SDN, *Software Defined Networks*) o a través de la virtualización de funciones de red (NFV, *Network Function Virtualization*). A diferencia del modelo tradicional, donde las redes disponen de hardware especial para servicios específicos, en las redes 5G esta funcionalidad pasa del hardware al software gracias a SDN/NFV. Esto permite a los operadores mayor flexibilidad a la hora de desplegar funciones de red, así como una gestión más eficiente de los recursos disponibles. En este contexto, los servicios serán ofrecidos a través de una cadena ordenada de funciones (SFC, *Service Function Chain*) que será un conjunto de módulos de software virtualizados que se ejecutarán en distintos nodos de la red [2].

Este trabajo se enfoca en este último mecanismo, planteando la composición de estas cadenas de servicios como un problema NP-Completo [3]. Dicho problema se define a partir de un grafo que modela una red y en el que se despliegan una serie de nodos, cada uno de

los cuales ofrece una o varias funciones de red. A su vez, los enlaces entre los nodos tendrán un ancho de banda máximo asociado para las transmisiones. En este modelo, se recibirán diferentes peticiones o conexiones para satisfacer determinado servicio a un usuario, cada uno de los cuales se realizará componiendo una cadena de funciones siguiendo un orden determinado. Además se deberá considerar la capacidad en recursos de que dispone cada nodo, así como los recursos consumidos por cada función a ejecutar en dicho nodo.

De este modo el problema se transformará en uno de búsqueda de camino óptimo dentro de dicho grafo, para seguir la secuencia de funciones deseada minimizando el número de saltos en la red, esto es, un problema de *routing*, aunque múltiple (se deben alcanzar distintos destinos). Se tendrá un problema de este tipo por cada conexión a resolver, que dependerá de las conexiones ya resueltas anteriormente, ya que éstas actualizarán el estado de los nodos y de los enlaces de la red.

Consideraremos para resolver este problema una metaheurística denominada Optimización basada en Colonias de Hormigas (OCH) [4], dada su efectividad contrastada en problemas de búsqueda de camino óptimo en grafos [5], [6], [7], [8].

Así pues, en el trabajo se ha implementado una variante de OCH adaptado para la resolución de este problema con restricciones, que hemos bautizado como *Ant-SFC*. Posteriormente se ha validado su corrección en dos instancias diferentes del problema, una con 6 nodos a modo de prueba de concepto y otra con 19 nodos que modela un escenario más cercano a la realidad.

II. PROBLEMA A RESOLVER: ROUTING PARA SFC

Gracias a las tecnologías de virtualización, NFV ofrece un nuevo mecanismo para diseñar, gestionar y desplegar redes y servicios. Sin embargo, junto con las nuevas posibilidades, aparecen nuevos retos a resolver, ligados a la optimización en la utilización y ubicación de recursos en este tipo de redes. En este contexto, dos de los principales retos se centran en el desarrollo de técnicas para la ubicación dinámica en la red de las diferentes instancias de las funciones virtuales [9], así como en algoritmos para la optimización del routing en la red [10], [11].

En este trabajo, nos centramos principalmente en el segundo de los retos, y que está relacionado con *construcción de una cadena de servicio (SFC) en la red*, al que denominaremos *Optimización del Routing en SFC (OR-SFC)*. En este proceso, se deberá determinar el camino que deben seguir los datos entre las funciones de red virtuales adyacentes para cada uno de los servicios solicitados.

La Figura 1 muestra un ejemplo sencillo de este problema a resolver.

En él, la conexión viene definida por una tupla, *C=(origen,destino,valor de la demanda, [funciones a ejecutar])*, en la que la demanda tiene como origen el nodo 1 y como destino el nodo 6, con un tamaño de demanda de 2 Mbps. Si miramos al grafo, el valor que aparece sobre

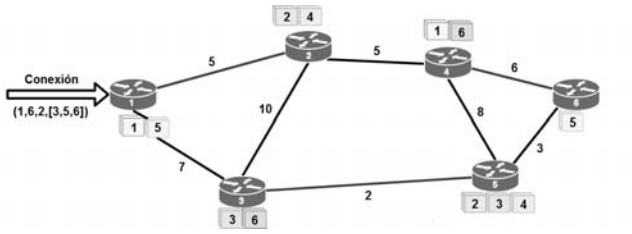


Fig. 1. Ejemplo de problema de Routing para SFC

los enlaces hace referencia al *ancho de banda* disponible en cada enlace, y en cada nodo se han señalado con cubos las funciones de red que pueden ejecutar. Además de eso, cada uno de los nodos tendrá asociados unos *recursos de computación* (CPU, Memoria, espacio en disco) agrupados en un único número por simplicidad. De la misma forma, se tendrá una lista de funciones que las asociará a cada una con unos requisitos en recursos para poder ser ejecutadas.

Por tanto, en este caso, se debe construir un camino que pase por las funciones 3, 5 y 6 (en ese orden estricto). Una solución podría ser el camino $[1 \rightarrow 3 \rightarrow 1 \rightarrow 2 \rightarrow 4 \rightarrow 6]$, pero también podría ser $[1 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 4 \rightarrow 6]$. La elección de una u otra alternativa afectará al coste de routing en la red, así como al rendimiento global de la misma. En este camino, habría que tener en cuenta determinadas restricciones, como que los enlaces y los nodos por los que pasa la conexión deben tener disponibles suficiente capacidad y recursos para soportar la demanda y los requisitos de ejecución, respectivamente.

III. ESTADO DEL ARTE

En la formación de cadenas SFC, uno de los mayores retos es la construcción del camino que debe seguir el tráfico y que permita ir ejecutando las diferentes funciones virtuales en el orden requerido. Una estrategia óptima supondrá un mayor aprovechamiento de los recursos de la red y, por tanto, un mayor rendimiento de la misma.

Este problema, que puede plantearse como un problema de optimización NP-Completo [3], ha está teniendo mucha atención por parte de la academia, pudiendo clasificar las distintas soluciones que lo abordan en dos categorías principales: soluciones exactas y algoritmos heurísticos.

La gran mayoría de soluciones exactas al problema OR-SFC se centran en modelos de optimización basados en técnicas de programación lineal. Por ejemplo, en [2] los autores desarrollan un modelo que resuelve el routing de SFC y la ubicación de las funciones virtuales para intervalos de horas puntas. O el trabajo desarrollado en [12], donde los autores formulan un modelo matemático exacto que resuelve la composición de cadenas mediante la descomposición de la misma, lo que permite minimizar uno de los principales inconvenientes de los trabajos propuestos en esta categoría, que es la incapacidad de resolver instancias de cualquier tamaño en un tiempo aceptable, debido al excesivo incremento de su complejidad.

La otra forma común de abordar este tipo de problemas, o como complemento de las anteriores, es mediante heurísticas (o metaheurísticas), que permiten obtener re-

sultados cercanos a los óptimos en tiempos de cómputo razonables. En muchos casos, estos algoritmos heurísticos se basan en algoritmos *greedy* como en [13] o en [14]. Hay que tener en cuenta que las heurísticas son más adecuadas en aquellos casos en los que tratemos con instancias grandes, en las que las soluciones exactas no son abordables y las heurísticas ofrecen un buen compromiso entre la calidad de la solución y el coste de cómputo [15]. Por ejemplo, el trabajo [16] presenta una propuesta de Algoritmo Genético para resolver el mismo problema que se plantea en este trabajo, pero no llega a mostrar una implementación y prueba del mismo, como aquí.

Tras la revisión de la literatura y, aunque las características del problema OR-SFC se adaptan bien a los problemas que puede resolver la técnica OCH, este trabajo es el primero en proponer una solución basada en Colonia de Hormigas para la optimización del routing en SFC. Si bien, este estudio pretende ser una primera propuesta de la que partir, que pueda conducir a otras soluciones más avanzadas en un futuro próximo.

IV. OCH

La Optimización basada en Colonias de Hormigas (OCH) es una metaheurística inspirada en el comportamiento colaborativo de algunas especies de hormigas para crear los caminos más cortos desde su nido a una fuente de alimento [17]. Este comportamiento se basa en la *estigmergia*, es decir, la comunicación entre agentes a través del medio en el que se encuentran. En este caso cada hormiga, mientras camina, deposita en el suelo una sustancia llamada *feromona* que las demás pueden detectar (oler). Cada una de ellas tiende a seguir (ante una bifurcación con varias posibilidades) las concentraciones más altas de dicha sustancia (la cual se evapora tras cierto tiempo) y a su vez hace su propio aporte. Esto acaba por constituir rastros de feromona que marcan el mejor camino (el más corto) entre el nido y la fuente de alimento.

Los algoritmos de OCH [5] se inspiran en este comportamiento para resolver problemas de optimización combinatoria, de modo que usan una colonia de *hormigas artificiales*, las cuales son agentes computacionales que se comunican entre sí usando una *matriz de feromonas*. Estos agentes trabajan en problemas formulados sobre un grafo con pesos en sus arcos. En cada iteración, cada hormiga construirá un camino completo (solución) moviéndose a través de él. Una vez construido dicho camino (o durante su construcción), la hormiga irá depositando un rastro de feromona que, generalmente, será función de la bondad de la solución que esté construyendo o haya construido. Por tanto, dicho rastro será una medida (informativa para las demás) de lo deseable que es seguir el mismo camino que la susodicha hormiga.

Para moverse en el grafo, cada hormiga manejará dos tipos de información: la mencionada feromona o *información memorística* y cierta *información heurística*, la cual depende del problema y se basa en el aprovechamiento de un conocimiento previo del mismo (cuyo valor no cambia durante la ejecución).

En su movimiento por el grafo, las hormigas elegirán normalmente los nodos con un mejor valor para las dos informaciones (las cuales se combinan), pero existirá un componente estocástico por el que una hormiga podrá moverse a nodos con valores menores, pues es posible que la solución final obtenida siguiendo esos nodos sea mejor. Se trata pues, de un componente para potenciar la exploración, la cual es fundamental para la resolución de problemas combinatorios complejos.

De esta forma todas las hormigas colaborarán para encontrar la mejor solución para el problema (el mejor camino dentro del grafo), lo que modela un comportamiento emergente global.

En este trabajo se ha utilizado una implementación básica de OCH, el llamado Sistema de Hormigas (*Ant System* en inglés) o SH [18], cuya estructura se puede ver en los Algoritmos 1 y 2.

Algorithm 1 SH ( )
Algoritmo principal de un SH

```
Inicializar_parametros()
while criterio_de_terminacion_no_satisfecho do
  for cada hormiga h do
    s[h]=Construir_Solucion(h)
  end for
  /* En todos los arcos del grafo */
  Evaporacion_de_Feromona()
  /* Sólo en arcos recorridos por la mejor hormiga */
  s*=Elegir_Mejor_Solucion(s[h])
  Actualizacion_Global_de_Feromona(s*)
end while
```

Como se puede ver, el algoritmo se basa en una repetición continua (hasta alcanzar un número determinado de iteraciones, normalmente) de un proceso sencillo, en el que cada hormiga construirá una solución recorriendo el grafo. Una vez todas hayan construido su solución, se actualizará la feromona de los arcos de dicho grafo mediante una evaporación (un decremento porcentual de su valor de feromona) y un posterior aporte o refuerzo que realizará únicamente la mejor hormiga sobre los arcos que constituyen su solución y de forma proporcional a la calidad de dicha solución. Con este proceso se irán reforzando los arcos más deseables del grafo para que las hormigas que lo exploren posteriormente tiendan a seguirlos también.

En la construcción de la solución los estados se corresponderían con los nodos que se pueden visitar en el grafo que se está explorando, algunos de los cuales no serán alcanzables desde el nodo actual en el que se encuentre una hormiga. La decisión se moverse a un nodo u otro, es decir, seguir un arco u otro, dependerá de una combinación del valor actual de feromona en el arco, de su valor heurístico (su deseabilidad considerando el problema a resolver) y de las posibles restricciones que tenga el problema. Con esto se asignará una probabilidad de selección de cada arco, que se considerará posteriormente en una política de selección del mismo. De modo que no siempre se elegirán

Algorithm 2 Construir_Solucion (id_hormiga)
Algoritmo de construcción de una solución por parte de una hormiga en un SH genérico

```
inicializar_hormiga(id_hormiga)
estado_actual = estado_inicial
L = guardar(estado_inicial) /* Lista de estados visitados */
while estado_actual ≠ estado_objetivo do
  /* A: lista de estados alcanzables, P: probabilidad de cada estado alcanzable, Ω: restricciones del problema */
  P = calcular_probabilidades_de_transicion(estado_actual, A, L, Ω)
  estado_siguiente = aplicar_politica_decision(P, Ω)
  mover_al_siguiente_estado(estado_siguiente)
  L = guardar(estado_siguiente)
  estado_actual = estado_siguiente
end while
```

los mejores arcos (según su probabilidad). Con esto el algoritmo consigue no estancarse en óptimos locales y elegir arcos menos prometedores en un punto determinado de la búsqueda para obtener presumiblemente una solución global mejor.

V. ALGORITMO IMPLEMENTADO: ANT-SFC

Como se ha comentado anteriormente, el algoritmo implementado para resolver este problema ha sido una adaptación de un Sistema de Hormigas [18] clásico para resolver un problema de camino óptimo múltiple en un grafo. Se han añadido una serie de restricciones de acuerdo con la definición del problema (ver Sección II), es decir:

- Se deberá definir un camino dentro del grafo que modela la red para resolver cada conexión (petición de servicio). Éste deberá pasar por nodos que sean capaces de servir cada una de las funciones de red requeridas en el orden definido para realizar el servicio.
- Cada enlace por el que pasen los caminos deberá tener ancho de banda restante suficiente para satisfacer la demanda de tráfico asociada a cada conexión. A este respecto, el ancho de banda se irá decrementando cada vez que un camino pase por un enlace.
- Los nodos deberán tener recursos suficientes para ejecutar la función deseada en cada momento. Del mismo modo que en el caso anterior, los recursos disponibles en el nodo se irán reduciendo de acuerdo a los requisitos que implique cada una de las funciones a ejecutar (según la conexión).

Dado que se trata de un problema de SFC, hemos denominado al algoritmo *Ant-SFC*. Se ha planteado la resolución de cada una de las *conexiones* (peticiones de servicios) como un problema de búsqueda de camino óptimo por separado, aunque son dependientes, debido a los consumos de ancho de banda en los enlaces y recursos en los nodos comentados previamente.

El cuerpo principal de Ant-SFC quedaría entonces como se define en el Algoritmo 3:

La adaptación el algoritmo se ha centrado en varios aspectos que se explican a continuación:

- *Inicialización de red*: Antes de que una hormiga comience la construcción de una solución, la red (anchos de banda y recursos) se inicializa al estado que tenía antes de que la hormiga anterior la modificara al construir su solución.
- *Heurística*: No se ha considerado una heurística que guíe la búsqueda como tal, pero sí se ha considerado asignar una mayor probabilidad de ser elegidos los enlaces con mayor ancho de banda disponible, para minimizar el riesgo de agotar alguno y dejar un nodo o subred incomunicados. Además, se ha incluido una condición en la selección del nodo siguiente al construir las soluciones y es que si alguno de los nodos es capaz de servir la siguiente función que se espera en la cadena, se duplicará la probabilidad de que ese nodo sea elegido.
- *Nodos alcanzables*: A se ha definido como la lista de aquellos nodos para los que existe un enlace desde el actual, dicho enlace tiene ancho de banda disponible suficiente (considerando la demanda de tráfico de la conexión) y dicho nodo tiene recursos suficientes para ejecutar la siguiente función en la cadena (si es que

Algorithm 3 Ant-SFC ( )
Algoritmo principal de Ant-SFC

```
Inicializar_parametros()
Leer_configuracion_red()
Leer_conexiones()
/* Se busca una solución por cada conexión */
for cada conexion c do
  while criterio_de_terminacion_no_satisfecho do
    for cada hormiga h do
      s[h]=Construir_Solucion(c,h)
    end for
    /* En todos los enlaces del grafo */
    Evaporacion_de_Feromona()
    /* Enlaces recorridos por la mejor hormiga */
    s*=Elegir_Mejor_Solucion(s[h])
    Actualizacion_Global_de_Feromona(s*)
  end while
  /* Se actualizan anchos de banda en los enlaces y
  recursos disponibles en los nodos */
  Actualizar_Red(c,s*)
end for
```

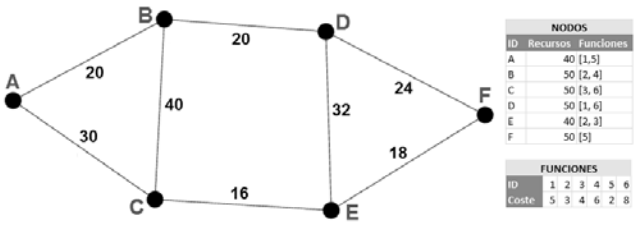
Algorithm 4 Construir_Solucion (conex, id_hormiga)
Algoritmo de construcción de una solución en Ant-SFC

```
inicializar_hormiga(id_hormiga)
inicializar_red() /* Fijar valores actuales de la red */
nodo_actual = conex.nodo_inicial
funcion_actual = conex.funciones[inicio]
L = guardar(nodo_actual) /* Lista de estados visitados */
F = guardar(funcion_actual) /* Lista de funciones servidas */
while (nodo_actual ≠ conex(nodo_final)) AND (funcion_actual ≠ conex.funciones[end]) do
  /* A: lista de nodos alcanzables, P: probabilidad de moverse a cada nodo alcanzable, Ω: restricciones del
  problema */
  P = calcular_probabilidades_de_transicion(nodo_actual, A, F, L, Ω)
  nodo_siguiente = ruleta_probabilidad(P, Ω)
  /* Se actualizan anchos de banda en los enlaces */
  Actualizar_Enlace(nodo_siguiente)
  L = guardar(nodo_siguiente)
  nodo_actual = nodo_siguiente
  /* Si está la función se sirve y se actualizan los recursos en el nodo */
  if funcion_actual in nodo_actual.funciones[] then
    Actualizar_Nodo(funcion_actual)
    F = guardar(funcion_actual)
    funcion_actual = conex.siguiente(funciones[])
  end if
end while
```

la puede servir).

- *Ruleta de probabilidad*: Una vez asignada la probabilidad de moverse a cada nodo desde el actual en la construcción de una solución, se utilizará una ruleta de probabilidad como política de decisión de estado siguiente. Dicha ruleta consistirá en la asignación de un espacio proporcional a la probabilidad de cada nodo en una 'ruleta virtual' y en su giro aleatorio para obtener el nodo elegido.
- *Restricción de ancho de banda en el enlace*: En esta restricción se basa la construcción de la lista de nodos alcanzables. Se ha comentado en el punto anterior.
- *Restricción de recursos en el nodo*: Igualmente, en esta restricción se basa la construcción de la lista de nodos alcanzables. Se ha comentado en el punto de Nodos Alcanzables.
- *Actualizaciones de enlaces/nodos (construcción de solución)*: Cada vez que una hormiga se mueve a un nodo en la red mientras construye una solución (un camino) para resolver una conexión concreta, se actualiza el ancho de banda disponible en el enlace (con la demanda de tráfico de la conexión) y se actualizan los recursos restantes del nodo, si éste sirve una función concreta (con el coste que tenga esa función en recursos). De esta forma se evita que se produzcan bucles infinitos en la construcción de una solución.
- *Actualizaciones de red (conexiones)*: Cada vez que se encuentra una solución para una cadena determinada, la red se actualiza considerando el camino definido en la solución. Por tanto los enlaces y nodos que atraviesa dicho camino se actualizan siguiendo el mismo esquema que en el caso anterior.
- *Restricción de camino completo*: Sólo se considerará como válida una solución/camino si ésta comienza y termina en los nodos indicados en la conexión y si dicho camino pasa por nodos que sirvan las funciones en el orden estipulado por dicha conexión. Cualquier solución que no cumpla estos criterios será descartada.
- *Coste de un camino (conexión)*: El coste que consideraremos para un camino será el *número de saltos* requeridos en el grafo para componer la cadena de funciones necesaria para satisfacer una conexión.
- *Coste global de una solución*: Una solución completa consistirá en la unión de varios caminos mínimos, uno para resolver cada una de las conexiones que se hayan solicitado para una instancia concreta en un intervalo de tiempo determinado. Por tanto, el coste de una solución será la suma de los de todas las conexiones que contenga.
- *Actualización de Feromona*: Aparte de la *evaporación de feromona* pertinente realizada a todos los enlaces del grafo tras la construcción de todas las soluciones (una por cada hormiga) para resolver una conexión determinada, se efectuará un *aporte de feromona* únicamente en los enlaces de la mejor solución (el camino de menor coste de los encontrados) que será proporcional al coste en saltos de dicho camino. De modo que a menor coste total mayor será el aporte realizado en dichos enlaces.

Una vez descrito el algoritmo Ant-SFC, en la sección siguiente evaluaremos y analizaremos su utilidad y valía en dos experimentos.



NODOS		
ID	Recursos	Funciones
A	40	[1,5]
B	50	[2, 4]
C	50	[3, 6]
D	50	[1, 6]
E	40	[2, 3]
F	50	[5]

FUNCIONES						
ID	1	2	3	4	5	6
Coste	5	3	4	6	2	8

Fig. 2. Instancia de 6 nodos: Topología de la red, nodos y enlaces con su ancho de banda disponible. Nodos con funciones asociadas y recursos totales disponibles para ejecutarlas. Funciones con su coste en recursos requerido.

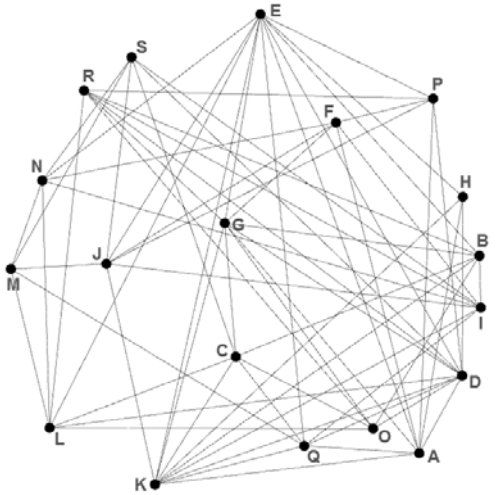
VI. EXPERIMENTOS Y RESULTADOS

En esta sección describiremos en primer lugar las dos instancias abordadas para probar el algoritmo. Posteriormente mostraremos y analizaremos los resultados obtenidos en los experimentos llevados a cabo.

A. Descripción de Instancias

Se han evaluado dos instancias diferentes del problema:

- Instancia 6N: se trata de un grafo de 6 nodos, que se ha usado a modo de prueba de concepto y en el que se ha podido evaluar y validar el algoritmo construido de manera más intuitiva. Las características de esta instancia se pueden ver en la Figura 2, incluyendo las funciones de red que puede satisfacer cada nodo y el ancho de banda asociado a cada enlace.
- Instancia 19N: se trata de un grafo de 19 nodos, que modela un escenario mucho más cercano a los que se resolverán en la realidad, si bien es cierto que la distribución de las cadenas de servicios en la topología de la red serán más complejas de lo que tendríamos en un escenario real, a fin de probar mejor la utilidad del algoritmo propuesto. Las Figuras 3 y 4 muestran la topología, así como las características de nodos y enlaces de la instancia.



NODOS		
ID	Recursos	Funciones
A	25	[5, 2, 1]
B	26	[1]
C	22	[5]
D	26	[4]
E	27	[2, 5, 1]
F	25	[3]
G	33	[1, 3]
H	22	[5, 4]
I	29	[4, 2, 3]
J	29	[1, 4]
K	29	[1]
L	33	[2, 5, 3]
M	22	[3, 5, 4]
N	39	[3, 5, 2]
O	24	[1, 4, 2]
P	25	[5, 4]
Q	26	[3]
R	25	[2, 3, 4]
S	23	[2]

FUNCIONES					
ID	1	2	3	4	5
Coste	7	7	6	10	8

Fig. 3. Instancia de 19 nodos: Topología de la red. Nodos con funciones asociadas y recursos totales disponibles para ejecutarlas. Funciones con su coste en recursos requerido.

ENLACES			
Nodo Orig	Nodo Dest	Peso	Ancho Banda
A	D	1.0	186
A	E	1.0	105
A	F	1.0	113
A	G	1.0	146
A	H	1.0	185
A	K	1.0	116
A	P	1.0	101
A	Q	1.0	137
B	C	1.0	143
B	E	1.0	190
B	G	1.0	141
B	I	1.0	173
B	K	1.0	142
B	O	1.0	147
B	R	1.0	113
C	G	1.0	132
C	K	1.0	164
C	L	1.0	184
C	O	1.0	118
C	Q	1.0	178
C	S	1.0	145
D	E	1.0	180
D	G	1.0	123
D	H	1.0	122
D	K	1.0	139
D	L	1.0	127
D	O	1.0	165
D	P	1.0	179
D	Q	1.0	111
D	R	1.0	137
D	S	1.0	188
E	G	1.0	160
E	H	1.0	200
E	J	1.0	186
E	K	1.0	150

ENLACES			
Nodo Orig	Nodo Dest	Peso	Ancho Banda
E	L	1.0	195
E	N	1.0	194
E	P	1.0	176
E	Q	1.0	120
F	G	1.0	164
F	I	1.0	189
F	J	1.0	136
F	N	1.0	136
F	P	1.0	125
G	K	1.0	150
G	O	1.0	114
G	R	1.0	172
H	K	1.0	194
I	J	1.0	193
I	K	1.0	108
I	N	1.0	112
I	Q	1.0	155
I	R	1.0	162
I	S	1.0	104
J	K	1.0	100
J	M	1.0	110
J	P	1.0	162
J	S	1.0	111
K	Q	1.0	141
L	M	1.0	160
L	N	1.0	176
L	O	1.0	135
L	R	1.0	196
M	N	1.0	115
M	Q	1.0	159
M	S	1.0	161
N	S	1.0	166
O	R	1.0	191
P	R	1.0	124

Fig. 4. Instancia de 19 nodos: Anchos de banda asociados a los enlaces.

Tabla I
PARÁMETROS CONSIDERADOS EN LOS EXPERIMENTOS.

Parámetro	Instancia 6N	Instancia 19N
$Num.Iteraciones$	6	19
$Num.Hormigas$	12	38
α (peso feromona)	$1,2$	$1,2$
β (peso heurística)	$2,0$	$2,0$
ρ (factor evaporación)	$0,3$	$0,3$

B. Resultados Obtenidos

Para la ejecución del algoritmo se ha utilizado una máquina con procesador Intel Core i7 4510-U de 2 núcleos y 4 hilos a 2.00 GHz, 8GB de memoria RAM DDR-3, con S.O. Windows 10 de 64 bits.

La configuración utilizada en el algoritmo en cada una de las instancias se muestra en la Tabla I.

Dichos valores han sido fijados a partir de recomendaciones de la literatura sobre OCH, como los pesos de feromona y heurística en el cálculo de la probabilidad de elección del nodo siguiente, aunque posteriormente han sido ajustados mediante experimentación sistemática. El número de iteraciones y conexiones se ha fijado de forma que se obtengan buenas soluciones en un tiempo aceptable.

En primer lugar se mostrarán los resultados obtenidos para la instancia 6N, considerando 3 conexiones a resolver, en concreto (ver formato en Sección II):
- Conexión 1: (A, F, 2, [3,5,6])
- Conexión 2: (A, E, 8, [1,2,4])

- Conexión 3: (A, D, 5, [2,4,5])

Dado que se trata de un algoritmo no determinista, se han realizado 10 ejecuciones independientes del mismo resolviendo la misma instancia del problema (con las mismas conexiones). En la Tabla II se muestran los resultados obtenidos.

Tabla II
RESULTADOS PARA LA INSTANCIA 6N CON TRES CADENAS DE CONEXIÓN. COSTE DE ROUTING REQUERIDO PARA RESOLVER CADA UNA DE LAS CONEXIONES EN CADA EJECUCIÓN Y COSTE GLOBAL DE LA SOLUCIÓN, ADEMÁS DE LA MEDIA Y DESVIACIÓN TÍPICA DE ÉSTE. EN AZUL Y NEGRITA LA MEJOR SOLUCIÓN OBTENIDA

Ejc	Conex1	Conex2	Conex3	COSTE
1	8	3	5	16
2	12	4	4	20
3	11	4	4	19
4	**5**	**4**	**4**	**13**
5	7	4	4	15
6	8	5	5	18
7	6	3	5	14
8	7	3	4	14
9	12	4	6	22
10	20	3	4	27
			Media	17,8
			Desv. Típica	4,37

Como se puede ver en la tabla todas las ejecuciones llegan a una solución bastante competente, siendo todas ellas además válidas, lo cual es un requisito para el algoritmo. La mejor solución obtenida, con 13 saltos se podría considerar óptima para este grafo e igual a la que un humano podría definir, con cierto esfuerzo en los cálculos. Los caminos para resolver las conexiones de dicha solución se han proyectado en la Figura 5. La desviación típica, eso sí, resulta un poco alta, lo cual es un indicativo de que se debería definir algún mecanismo que mejore el desempeño del algoritmo en cuanto a la calidad de la solución obtenida, como una heurística que realmente guíe la búsqueda o algún mecanismo de búsqueda local [19], que han demostrado ampliamente en la literatura su efectividad para mejorar el rendimiento de todo tipo de algoritmos.

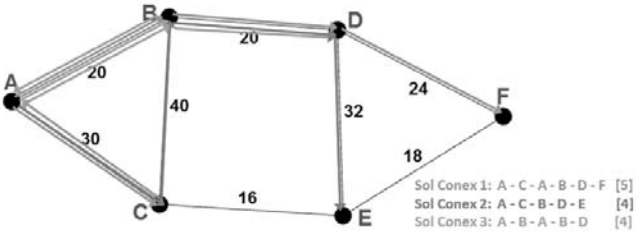


Fig. 5. Mejor solución encontrada para la instancia de 6 nodos con 3 conexiones. Coste en número de saltos junto a cada conexión.

En cuanto a la resolución de la Instancia 19N, se han definido las siguientes conexiones:
- Conexión 1: (H, J, 8, [5,1,2])
- Conexión 2: (B, D, 8, [4,3,1])
- Conexión 3: (Q, B, 1, [2,3,1])
- Conexión 4: (R, J, 3, [5,2,3])
- Conexión 5: (J, S, 8, [4,1,3])

Tras ejecutar el algoritmo nuevamente 10 veces, se han obtenido los resultados mostrados en la Tabla III.

Tabla III
RESULTADOS PARA LA INSTANCIA 19N CON TRES CADENAS DE CONEXIÓN. COSTE DE ROUTING REQUERIDO PARA RESOLVER CADA UNA DE LAS CONEXIONES EN CADA EJECUCIÓN Y COSTE GLOBAL DE LA SOLUCIÓN, ADEMÁS DE LA MEDIA Y DESVIACIÓN TÍPICA DE ÉSTE. EN AZUL Y NEGRITA LA MEJOR SOLUCIÓN OBTENIDA

Ejc	Conex1	Conex2	Conex3	Conex4	Conex5	COSTE
1	**4**	**4**	**2**	**4**	**5**	**19**
2	5	6	2	3	6	22
3	5	6	2	6	7	26
4	4	4	2	3	6	19
5	5	6	2	5	5	23
6	4	6	2	4	6	22
7	5	6	2	5	7	25
8	4	5	2	5	4	20
9	5	5	2	4	5	21
10	4	5	2	5	6	22
					Media	21,9
					Desv. Típica	2,33

Como se puede ver en la tabla y en el grafo (Figura 6), se obtienen soluciones válidas según las restricciones del problema. Además, en este caso, la desviación típica es mucho menor, por lo que podemos considerar que las soluciones son todas útiles. Aunque esto normalmente sería un indicativo de la robustez del algoritmo, nos inclinamos a pensar que las cadenas propuestas son en este caso más fáciles de resolver o que hay menos alternativas para su resolución, pese a ser un grafo de mucho mayor tamaño que el de la instancia anterior.

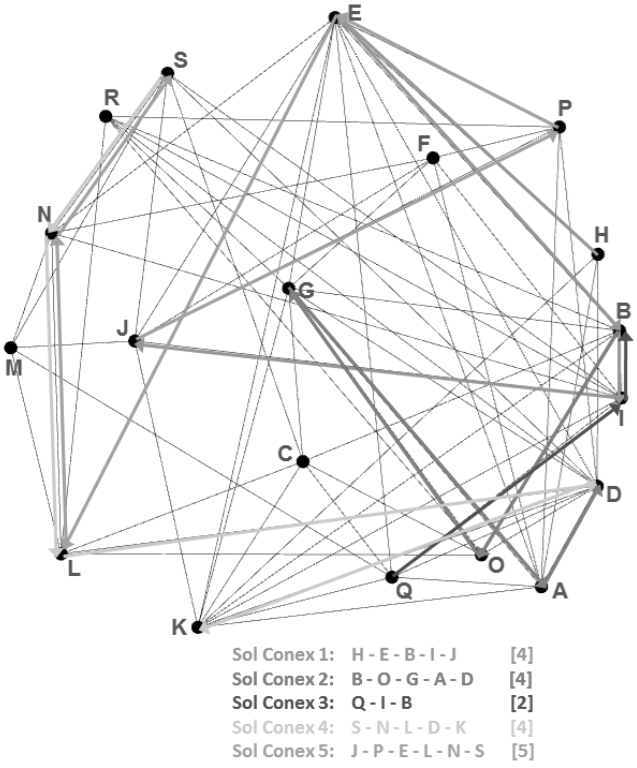


Fig. 6. Mejor solución encontrada para la instancia de 19 nodos con 5 conexiones. Coste en número de saltos junto a cada conexión.

Por último, quisiéramos señalar que los tiempos de ejecución han sido inferiores al medio segundo en todos

los casos, incluso la instancia de tamaño no abordable por un humano. En concreto, y considerando que se ha usado una máquina de nivel usuario, los tiempos medios requridos han sido 0,138 segundos y 0,389 segundos.

VII. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo ha presentado una adaptación de la metaheurística Optimización basada en Colonias de Hormigas (OCH) para la resolución del problema de Routing para SFC (Service Function Chaining), o composición de cadenas de servicios en red.

El algoritmo se ha aplicado sobre dos instancias: una de tamaño pequeño (6 nodos) y una de tamaño medio (19 nodos), que se aproxima a un escenario realista, si bien se ha considerado una topología más compleja en la distribución de las cadenas de servicios de la que se tendría normalmente en una red real.

Los experimentos realizados nos permiten concluir que esta primera aproximación es muy prometedora, puesto que los resultados obtenidos en ambos casos son óptimos o casi óptimos. A su vez, el tiempo requerido para obtener las soluciones ha estado en torno a los 100 ms y a los 300 ms respectivamente, lo cual es un tiempo aceptable en un escenario real. En cualquier caso, una de las ventajas de los algoritmos de OCH es que éstos ofrecen soluciones completas aceptables (aunque no óptimas) desde la primera iteración del mismo, por lo que, si existieran limitaciones de tiempo más estrictas, éstos se adaptarían a ellas fácilmente. Del mismo modo, el algoritmo sería capaz de adaptarse a cambios en la red o los nodos, encontrando caminos alternativos si se agotasen los recursos de un nodo o quedase incomunicado, por ejemplo.

Además, esta propuesta resulta una contribución interesante para el estado del arte, ya que hasta la fecha no se habían aplicado este tipo de metaheurísticas para resolver este problema en concreto.

A partir de estos resultados, continuaremos mejorando el algoritmo en varios aspectos. En primer lugar implementaremos una heurística más efectiva para guiar realmente la búsqueda hacia nodos que puedan servir la siguiente función en la cadena deseada (conexión). A su vez aplicaremos algún mecanismo simple de Búsqueda Local, que permita mejorar las soluciones fácilmente y ahorrar iteraciones para llegar al óptimo, mejorando el rendimiento global del algoritmo. Con estos dos mecanismos nos aseguraremos además de que los resultados son más robustos entre ejecuciones.

Otra posible línea de acción pasa por la implementación de otro modelo de OCH que pudiera ser más efectivo que el elegido, o incluso de otra metaheurística, como un Algoritmo Evolutivo.

Una vez implementados estos nuevos algoritmos, tendríamos una mayor base comparativa de métodos para aplicarlos sobre las mismas instancias y otras de tamaño mayor y con un mayor número de conexiones, intentado acercar los modelos a escenarios más realistas que los abordados en este estudio. En esta última línea estudiaremos también la posibilidad de manejar restricciones de tiempo mucho más exigentes.

AGRADECIMIENTOS

Este trabajo ha sido financiado en parte por los proyectos TIN2017-85727-C4-2-P, RTI2018-102002-A-I00, (Ministerio de Ciencia, Innovación y Universidades), TEC2015-68752 (Ministerio de Economía y Competitividad y fondos FEDER), así como IB18003 (FEDER y Consejería de Economía e Infraestructuras de la Junta de Extremadura) y B-TIC-402-UGR18 (FEDER y Junta de Andalucía).

REFERENCIAS

[1] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[2] V. Eramo, E. Miucci, M. Ammar, and F. G. Lavacca, "An approach for service function chain routing and virtual function network instance migration in network function virtualization architectures," *IEEE/ACM Trans. Networking*, vol. 25, no. 4, pp. 2008–2025, 2017.

[3] T. Lukovszki, M. Rost, and S. Schmid, "It's a match!: Near-optimal and incremental middlebox deployment," *SIGCOMM Comput. Commun. Rev.*, vol. 46, pp. 30–36, Jan. 2016.

[4] M. Dorigo and G. D. Caro, "The ant colony optimization meta-heuristic," in *New Ideas in Optimization* (D. Corne, M. Dorigo, and F. Glover, eds.), pp. 11–32, McGraw-Hill, 1999.

[5] M. Dorigo and T. Stützle, "The ant colony optimization meta-heuristic: Algorithms, applications, and advances," in *Handbook of Metaheuristics* (G. K. F. Glover, ed.), pp. 251–285, Kluwer, 2002.

[6] D. Sudholt and C. Thyssen, "Running time analysis of ant colony optimization for shortest path problems," *Journal of Discrete Algorithms*, vol. 10, pp. 165 – 180, 2012.

[7] A. Mora, J. Merelo, P. Castillo, and M. Arenas, "hchac: A family of moaco algorithms for the resolution of the bi-criteria military unit pathfinding problem," *Computers and Operations Research*, vol. 40, no. 6, pp. 1524 – 1551, 2013.

[8] Chandana M. and S. Thakur, "Ant-net: An adaptive routing algorithm," in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, pp. 1–4, 2016.

[9] A. Laghrissi and T. Taleb, "A Survey on the Placement of Virtual Resources and Virtual Network Functions," *IEEE Communications Surveys and Tutorials*, pp. 1–1, 2018.

[10] J. Gil Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," *IEEE Transactions on Network and Service Management*, vol. 13, pp. 518–532, sep 2016.

[11] B. Yi, X. Wang, K. Li, S. k. Das, and M. Huang, "A comprehensive survey of Network Function Virtualization," *Computer Networks*, vol. 133, pp. 212–262, mar 2018.

[12] N. Huin, B. Jaumard, and F. Giroire, "Optimal Network Service Chain Provisioning," *IEEE/ACM Transactions on Networking*, vol. 26, pp. 1320–1333, jun 2018.

[13] Z. Allybokus, N. Perrot, J. Leguay, L. Maggi, and E. Gourdin, "Virtual function placement for service chaining with partial orders and anti-affinity rules," *Networks*, vol. 71, no. 2, pp. 97–106, 2018.

[14] L. Qu, M. Khabbaz, and C. Assi, "Reliability-Aware Service Chaining in Carrier-Grade Softwarized Networks," *IEEE Journal Selec. Areas in Communications*, vol. 36, no. 3, pp. 558–573, 2018.

[15] T.-M. Nguyen, M. Minoux, and S. Fdida, "Optimizing resource utilization in NFV dynamic systems: New exact and heuristic approaches," *Computer Networks*, vol. 148, pp. 129–141, jan 2019.

[16] I. Jo and G.-I. Kwon, "Genetic algorithm for service function chaining in NFV," in *Mechanical Engineering*, pp. 223–228, 2016.

[17] J.L.Deneubourg, J.M.Pasteels, and J.C.Verhaeghe, "Probabilistic behaviour in ants: a strategy of errors?," *J. Theor. Biol.*, vol. 105, pp. 259–271, 1983.

[18] M. Dorigo, V. Maniezzo, and A. Colorni, "The ant system: Optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 26, no. 1, pp. 29–41, 1996.

[19] H. Hoos and T. Stützle, *Stochastic Local Search. Foundations and Applications*. The Morgan Kaufmann Series in Artificial Intelligence, Elsevier", 1st ed., September 2004.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Mecanismos de ahorro de energía para WiFi en ns-3

Vicente Mayor, Antonio Estepa, Rafael Estepa
Departamento de Ingeniería Telemática,
Universidad de Sevilla
C/ Camino de los Descubrimientos s/n, 41092, Sevilla, España
vmayor@trajano.us.es, aestepa@trajano.us.es, rafa@trajano.us.es

***Resumen*—Las actividades de investigación relacionadas con los mecanismos de ahorro de energía definidos para redes IEEE 802.11 necesitan validar sus propuestas mediante simulaciones o bien mediante experimentación. Sin embargo, el simulador ns-3, ampliamente extendido en la comunidad investigadora, no cuenta con implementaciones documentadas de estos mecanismos de ahorro energético. Este artículo propone la implementación en curso de PSM (Power Saving Mode) que estamos realizando para el simulador ns-3. Esta implementación y la documentación asociada facilita a los investigadores la validación de resultados mediante simulación en ns-3 con un coste menor que la experimentación con terminales reales.**

***Palabras Clave*—ns-3, wifi, psm, energía, simulación**

I. INTRODUCCIÓN

Durante las últimas décadas, el uso y despliegue de las redes de área local inalámbricas (WLAN) IEEE 802.11 [1] ha ganado gran popularidad y, en la actualidad, la casi totalidad de los dispositivos móviles (e.g. smartphones, IoT) disponen de interfaces WiFi que permiten a los usuarios disfrutar de aplicaciones (e.g. mensajería, multimedia, juegos) conectadas a Internet. El gasto energético de la tarjeta WiFi, es un componente importante en dispositivos portátiles impulsados por baterías [2]. Sin embargo, el mecanismo de acceso al medio (MAC) definido en la versión original del standard resulta ineficiente en cuanto al consumo energético [3].

El consumo energético de una interfaz WiFi se puede calcular a partir de la potencia (en Watios) que requiere en cada uno de sus estados (e.g. transmisión, recepción, reposo, etc.) y el tiempo pasado en cada estado [4], [5], [6]. Cuanto menos tiempo se invierta en los estados más exigentes (p.ej. transmisión o recepción), menor será el consumo energético medio de la tarjeta.

A lo largo de sucesivas revisiones, el estándar IEEE 802.11 ha ido incorporando nuevos mecanismos de ahorro energético como PSM (power saving mode) o APSD (Unscheduled Automatic Power Save Delivery), que incluyen un nuevo estado de sueño o *sleep* que puede producir un ahorro de energía muy significativo, aunque a su vez también puede afectar al rendimiento de la red o aplicación pues implica un incremento del retardo [7], [8], [9], [10].

Sin embargo, a pesar de que estos mecanismos se encuentran muy extendidos en los terminales y puntos de acceso actuales, su incorporación en los simuladores más utilizados como ns-3 aun es pobre [11], [12]. Las implementaciones de PSM existentes [13], [14] hacen referencia a versiones antiguas del simulador (ns-2) o no se exponen en suficiente detalle como para poder ser replicadas. Esta ausencia de implementaciones abiertas dificultas las labores de investigación y conduce en muchos casos a la experimentación con terminales reales como en [6], lo cual puede llegar a ser complejo en algunos escenarios.

En este documento se detallan las modificaciones necesarias en el módulo WiFi del simulador ns-3 para incorporar el mecanismo de ahorro de energía PSM. Para ello, se introducen en primer lugar el mecanismo de acceso al medio que propone el estándar IEEE 802.11, seguido de los protocolos de ahorro de energía a implementar, se presenta la estructura ya existente del módulo WiFi en ns-3 y, finalmente se propone un análisis de las consideraciones y cambios mínimos necesarios para la implementación de ambos protocolos. Este es un trabajo en curso cuyo objetivo final es ofrecer una implementación pública y abierta de los modos de ahorro energético de WiFi que pueda ser utilizada en investigaciones futuras.

II. IEEE 802.11 MAC Y AHORRO DE ENERGÍA

Los mecanismos de acceso al medio (MAC) de IEEE 802.11 [1] presentan algunas variantes [15], pero en este estudio nos centraremos en DCF (Distributed Coordination Function) ya que es el mecanismo más común sobre el que se utilizará PSM (Power Saving Mode). Adicionalmente nos centraremos únicamente en el redes

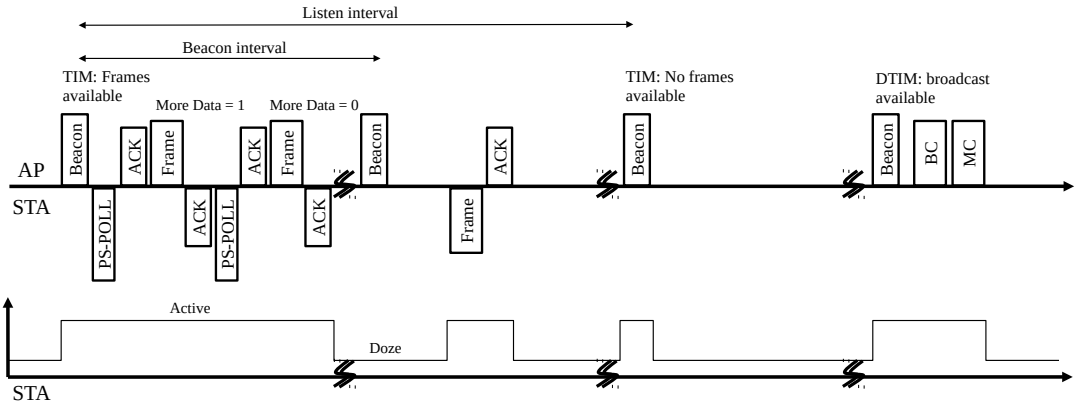


Fig. 1. Ejemplo de PSM

de tipo infraestructura, en el que existe un punto de acceso que actúa de mediador entre las estaciones.

DCF propone un control de acceso al medio (CSMA/CA) que consiste en que cada estación debe escuchar el medio durante un periodo de tiempo para asegurarse de que esté libre antes de transmitir. Si el canal está ocupado, la estación debe esperar un tiempo aleatorio (backoff) antes de volver a intentarlo. Los tiempos de espera (backoff) crecen exponencialmente hasta un número máximo. Las estaciones tienen además un límite de reintentos. El rendimiento de DCF ha sido estudiado analíticamente con anterioridad en [16].

Según el estándar, las estaciones pueden operar tanto en modo activo, como en modo de ahorro de energía (PSM) [3]. En el primero, las estaciones deben mantenerse despiertas para poder transmitir y recibir paquetes. En el último, las estaciones que no estén transmitiendo o recibiendo tramas pueden pasar a un estado de muy bajo consumo energético (sueño).

En el modo de ahorro de energía, las estaciones pueden notificar al punto de acceso cuando van a pasar a un estado de sueño a través del campo PWR MGT de la cabecera. En este caso, el punto de acceso (AP) debe retener y almacenar todo el tráfico que interese a dicha estación ya sea unicast, multicast o difusión.

Periódicamente, los puntos de acceso envían tramas a difusión (beacons) con un nuevo campo llamado TIM (Traffic Indicator Map) que incluye información sobre el tráfico almacenado para las estaciones PSM. Cada cierto tiempo (DTIM period) se enviará una beacon con información del tráfico de difusión disponible (DTIM o Delivery Traffic Indication Message) para las estaciones, seguido de las tramas a difusión almacenadas.

Las estaciones deberán programar intervalos de escucha (listen interval) para la recepción de beacons (no necesariamente todas), siendo estrictamente necesario la escucha de aquellas que contengan DTIM. Si en el campo TIM se indica tráfico disponible, la estación afectada deberá mantenerse despierta para recibir el tráfico si es de difusión, o para solicitar la recepción de un paquete mediante el envío de una trama PS-POLL siguiendo el mecanismo de acceso al medio.

Al recibir el PS-POLL, el punto de acceso deberá asentirlo y solicitar el acceso al medio mediante DCF para enviar un paquete de datos indicando (en el campo More Data) si hay más tramas disponibles para la estación. El proceso se repetirá hasta recibir todas las tramas disponibles y luego la estación podrá volver a dormir.

El procedimiento anterior queda reflejado en la Figura 1. La primera beacon anuncia tráfico disponible para la estación, por lo que la STA envía tramas PS-POLL hasta recibir todo el tráfico disponible. A pesar de que la segunda beacon no es recibida por la estación por su intervalo de escucha, esta se despierta sin restricción para enviar tráfico ascendente. La siguiente beacon es recibida de nuevo por la estación, pero no hay tramas almacenadas. La última beacon anuncia tráfico a difusión, por lo que las estaciones la escuchan y esperan a recibir las tramas.

III. MÓDULO WIFI EN NS-3

En ns-3[1] [11], la representación de los diferentes dispositivos dentro de la red es posible gracias a la clase `Node`. Cada `Node` puede disponer diferentes interfaces de red definidas a partir de la clase `NetDevice`. Por ejemplo, en el caso de WiFi se tendría la clase `WifiNetDevice` y, en el caso de los enlaces punto a punto, se utilizaría `PointToPointNetDevice`.

Un `WifiNetDevice` ofrecerá métodos para que las capas superiores puedan enviar y recibir paquetes. La arquitectura [17] de esta clase es modular (ver Figura 2) y separa sus componentes en tres capas: capa MAC de alto nivel, capa MAC de bajo nivel, y capa PHY.

En la capa MAC de alto nivel se ofrecen diferentes clases que se adaptan a diferentes modos de operación o funcionamiento. Por ejemplo, `ApWifiMac` para los puntos de acceso o `StaWifiMac` para las estaciones. En ellas se tratan las tareas de asociación y generación o recepción de tramas beacon.

En la capa MAC de bajo nivel se distinguen tres componentes principales:

- `MacLow` se encarga del intercambio de paquetes más básico como, por ejemplo, el envío de asentimientos (ACKs). Además gestiona la agregación de paquetes.

[1]En este trabajo se utiliza la versión estable más reciente en la actualidad (ns-3.29). La nomenclatura y funciones pueden variar según la versión.

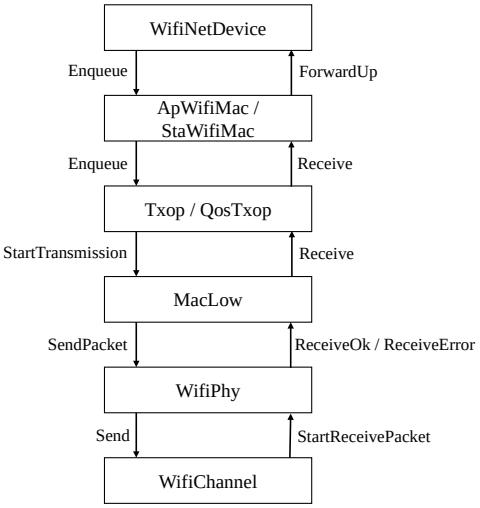


Fig. 2. Arquitectura simplificada de WifiNetDevice

- `ChannelAccessManager` implementa las funciones de acceso al medio (e.g. DCF).
- `Txop` y `QosTxop` gestionan las colas de acceso, fragmentado de paquetes y retransmisiones. En el caso de DCF y envío de tramas sin QoS se utiliza `Txop` mientras que el uso de `QosTxop` queda relegado a EDCA (Enhanced Distributed Channel Access).

La relación entre ellos se resume del siguiente modo: Cuando algún paquete llegue al `Txop` de una estación, este solicita el acceso al canal al `ChannelAccessManager` asociado y esperará hasta obtener permiso. Una vez permitido el acceso al canal, el `Txop` envía el paquete a través del componente `MacLow`.

Finalmente, la capa PHY se encarga de las tareas de envío y recepción de paquetes a través del canal WiFi teniendo en cuenta las posibles pérdidas (e.g. por ruido, interferencias, etc.). Además, incorpora un seguimiento del estado de la NIC (e.g. transmisión, recepción, reposo, sueño) y permite la transición entre dichos estados, lo cual resulta de especial interés para este trabajo [18].

Los componentes anteriores además se apoyan en otros elementos como la definición de cabeceras y métodos para su serialización y deserialización (e.g. `WifiMacHeader`, etc.).

IV. CAMBIOS NECESARIOS

En esta sección se presentan los cambios mínimos necesarios para la inclusión del mecanismo de ahorro de energía PSM en el módulo WiFi de ns-3. Para ello se agruparán en tres apartados: cambios preliminares, comportamiento del punto de acceso y comportamiento de las estaciones.

A. Cambios preliminares

Antes de modificar el comportamiento de los dispositivos, es necesario incorporar al simulador los nuevos tipos de trama y cabeceras que se utilizarán en PSM.

En primer lugar, es necesario implementar las tramas PS-POLL como un nuevo tipo de trama de control en el fichero *wifi-mac-header.**. Además, es necesario modificar el componente `MacLow` (en el fichero *mac-low.**) para que se asientan los PS-POLL (e.g. ACK).

En segundo lugar, es necesario añadir el bit PWR MGT en la cabecera de control. Esto implica la incorporación de métodos que permitan editar e interpretar el valor de este campo y puede hacerse a través del fichero *wifi-mac-header.**.

Finalmente, es necesario crear el elemento TIM para que pueda ser incluido en las beacons, para ello es necesario crear una nueva clase `TimElement` (con ficheros *tim-element.**) y modificar el fichero *mgt-headers.** para que se incluya este campo en las beacons.

B. Comportamiento del punto de acceso

En primer lugar, es necesario seguir un control de las estaciones que se encuentran en el modo de ahorro de energía. Para ello, se ha optado por crear una nueva clase `PwrBlockedDestinations` (ficheros *pwr-blocked-destinations.**) que permita marcar aquellas estaciones cuyo tráfico debe ser retenido por cuestiones de ahorro de energía[2]. Para ello, se modifica `ApWifiMac` (en ficheros *ap-wifi-mac.**) de modo que tras recibir un paquete de datos interprete el bit PWR MGT de la cabecera y actualice el estado de la estación.

Por otro lado, es necesario modificar `ApWifiMac` para que se personalice e incluya el elemento TIM en las beacons transmitidas según el tráfico disponible, así como gestionar el periodo de envío de DTIMs.

Por último, es necesario modificar la clase `Txop` (y `QosTxop`) para que retenga las tramas indicadas en `PwrBlockedDestinations`. Además, es necesario implementar dos nuevos métodos que permitan:

- El envío de un paquete a una dirección concreta tras solicitar el acceso al canal. Además se debe configurar el bit More Data según el caso.
- Y el envío de las tramas a difusión disponibles.

Finalmente, se debe modificar `ApWifiMac` para que haga uso de los métodos anteriores al recibir un PS-POLL, o tras enviar un DTIM con tráfico a difusión disponible, respectivamente.

C. Comportamiento de las estaciones

El comportamiento general de las estaciones debe ser modificado para que incorporen una máquina de estados que regule los eventos de sueño y despertar. Todos los cambios podrán hacerse directamente sobre la clase `StaWifiMac` (y sus ficheros *sta-wifi-mac.**) y se deberán considerar las siguientes transiciones:

- Las estaciones podrán despertarse sin restricción para transmitir.
- Las estaciones deberán escuchar las beacons que le correspondan (según su listen interval y DTIM).
- Las estaciones podrán pasar a un estado de sueño cuando se cumplan todas las condiciones a continuación:

[2]El mecanismo que se propone se inspira en la clase `QosBlockedDestinations`, utilizada para bloquear destinatarios por cuestiones de calidad de servicio.

2.396130	00:00:00_00:00:01	Broadcast	802.11	Beacon frame, SN=120, FN=0, Flags=........, BI=20, SSID=ns3-80211n
2.397697	00:00:00_00:00:02 …	00:00:00_00:00:01 …	802.11	Power-Save poll, Flags=.......T
2.397707		00:00:00_00:00:02 …	802.11	Acknowledgement, Flags=........
2.397803	192.168.1.1	192.168.1.2	UDP	49153 → 9 Len=172
2.398205		00:00:00_00:00:01 …	802.11	Acknowledgement, Flags=........
2.398377	00:00:00_00:00:02 …	00:00:00_00:00:01 …	802.11	Power-Save poll, Flags=.......T
2.398387		00:00:00_00:00:02 …	802.11	Acknowledgement, Flags=........
2.398519	192.168.1.1	192.168.1.2	UDP	49153 → 9 Len=172
2.398921		00:00:00_00:00:01 …	802.11	Acknowledgement, Flags=........
2.400312	192.168.1.2	192.168.1.1	UDP	49153 → 9 Len=172
2.400322		00:00:00_00:00:02 …	802.11	Acknowledgement, Flags=........
2.416610	00:00:00_00:00:01	Broadcast	802.11	Beacon frame, SN=121, FN=0, Flags=........, BI=20, SSID=ns3-80211n

Fig. 3. Captura en Wireshark de la implementación

- Se está asociado a una red WiFi.
- No se está esperando una beacon, además es necesario establecer un temporizador (timeout).
- No tengan ninguna trama disponible para transmitir.
- No tenga constancia de tráfico disponible en el AP, es decir, que la última beacon recibida no indicó tráfico disponible o el último paquete de datos recibido tuviese el bit More Data a cero.
- No se tenga ningún evento de asentimiento en curso.

Las transiciones entre estados podrán hacerse mediante los métodos `SetSleepMode` y `ResumeFromSleep` disponibles en la clase `WifiPhy`.

V. ESTADO ACTUAL Y LÍNEAS FUTURAS

El estado actual de la implementación se encuentra en fases tempranas de depuración. Por ejemplo, se estan verificando el intercambio de tramas y la estructura de las mismas mediante Wireshark, interpretando las trazas (en formato pcap) generadas con ns-3. Los resultados obtenidos son positivos y no muestran evidencias de errores ni en la construcción de las tramas ni en el patrón de funcionamiento. En la Figura 3 se observa un pequeño fragmento del experimento que demuestra que el intercambio de tramas sigue un patrón lógico.

No obstante, la prueba anterior no es suficiente para asegurar la validez de la implementación, ya que sólo nos permite validar el patrón y estructura de tramas y no comprueba que la estación ejecute su ciclo de despertar y sueño. Para solventarlo, nuestra continuación inmediata pasa por validar el consumo energético obtenido frente a mediciones en laboratorio con dispositivos reales.

En el *roadmap* del proyecto[3] se plantean dos líneas:

- En primer lugar, se propone la implementación de U-APSD en ns-3 siguiendo una metodología similar a la utilizada en este trabajo. Este mecanismo de ahorro de energía es muy recomendable para tráfico multimedia en el que el retardo sea determinante como VoIP, por lo que en conjunto ambas implementaciones cubren un gran porcentaje de los mecanismos de ahorro de energía utilizados en la actualidad.
- En segundo lugar, se pretende utilizar la implementación para validar otros trabajos en desarrollo como, por ejemplo, modelos analíticos sobre el rendimiento de ambos mecanismos.

[3]SPECTRA LI1.5: Vehículo como propulsor de la *smartcity* (PI-1479/22/2015).

REFERENCIAS

[1] IEEE 802.11 WG, "IEEE Standard for Local and Metropolitan Networs Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Applications, IEEE 802.11-2007.

[2] Jones, Christine E., et al. "A survey of energy efficient network protocols for wireless networks." wireless networks 7.4 (2001): 343-358.

[3] Tsao, Shiao-Li, and Chung-Huei Huang. "A survey of energy efficient MAC protocols for IEEE 802.11 WLAN." Computer communications 34.1 (2011): 54-67.

[4] Mayor, Vicente, et al. "Deploying a Reliable UAV-Aided Communication Service in Disaster Areas." Wireless Communications and Mobile Computing 2019 (2019).

[5] Ebert, J-P., et al. "Measurement and Simulation of the Energy Consumption of a WLAN Interface." (2002).

[6] Garcia-Saavedra, Andres, et al. "Energy consumption anatomy of 802.11 devices and its implication on modeling and design." Proceedings of the 8th international conference on Emerging networking experiments and technologies. ACM, 2012.

[7] Pérez-Costa, Xavier, Daniel Camps-Mur, and Albert Vidal. "On distributed power saving mechanisms of wireless LANs 802.11 e U-APSD vs 802.11 power save mode." Computer Networks 51.9 (2007): 2326-2344.

[8] Swain, Pravati. "A survey on performance modeling of IEEE 802.11 DCF in Power Save Mode." 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE). IEEE, 2013.

[9] Pérez-Costa, Xavier, and Daniel Camps-Mur. "IEEE 802.11 E QoS and power saving features overview and analysis of combined performance [Accepted from Open Call]." IEEE Wireless Communications 17.4 (2010): 88-96.

[10] Assem, Haytham, et al. "Monitoring VoIP call quality using improved simplified E-model." 2013 International Conference on Computing, Networking and Communications (ICNC). IEEE, 2013.

[11] Nsnam. Network Simulator 3, www.nsnam.org/.

[12] Carneiro, Gustavo. "NS-3: Network simulator 3." UTM Lab Meeting April. Vol. 20. 2010.

[13] Chumchu, Prawit. "An extension to IEEE 802.11 power save mode for NS-3." 2015 Seventh International Conference on Ubiquitous and Future Networks. IEEE, 2015.

[14] Chen, Xijian, Yi Xie, and Chengyan Wang. "Implementation and analysis of IEEE 802.11 PSM in NS-2." 2011 International Conference on Machine Learning and Cybernetics. Vol. 3. IEEE, 2011.

[15] Malik, Aqsa, et al. "QoS in IEEE 802.11-based wireless networks: a contemporary review." Journal of Network and Computer Applications 55 (2015): 24-46.

[16] Bianchi, Giuseppe. "Performance analysis of the IEEE 802.11 distributed coordination function." IEEE Journal on selected areas in communications 18.3 (2000): 535-547.

[17] "Design Documentation." Design Documentation - Model Library, Nsnam, www.nsnam.org/docs/models/html/wifi-design.html.

[18] Wu, He, Sidharth Nabar, and Radha Poovendran. "An energy framework for the network simulator 3 (ns-3)." Proceedings of the 4th international ICST conference on simulation tools and techniques. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Metodología para la gestión de la QoX basada en el aprendizaje automático

Leire Cristobo, Luis Zabala, Eva Ibarrola, Armando Ferro, Fidel Liberal
Departamento de Ingeniería de Comunicaciones
Universidad del País Vasco - UPV/EHU
ESI de Bilbao-Plaza Ingeniero Torres Quevedo, 1 - 48013 BILBAO.
{leire.cristobo, luis.zabala, eva.ibarrola, armando.ferro, fidel liberal}@ehu.eus

***Resumen*- La llegada del 5G va a suponer una auténtica revolución, tanto en la evolución de las redes y servicios como en otros muchos sectores productivos. Esta tecnología traerá consigo nuevas oportunidades de negocio y grandes retos a afrontar. Uno de estos retos será, sin duda, la gestión de la calidad de servicio (QoS). Este concepto ha evolucionado en los últimos tiempos, complicando su gestión. Ya no se habla tan sólo de la calidad del funcionamiento de la red (NP) sino que cobran importancia nuevas dimensiones, como la calidad experimentada (QoE) o la calidad de negocio (QoBiz).**

En este artículo se presenta una metodología para la implantación de un modelo para la gestión global de la QoS en todas sus dimensiones (QoX). La metodología propuesta se basa en la utilización del aprendizaje automático. Mediante el uso de algoritmos de aprendizaje, tanto supervisado como no supervisado, se propone automatizar y dinamizar los procesos que permitan una gestión adecuada de la QoX en escenarios tan complejos como los que se plantean con la tecnología 5G.

***Palabras Clave*- QoX, QoS, QoE, QoBiz, Machine Learning**

I. INTRODUCCIÓN

La era de Internet ha traído consigo un sinfín de nuevas posibilidades cambiando los hábitos de nuestras actividades más cotidianas. Es por ello que los usuarios de Internet son cada vez más exigentes, tanto en términos del acceso a la red como en cuanto a la calidad de los servicios ofrecidos a través de la misma. La llegada de 5G abre nuevas posibilidades y oportunidades, facilitando mejores coberturas, movilidad global, mayores velocidades de transmisión y latencias mínimas. La coexistencia de redes heterogéneas (HetNet) permitirá tener redes de acceso (RAN) totalmente agnósticas respecto a la tecnología. Así mismo, la utilización de plataformas como SDN (Software Defined Networks) y NFV (Network Functions Virtualization) permitirá desplegar, parametrizar y gestionar recursos de forma dinámica, de acuerdo a los requerimientos de los usuarios o las necesidades para garantizar la rentabilidad del proveedor. Sin embargo, en este escenario tan complejo, puede resultar complicado establecer una gestión global y adecuada de la calidad de servicio que permita cumplir con las expectativas de los usuarios y responder a las necesidades de calidad de negocio (QoBiz) que dicte el mercado.

En este contexto, se enmarca la presente propuesta que, tomando como base la minería de datos (Big Data) y las técnicas de aprendizaje automático (Machine Learning - ML), establece una metodología para la implantación de un modelo para la gestión global de la QoX en las redes de telecomunicaciones actuales y venideras.

II. ANTECEDENTES

En la literatura científica más actual se pueden encontrar algunos trabajos que proponen la utilización de las técnicas de aprendizaje automático para la gestión de la QoE [1-5]. El marco presentado por Yusuf-Asaju en [1] contempla un proceso para estimar o predecir la QoE mediante la utilización de ML (Fig.1). En este trabajo se incide en la importancia de determinar los aspectos que deben ser tenidos en cuenta de cara a un modelado adecuado de la QoE.

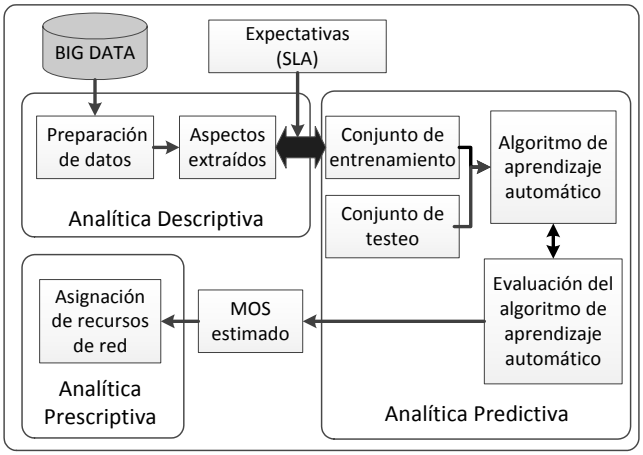


Fig. 1. Marco para modelar la QoE mediante ML y Big Data [1].

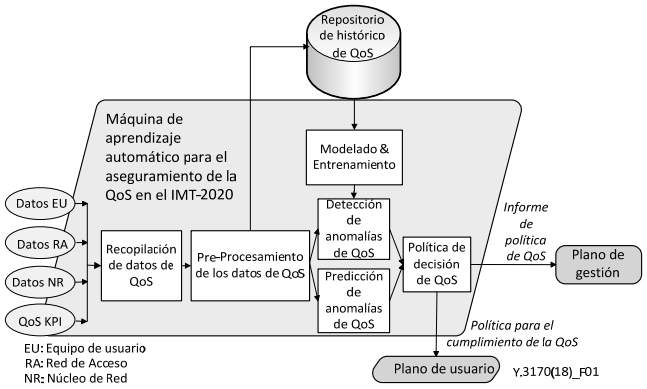


Fig. 2. Marco para el aseguramiento de la QoS en base a ML de la UIT-T

Por ello, se propone la utilización de la minería de datos (Big Data) para la identificación de estos aspectos de influencia. Se especifica cómo algunos de estos aspectos pueden ser extraídos de los datos obtenidos de la red (localización, hora de uso, etc.). También se destaca la importancia de otros aspectos de influencia más subjetivos, como las expectativas y requerimientos de los usuarios, que suelen capturarse por medio de otros métodos más laboriosos, caros y complejos, como son las encuestas. Por todo ello, se determina que la identificación de los factores de influencia en la QoE es crucial en el proceso de estimación y predicción de la misma. Existen otros trabajos [6-8] que proponen técnicas para la determinación de estos factores de influencia (FI) y, en general, parece haber consenso en categorizarlos en los tres grupos que a continuación se detallan:

- *Factores de contexto*: Localización, escenario (comercial, residencial, público...), movilidad, tiempo/días/horas de uso, coste del servicio, etc.
- *Factores humanos*: Edad, género, estudios, empleo, experto/no experto, emociones, expectativas, etc.
- *Factores del sistema*: rendimiento/caudal/velocidad, retardo, pérdidas, seguridad, tipo de dispositivo, monitor/pantalla, etc.

Otro de los aspectos en los que parecen confluir los estudios más recientes es en la idoneidad de utilizar técnicas de aprendizaje automático supervisado/semi-supervisado para el establecimiento de la correlación entre dos de sus dimensiones más importantes: la calidad de la red (Network Performance-NP) y la calidad experimentada por los usuarios (Quality of Experience-QoE). Los trabajos de Alreshoodi y Aroussi [5, 9, 10] describen la necesidad de sistemas con entrenamiento para llegar a modelar estas relaciones. En concreto, el estudio de Aroussi presentado en [10], desarrolla un excelente resumen de las diferentes aproximaciones propuestas por distintos autores

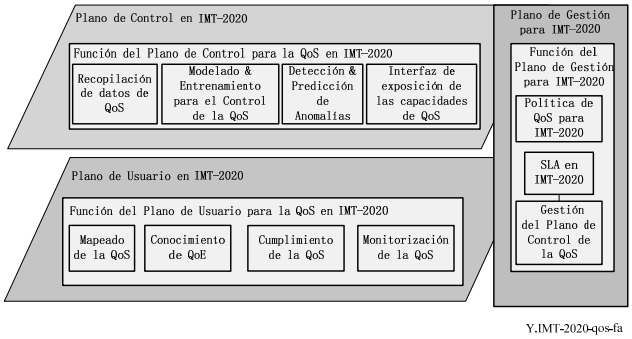


Fig. 3. Marco para el aseguramiento de la QoS en IMT-2020 de la UIT-T

Se citan, entre otros, la regresión de mínimos cuadrados, las redes neuronales artificiales y los métodos de clasificación (máquinas de vectores de soporte, árboles de decisión, NaiveBayes, k-vecinos más cercanos, etc.).

III. MARCO DE REFERENCIA

Si bien los trabajos mencionados en la sección anterior han contribuido enormemente a la especificación de la metodología objeto de este estudio, se ha considerado fundamental tomar como referencia un marco que esté contemplado en los estándares. En este sentido, cabe destacar que también los organismos de estandarización han tomado conciencia de las posibilidades que el ML puede proporcionar para la gestión de la QoX.

En concreto, la UIT-T fundó en enero de 2018 el grupo de trabajo "*Focus Group on Machine Learning for Future Networks including 5G"* que está trabajando activamente en el ámbito del ML para 5G (*IMT-2020 en la UIT-*).

Así mismo, el grupo de trabajo SG-13 de la UIT-T ha editado varias recomendaciones relacionadas con el uso del ML para el aseguramiento de la QoS [. El marco de trabajo descrito en estas recomendaciones (Fig. 2 y 3) es el que se ha establecido como marco de referencia para la definición de la metodología propuesta.

IV. MODELO Y METODOLOGÍA

El modelo para la gestión de la QoX a utilizar, es el modelo QoXfera [14]. En la Fig. 4 se muestra la arquitectura en capas del modelo y los aspectos y dimensiones de la QoS que se contemplan en cada capa. La arquitectura se define de tal forma que permite relacionar las diferentes vertientes de la QoS (QoX) y establecer las dependencias entre ellas. Además, al definirlo se han tenido en cuenta los estándares de referencia para la gestión de la QoS [15-18] y encaja perfectamente con los nuevos marcos para el aseguramiento de la calidad en escenarios 5G contemplados recientemente por la UIT-T [11, 13].

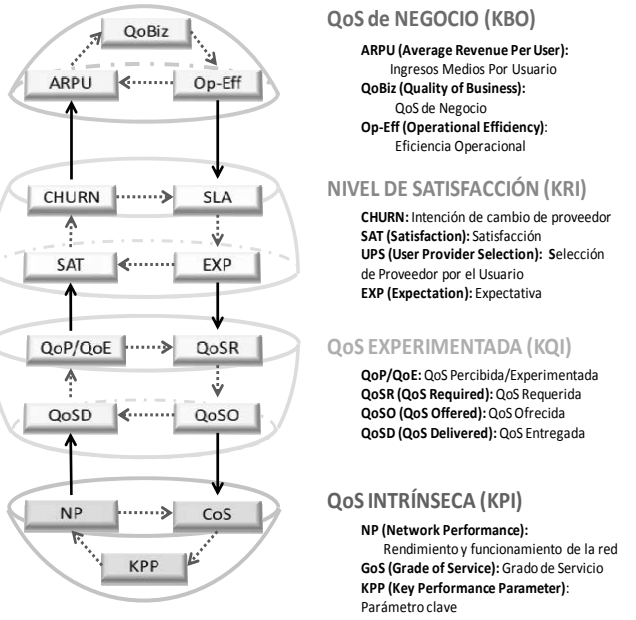


Fig. 4. Modelo QoXfera

El objetivo principal del modelo es el desarrollo de un sistema orientado a la mejora de la satisfacción de los usuarios garantizando el beneficio de los proveedores. De este modo, la arquitectura de QoXfera contempla el análisis y evaluación de la QoS desde todas sus vertientes: la QoS intrínseca, la QoS experimentada (QoE), la satisfacción del usuario y la QoS de negocio (QoBiz). Este análisis se establece en base a un proceso iterativo de traslación y rotación de la arquitectura esférica, que permite una gestión efectiva de la QoX basada en la búsqueda de la mejora continua de la satisfacción final del usuario con el servicio, a través de las sucesivas rotaciones y traslaciones hacia la convergencia de las cuatro vertientes de la QoS. De esta forma se persigue garantizar la fidelidad de cliente y, por ende, la rentabilidad del proveedor.

Sin embargo, la definición de una metodología para poder automatizar las iteraciones mencionadas lleva siendo, desde hace mucho tiempo, el caballo de batalla de esta investigación. Esto se debe a la multitud de dependencias entre los diferentes aspectos contemplados en el modelo y la naturaleza dinámica de muchos de ellos. Es por ello, que se ha encontrado en el aprendizaje automático (ML) una vía de solución a este problema dadas las posibilidades que estos algoritmos pueden proporcionar. Adicionalmente, los nuevos estándares de aseguramiento de la QoS propuestos por la UIT-T para el IMT-2020 han contribuido a confirmar la idoneidad de esta solución. De este modo, se ha definido una metodología (Fig. 5) basada en la utilización de mecanismos de ML. Estos algoritmos ayudarán en la identificación de los aspectos de influencia e indicadores de interés en la gestión de la QoX.

Por otro lado, el aprendizaje automático facilitará dinamizar las relaciones entre las diferentes vertientes de la QoX (NP/QoE/QoBiz). Recopilar y almacenar datos (Big Data) para alimentar los algoritmos de aprendizaje es esencial para inferir las reglas que se aplicarán, tanto para la identificación de los indicadores clave como en la especificación y dinamización de las relaciones entre las distintas capas del modelo.

Por último, destacar que ha sido fundamental en el desarrollo de la metodología la arquitectura marco para la gestión de la QoS que está actualmente definiendo la UIT-T [13] así como los requisitos para la aplicación del ML en la gestión de la QoS recogidos en la Rec. Y.3170 [11] de este organismo, ambos resumidos en las Fig. 2 y 3.

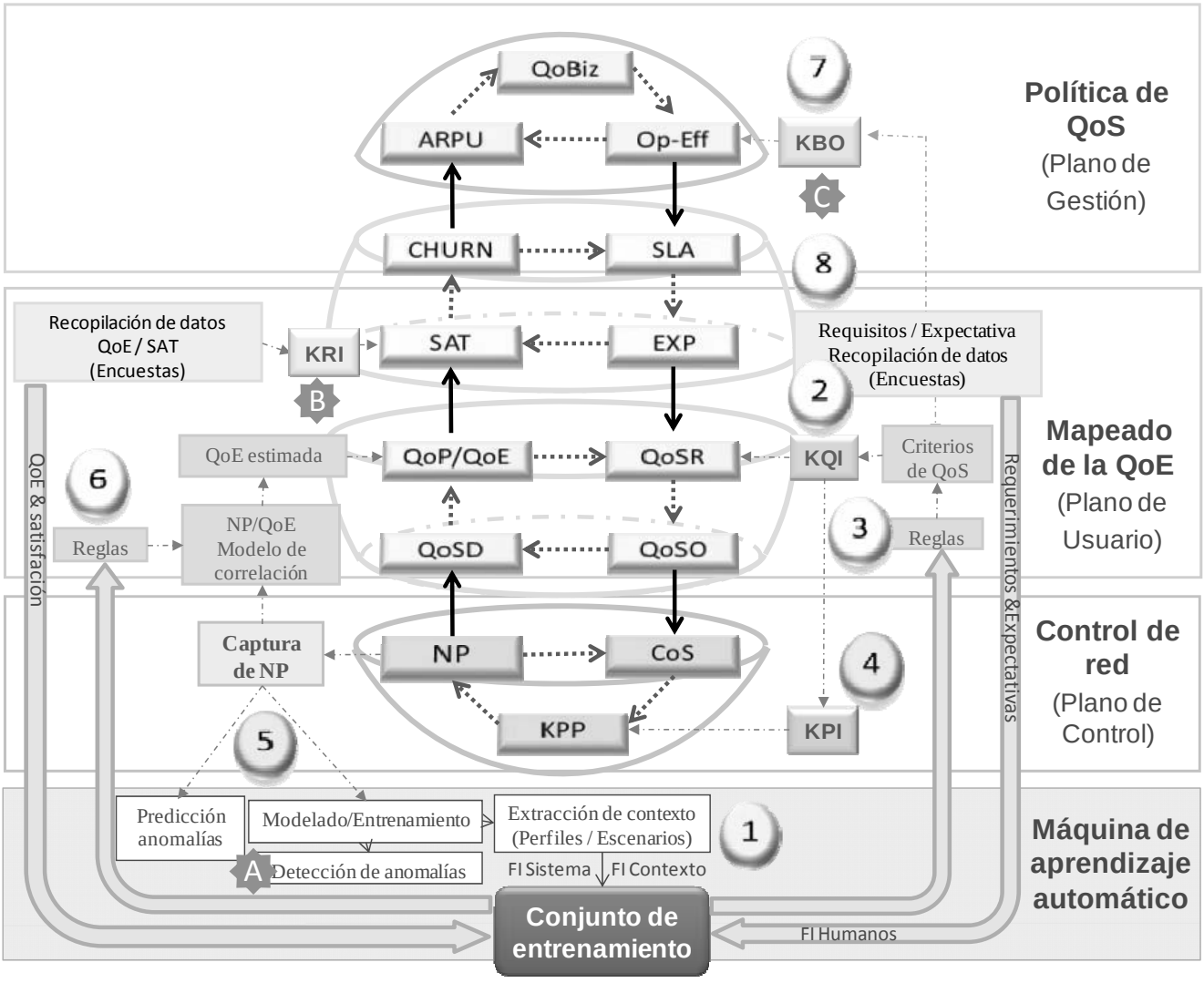


Fig. 5. Metodología para la gestión de la QoX basada en el aprendizaje automático

A. Descripción de la metodología

Como se recoge en la Recomendación UIT-T E.802 [16], para que cualquier modelo de gestión de la QoS sea eficaz, es crucial la identificación de los criterios de QoS relevantes para los usuarios (función de sus requisitos y expectativas).

Este proceso, que puede parecer trivial, puede llegar a ser muy complejo, sobre todo en entornos de redes inalámbricas y heterogéneas como 5G. En estos escenarios la respuesta de la red puede depender enormemente del comportamiento de los usuarios y de muchos otros factores contextuales y no contextuales no controlados por los proveedores. Es por ello que se considera crucial un análisis preliminar para el establecimiento de los factores de influencia asociados a cada escenario concreto y cada tipología de usuarios.

En base a ello, el primer paso propuesto en la metodología es comprender el comportamiento de los usuarios en cada uno de los escenarios posibles. Sólo de esa forma será posible identificar sus requerimientos y expectativas que nos permitan determinar los criterios de QoS relevantes para ellos y los indicadores clave de calidad pertinentes (Key Quality Indicator - KQI).

Dado que este primer paso es fundamental, se plantea recopilar tanto información contextual como no contextual a través de la red (paso #1 en la Fig. 5). Se proponen técnicas de ML no supervisadas (en concreto análisis por agrupamiento/clustering) para inferir los diferentes escenarios y perfiles de usuarios a partir de los datos registrados en la red (extracción de contexto en la Fig. 5). Estos datos nos permitirán obtener los factores de influencia relacionados con el contexto del usuario (como la ubicación, tipo de escenario, día/hora/tiempo de uso...), así como los denominados factores de influencia debido al sistema (tipo de dispositivo, sistema operativo, etc).

Por otro lado, se propone capturar los factores de influencia que no han podido ser extraídos de la red, es decir, los categorizados como humanos (edad, género, experiencia, emociones, requerimientos, expectativas, etc.) a través de encuestas (paso #2 en la Fig. 5). Una vez establecidos todos los factores de influencia y teniendo en cuenta los resultados de la encuesta en cuanto a expectativas y requerimientos, se plantea la utilización del aprendizaje supervisado inductivo, utilizando estos datos como conjunto de entrenamiento, para inferir las reglas que nos permitan identificar los KQI relevantes para cada tipología de usuarios en cada escenario (paso # 3 en la Fig. 5). Esta automatización evitará la necesidad de desarrollar el proceso tan laborioso de las encuestas, excepto para la fase inicial de entrenamiento.

Una vez determinados los KQI, se identificarán los indicadores clave de funcionamiento asociados (Key Performance Indicator - KPI) y sus parámetros clave de rendimiento asociados (Key Performance Parameter - KPP) de acuerdo con la normativa establecida [19, 20].

Se definirán los sistemas de medición adecuados para llevar a cabo las mediciones asociadas al funcionamiento de la red de acuerdo con los KPP/KPI establecidos (paso #4 en la Fig. 5). Comenzará de este modo la función del *plano de control* con la recopilación de datos de NP, tal y como se establece en el marco de referencia, y que en nuestro modelo QoXfera queda reflejado en la capa “QoS intrínseca” (paso #5 en la Fig. 5).

En base a los datos de QoS asociados al NP, se recomienda usar técnicas de ML no supervisadas para detectar y predecir anomalías en la red, tal y como establece también el marco de referencia. El resultado de este análisis constituye el primer punto de intervención donde se pueden implementar acciones correctivas para mejorar la QoS (estrella roja A en la Fig. 5).

En el *plano de usuario* se encuentran ubicadas las dos capas intermedias de nuestra QoXfera (mapeado de la QoE en la Fig. 5). El método que se propone para la gestión de la QoX en este plano sigue también las pautas marcadas en el marco de referencia. Se considera crucial el mapeado automatizado de la NP con la QoE. Para la automatización de esta correlación se sugiere la utilización de algoritmos supervisados de aprendizaje automático (modelos de regresión, etc.) [5]. En este caso es necesario, una vez más, utilizar tanto los datos objetivos (de la calidad de la red) como los datos subjetivos, relacionados con la experiencia y la satisfacción del usuario, recopilados por medio de encuestas (paso # 6 y modelo de correlación NP / QoE en la Fig. 5). Los factores de influencia contextuales y no contextuales también serán cruciales cuando se analice la QoE y la satisfacción del usuario y, por esta razón, nuevamente, se incorporan al conjunto de entrenamiento para aprender las reglas que proporcionarán la QoE prevista.

El aprendizaje automatizado evitará repetir el proceso de encuesta para capturar la QoE en cada momento excepto en el caso de la encuesta inicial que permitirá entrenar el sistema. Este conjunto de entrenamiento se utilizará para deducir las reglas que controlarán el modelo de correlación NP/QoE.

Sobre la base de los resultados de la QoE, el modelo de satisfacción (CSAT) [21] estimará la satisfacción del usuario con el servicio. Esto constituye el segundo punto de actuación (estrella roja B en la Fig. 5) en el que pueden ser necesarias acciones correctivas basadas en los indicadores de riesgo (Key Risk Indicator – KRI) detectados que pueden provocar la pérdida de clientes (churn) y afectar al modelo de negocio.

Finalmente, la decisión de la política de QoS se toma en el *plano de gestión*, como establece el marco de referencia. Sobre la base de los resultados de la detección/predicción de anomalías en la red y los resultados en la QoE se actualizará (paso # 7 en la Fig. 5) el indicador clave de negocio (Key Business Objectives – KBO). Los KBO se derivan de las áreas comerciales que se consideran importantes para cada compañía y deben ajustarse a través de la eficiencia operativa para aumentar los ingresos, reducir los costos y mejorar la experiencia del cliente. Esto constituye el último punto de intervención donde se pueden requerir acciones correctivas (estrella roja C en la Fig. 5).

La facturación, la publicidad, el ajuste de los requisitos de QoS y otras medidas adicionales deben analizarse para actualizar el SLA en base a los resultados obtenidos del plano de usuario y el plano de control (paso #8 en la Fig. 5).

V. Conclusiones

En este artículo se presenta una metodología para la implementación de un modelo para la gestión global de la QoX (QoXfera) en base a la utilización de técnicas de minería de datos (Big Data) y aprendizaje automático (Machine Learning).

Teniendo en cuenta los estándares internacionales, la metodología propuesta hace uso de técnicas de ML supervisadas para identificar los indicadores de calidad y el establecimiento de relaciones entre las diferentes dimensiones de la QoX. Se propone, además, la utilización de mecanismos de ML no supervisados para la inferencia de los factores de influencia en la QoE y la detección y predicción de anomalías en la red.

El objetivo final que se persigue es la identificación de los puntos críticos donde se hacen necesarias acciones correctivas para conseguir la satisfacción del usuario y garantizar la rentabilidad de los proveedores.

Si bien la definición de la metodología y la validación de la misma se encuentran en una fase inicial, se han llevado a cabo algunos experimentos en escenarios reales que evidencian la validez de la propuesta.

Referencias

[1] A. W. Yusuf-Asaju, Z. M. Dahalin, and A. Ta'a, "Framework for modelling mobile network quality of experience through big data analytics approach," *Journal of Information and Communication Technology (JICT)*, vol. 17, pp. 79-113, 2018.

[2] S. Ayoubi, N. Limam, M. A. Salahuddin, N. Shahriar, R. Boutaba, F. Estrada-Solano, and O. M. Caicedo, "Machine Learning for Cognitive Network Management," *IEEE Communications Magazine*, vol. 56, pp. 158-165, 2018.

[3] L. Amour, M. I. Boulabiar, S. Souihi, and A. Mellouk, "An improved QoE estimation method based on QoS and affective computing," in *2018 International Symposium on Programming and Systems (ISPS)*, 2018, pp. 1-6.

[4] A. W. Yusuf-Asaju, Z. B. Dahalin, and A. Ta'a, "Mobile network quality of experience using big data analytics approach," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 658-664.

[5] M. Alreshoodi and J. Woods, "Survey on QoE\QoS correlation models for multimedia services," *arXiv preprint arXiv:1306.0221*, 2017.

[6] S. Baraković, J. Baraković, and H. Bajrić, "QoE Dimensions and QoE Measurement of NGN Services," in *18th Telecommunication Forum TELFOR*, Serbia Belgrade., 2010.

[7] S. Baraković and L. Skorin-Kapov, "Survey and Challenges of QoE Management Issues in Wireless Networks," *Journal of Computer Networks and Communications*, vol. 2013, p. 28, 2013.

[8] R. Stankiewicz and A. Jajszczyk, "A survey of QoE assurance in converged networks," *Computer Networks*, vol. 55, pp. 1459-1473, 2011/05/16/ 2011.

[9] S. Aroussi, T. Bouabana-Tebibel, and A. Mellouk, "Empirical QoE/QoS correlation model based on multiple parameters for VoD flows," in *2012 IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 1963-1968.

[10] S. Aroussi and A. Mellouk, "Survey on machine learning-based QoE-QoS correlation models," in *International Conference on Computing, Management and Telecommunications (ComManTel)*, 2014, pp. 200-204.

[11] ITU-T, "Y.3170 : Requirements for machine learning-based quality of service assurance for the IMT-2020 network," ed, 2018.

[12] ITU-T, "Draft new Recommendation Y.IMT-2020.qos-mon: IMT-2020 network QoS monitoring architectural framework," ed, 2018.

[13] ITU-T, "Draft new Recommendation Y.IMT2020-qos-fa: QoS framework architecture for IMT-2020 networks," ed, 2019.

[14] E. Ibarrola, E. Saiz, L. Zabala, L. Cristobo, and J. Xiao, "A new global quality of service model: QoXphere," *IEEE Communications Magazine*, vol. 52, pp. 193-199, 2014.

[15] ITU-T, "G.1000: Communications quality of service: A framework and definitions," ed, 2001.

[16] ITU-T, "E.802: Framework and methodologies for the determination and application of QoS parameters," ed, 2007.

[17] ITU-T, "E.800: Definitions of terms related to Quality of Service," ed, 2008.

[18] ITU-T, "E.800 SerSup10 : ITU-T E.800 series - QoS/QoE framework for the transition from network oriented to service oriented operations," ed, 2016.

[19] ETSI, "EG 202 009-1: Quality of telecom services; Part1: Methodology for identification of parameters relevant to the Users," ed, 2007.

[20] ETSI, "EG 202 009-2: Quality of telecom services; Part 2: User related parameters on a service specific basis," ed, 2007.

[21] J. Xiao and R. Boutaba, "Assessing Network Service Profitability: Modeling From Market Science Perspective," *Networking, IEEE/ACM Transactions on*, vol. 15, pp. 1307-1320, 2007.

Glosario de acrónimos

ARPU:	Ingresos medios por usuario
CHURN:	Tasa de cancelación de clientes
CoS:	Clase de servicio
EXP:	Expectativas del usuario
FI:	Factores de influencia
HetNet:	Redes heterogéneas
IMT:	Telecomunicaciones Móviles Internacionales
KBO:	Objetivo clave de negocio
KQI:	Indicador clave de calidad
KPI:	Indicador clave de funcionamiento
KPP:	Parámetro clave de funcionamiento
KRI:	Indicador clave de riesgo
ML:	Aprendizaje automatizado (Machine Learning)
NFV:	Virtualización de las funciones de red
MOS:	Nota media de opinión (Mean Opinion Score)
NP:	Funcionamiento de la red
Op-Eff:	Eficiencia operacional
QoBiz:	Calidad de negocio
QoE:	Calidad experimentada
QoP:	Calidad percibida
QoS:	Calidad de servicio
QoSO:	Calidad de servicio ofrecida (QoS Offered)
QoSR:	Calidad de servicio requerida (QoS Required)
QoSD:	Calidad de servicio entregada (QoS Delivered)
QoX:	Calidad global (en todas sus dimensiones)
RAN:	Redes de acceso inalámbricas
SAT:	Satisfacción del usuario
SDN:	Redes definidas por software
SLA:	Acuerdo de nivel de servicio

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Evaluación de experiencias de innovación docente en el Grado de Ingeniería Telemática de la Universitat de València

Antonio Soriano Asensi, Jaume Segura Garcia, Carmen Botella, Santiago Felici-Castell,
Joaquín Pérez y Miguel García Pineda
Departament d'Informàtica, Escola Tècnica Superior d'Enginyeria
Universitat de València
Avinguda de l'Universitat, s/n, 46100 Burjassot, Valencia.
{antonio.soriano-asensi,jaume.segura,carmen.botella,santiago.felici,joaquin.perez-soler,miguel.garcia-pineda}@uv.es

***Resumen*—Desde el punto de vista de la investigación estamos acostumbrados a diseñar experimentos, realizar medidas, analizar los resultados y difundirlos al resto de la comunidad científica. Pero en el ámbito docente no es tan frecuente completar todo ese ciclo hasta la difusión de los resultados. El objetivo de esta presentación es dedicar unos minutos a la reflexión, al intercambio de experiencias docentes y a la evaluación de su impacto en el aprendizaje de la Ingeniería Telemática. En el presente trabajo se pretende presentar el planteamiento seguido y los resultados obtenidos, para evaluar el beneficio de introducir nuevas prácticas en el laboratorio de Fundamentos de Sistemas de Comunicaciones del Grado en Ingeniería Telemática de la Universitat de València.**

***Palabras Clave*—Docencia en telemática, Evaluación de experiencias docentes, Radio definida por *software*, GNU Radio**

I. INTRODUCCIÓN

En este trabajo, se plantea una reflexión en torno al problema de introducir y evaluar de forma cuantitativa y cualitativa los resultados de un proyecto de innovación docente centrado en la asignatura de Fundamentos de Sistemas de Comunicaciones (FST), del Grado en Ingeniería Telemática (GIT) de la Escuela Tècnica Superior de Ingeniería de la Universitat de València (ETSE-UV). El proyecto de innovación docente de donde surge el presente trabajo, se inició en el curso académico 2015-2016 y tiene como propósito la introducción gradual de varias plataformas de Radio Definida por *Software* (SDR) en asignaturas de GIT y el Máster en Ingeniería de Telecomunicación (MITUV). En esta contribución, se explora el uso del dispositivo HackRF One de Great Scott Gadgets[1], junto con una herramienta de software de código abierto como GNU Radio[2]. HackRF One es un periférico SDR con un rango de frecuencia de 1 MHz a 6 GHz, y puede utilizarse como periférico USB o en modo de operación autónomo. Las referencias [1], [2], [3] analizan los beneficios, retos y casos de uso de las plataformas SDR en la enseñanza de las telecomunicaciones.

Debido a la especificidad y al elevado coste de la instrumentación, es frecuente que se aborde la enseñanza de las asignaturas de telecomunicaciones desde un punto de vista teórico. Estas asignaturas incluyen sesiones de laboratorio donde el alumnado simula partes de los sistemas de comunicaciones utilizando, por ejemplo, Matlab, Python, o C. Esta aproximación presenta dos puntos críticos desde nuestro punto de vista. En primer lugar, obliga a simplificar o idealizar el funcionamiento de los sistemas a evaluar por limitaciones de tiempo de laboratorio, fundamentalmente. En segundo lugar, en el entorno académico actual, es necesario acreditar la calidad de la titulación mediante las certificaciones otorgadas, por ejemplo, por el Consejo de Acreditación de Ingeniería y Tecnología (ABET) o la Red Europea para la Acreditación de la Enseñanza de la Ingeniería (ENAEE). Estos organismos, evalúan los Objetivos Educativos del Programa y los Resultados Estudiantiles, comprendiendo tanto los conocimientos, las habilidades como los comportamientos adquiridos en la titulación. Más aún, se intenta cuantificar la dimensión de "*lo que se espera que los estudiantes sean capaces de hacer*". De esta forma, aparece la necesidad de acercar la aproximación teórica al entorno profesional, donde un conocimiento más realista de los sistemas de comunicaciones es mucho mejor valorado.

La experiencia aprendida y aquí presentada, tiene como objetivos por un lado introducir una aproximación más aplicada en las asignaturas de comunicaciones. Por otro lado, se pretende la elaboración de material docente mejor adaptado al interés de los estudiantes con el objetivo

[1]https://greatscottgadgets.com/hackrf/
[2]https://www.gnuradio.org/

de mejorar su implicación en las actividades formativas. Para ello, se propone una metodología de Aprendizaje Basado en Proyectos (ABP) para mejorar adquisición de habilidades prácticas relacionadas con las Comunicaciones Digitales y el Procesado de Señales para cursos de Comunicaciones a través de dispositivos SDR de bajo coste, lo que ha demostrado mejorar la participación de estudiantes de máster y grado.

Las estructura del artículo es como sigue. En primer lugar, en la sección II se presenta una breve revisión de algunas plataformas SDR disponibles y de algunos de los proyectos llevados a cabo con las mismas. En la sección III se detalla la experiencia ABP llevada a cabo con los estudiantes de máster, cómo prepararla, qué actividades se han llevado a cabo y cómo ir cediendo el protagonismo a los estudiantes para que completen por sí mismos el trabajo. A partir de la experiencia ABP con los estudiantes de máster, en la sección IV se indica cómo se ha elaborado un material guiado para que los estudiantes de grado lleven a cabo una práctica demostración. En la sección V se detalla la evaluación del impacto de la actividad propuesta en la motivación de los estudiantes de grado en su formación docente. Finalmente en la sección VI se resumen los principales resultados de esta contribución.

II. PLATAFORMAS SDR

Existe una amplia gama de equipos SDR y programas de software para soportar y gestionarlos. Por ejemplo, en [1] y [2] utilizan Universal Software Radio Peripherals (USRP) [4] y el software Laboratory Virtual Instrumentation Engineering Workbench (LabVIEW), cuya interfaz gráfica facilita su uso. Pero los USRP son dispositivos bastante costosos y se está trabajando en encontrar soluciones de bajo coste. En [3] se desarrolló un curso abierto con dispositivos RTL-SDR, por menos de $20, donde se usa MATLAB y Simulink para modelar los sistemas de comunicación digital. Otro producto en el mercado es [5], con un coste unitario de alrededor de $200.

En el GIT, cada asignatura consta normalmente de 4 grupos de laboratorio con 16-20 estudiantes cada uno, lo que representa una media de 40 dispositivos SDR necesarios por curso, teniendo en cuenta que, a veces, dos cursos pueden impartirse en paralelo. En nuestro caso, los dispositivos RTL-SDR inicialmente se han utilizado junto con GNU Radio en cursos introductorios de Comunicaciones Digitales para aumentar la participación de los estudiantes, con un presupuesto asequible [6].

Sin embargo, estos dispositivos (RTL-SDR) plantean algunos problemas de fiabilidad, especialmente cuando se conectan durante períodos prolongados, debido a problemas de disipación de energía. Es por ello que se han adquirido dispositivos HackRF One, con un coste unitario de alrededor de $300, para complementarlos. Que son los mismos dispositivos que se emplean en otros cursos más avanzados a nivel de máster.

III. APRENDIZAJE DE COMUNICACIONES DIGITALES BASADO EN PROYECTOS

El ABP es un estilo de aprendizaje activo que coloca al estudiante en el centro del proceso de aprendizaje. La exploración activa de los retos del mundo real permite a los estudiantes adquirir un conocimiento más profundo. El proceso de aprendizaje en ABP es dirigido por los estudiantes, mientras que el papel del profesor es guiar la investigación de los estudiantes proponiendo preguntas y tareas que les ayuden a resolver el problema propuesto. Se espera que el profesor presente el problema y proporcione a los estudiantes los materiales y la documentación necesarios para iniciar el proyecto. A medida que avanza la actividad, los estudiantes asumen el papel principal del proyecto.

A. Preparación del proyecto por parte del profesor

El ABP se implementó en un curso de máster, debido a la mayor experiencia de los estudiantes en comparación con los de grado. Además, el número de estudiantes de máster por curso (normalmente de 8 a 10) es menor que el de los 16 a 20 estudiantes por grupo de laboratorio en los cursos de grado. Los estudiantes de máster se organizaron en parejas y se les pidió a todos que diseñaran un banco de pruebas para evaluar el rendimiento de un sistema 802.11. Se les proporcionó dos HackRF One por grupo, una máquina virtual Ubuntu 16.04 (VM) con GNU Radio y el enlace al proyecto *gr-ieee802-11*[3] GitHub con la implementación del estándar 802.11p, que se detalla en [7]. Durante el proyecto, los estudiantes tuvieron acceso al laboratorio de comunicaciones, donde había un puesto de trabajo con un ordenador, que tenía el mismo software que estaba instalado en la máquina virtual. El lugar de trabajo era compartido por todos los grupos, por lo que tuvieron que organizarse para poder acceder al laboratorio. La VM permitió a los estudiantes realizar parte de la actividad en casa, mientras que el lugar de trabajo en el laboratorio de comunicaciones les permitió probar sus avances con el hardware SDR.

B. Introducción del proyecto y tareas iniciales

Se esperaba que los estudiantes del máster realizaran parte de su investigación fuera de las horas lectivas, mientras que el tiempo en clase se dedicaba a comprobar la evolución de cada grupo y a ayudarles a continuar con su trabajo. La actividad requirió tres sesiones de clase (tres horas cada una). La primera sesión se utilizó para presentar el proyecto a los estudiantes. Se les presentó la VM, el proyecto GitHub, y su documentación citando a [7]. El proyecto SDR descrito en [7] está basado en dispositivos USRP, mientras que se esperaba que el proyecto en este caso se realizara con HackRF One. En esta primera sesión el profesor instruyó a los alumnos sobre las diferencias entre el USRP y el HackRF One. La tarea propuesta a los estudiantes en esta primera sesión fue adaptar el proyecto *gr-ieee802-11* para ser utilizado con el HackRF One. Una de las desventajas del HackRF One es que introduce un

[3]https://github.com/bastibl/gr-ieee802-11

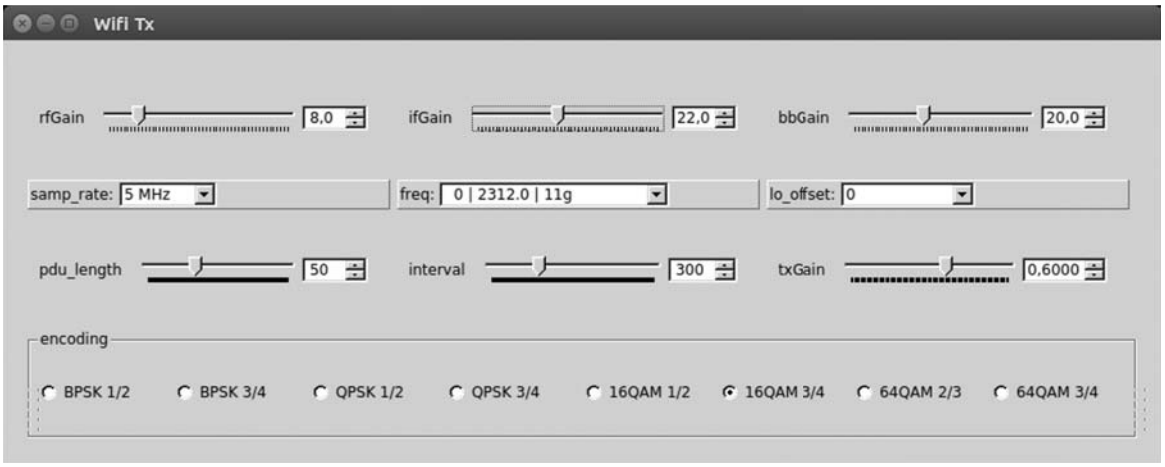


Fig. 1. Imagen de la GUI del transmisor, con los controles que permiten elegir las características de la transmisión.

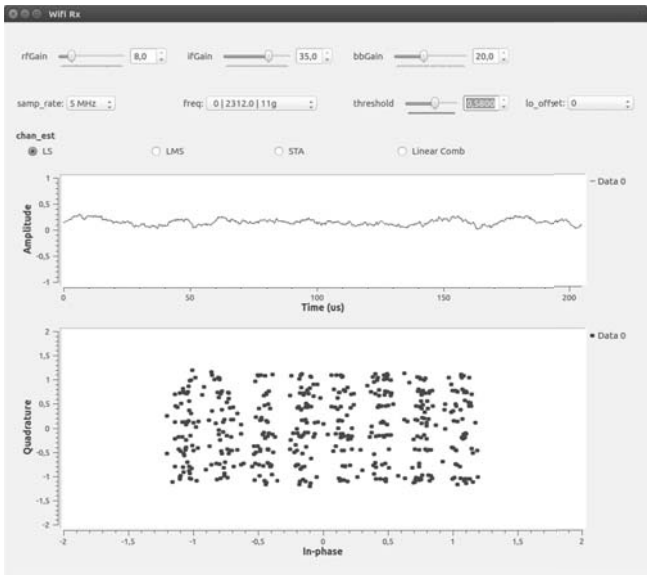


Fig. 2. Imagen de la interfaz gráfica del receptor. Los controles del receptor se colocan en la parte superior de la interfaz gráfica de usuario. La parte central muestra un gráfico de tiempo de la señal utilizada para determinar la llegada de un nuevo paquete. La parte inferior de la GUI corresponde a los símbolos recibidos en el último paquete. La imagen mostrada corresponde a una modulación 64-QAM.

pico de corriente continua (DC) en la señal IQ recibida. Se aconsejó a los estudiantes que instalaran el proyecto *gr-correctiq*[4] GitHub, y que usaran el bloque *correctIQ* en combinación con la fuente *osmocom-source* para eliminar el pico de DC introducido por HackRF One.

Las Interfaces Gráficas de Usuario (GUI) resultantes desarrolladas por los estudiantes para el transmisor y el receptor se muestran en Fig. 1 y Fig. 2, respectivamente. Fig. 1 muestra la GUI que permite ajustar los parámetros de transmisión, que consiste en cuatro filas de controles. El primero contiene las tres ganancias ajustables para las etapas de RF, frecuencia intermedia y banda base del HackRF One. La segunda fila de controles permite fijar las frecuencias de muestreo y de canal. Los controles de la tercera fila permiten ajustar el tamaño del paquete y el intervalo de tiempo entre paquetes consecutivos. Finalmente, la cuarta fila de controles permite elegir la modulación deseada. La Fig. 2 muestra la GUI con los controles del receptor, donde se observa que la parte superior contiene tres filas de controles para ajustar el rendimiento del receptor. Los controles de la primera y segunda fila de controles son similares a los que se muestran en Fig. 1, para el transmisor. Además se agregó un control *umbral* en la segunda fila y su función es establecer el umbral que activa la detección de un paquete entrante. La tercera fila de controles permite elegir el estimador de canal entre los tipos Least Squares (LS), Least Mean Squares (LMS), Spectral Temporal Averaging (STA) y COMB. La parte central de la GUI del receptor (Fig. 2) representa la variación temporal de la señal a la salida del detector de secuencia corta, que se utiliza para activar el receptor a la llegada de un nuevo paquete. Finalmente, la parte inferior de la interfaz gráfica de usuario muestra los símbolos recibidos.

Los controles en las GUIs de transmisión y recepción y la información mostrada en la GUI de recepción permitieron a los estudiantes de master ajustar los parámetros del transmisor y del receptor para cada esquema de modulación. Como resultado de esta actividad los estudiantes pudieron fijar las ganancias del transmisor y del receptor HackRF One. El proyecto *gr-ieee802-11* permite establecer la frecuencia de muestreo en 5 MHz o 10 MHz. Pero, considerando que la frecuencia máxima de muestreo del HackRF One es de 10 MHz y con el fin de reducir los requerimientos de computación del receptor, los estudiantes concluyeron que era más apropiado establecer la frecuencia de muestreo en 5 MHz. Los estudiantes también evaluaron cuál era el canal más adecuado para realizar sus experimentos, ya que la presencia de otros puntos de acceso WiFi (AP) interfería con los experimentos. En general, se obtuvieron mejores resultados cuando se utilizaron los canales más bajos o más altos en la banda de $2,4$ GHz. El número de AP en la banda de 5 GHz era menor que en la banda de $2,4$ GHz. Sin embargo, se

[4]https://github.com/ghostop14/gr-correctiq

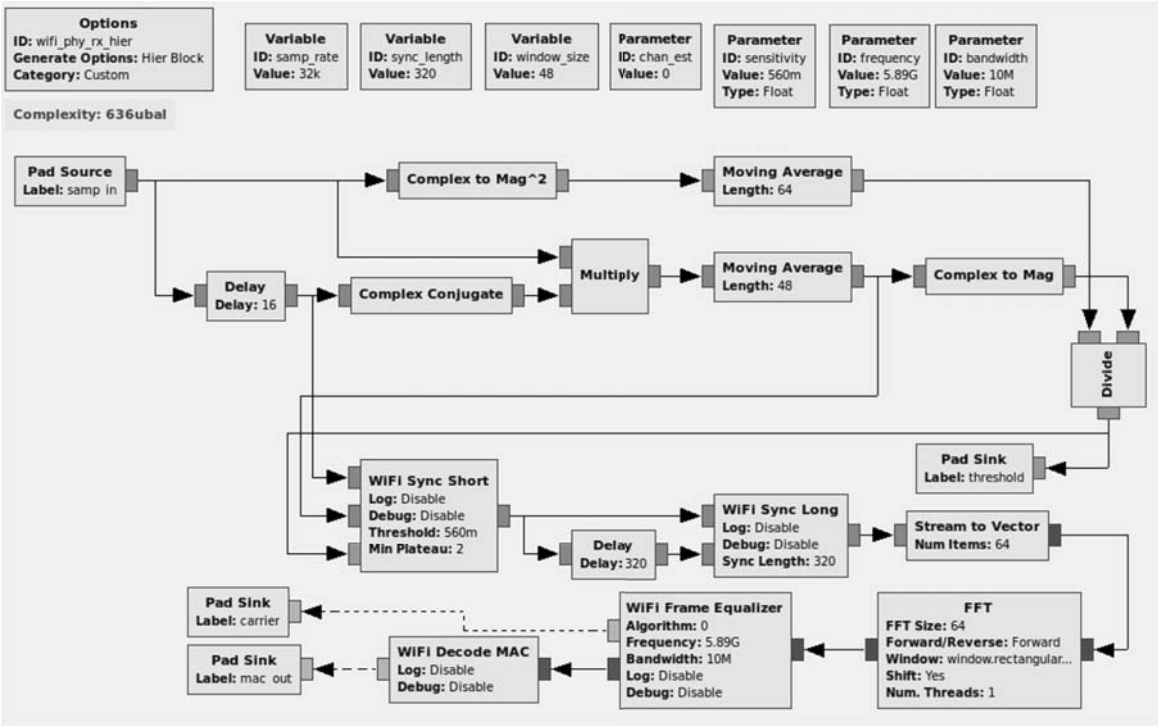


Fig. 3. Diagrama de bloques de la capa física del receptor WiFi.

Tabla I
TASA DE ERROR PARA LAS MODULACIONES DISPONIBLES CONSIDERANDO LA NUEVA LÍNEA DE VARIOS TAMAÑOS DE PAQUETES.

	Packet size (Bytes)		
Modulation	500	1000	1500
BPSK 1/2	0.11	0.21	0.31
BPSK 3/4	0.08	0.14	0.22
QPSK 1/2	0.06	0.11	0.15
QPSK 3/4	0.04	0.08	0.12
16 QAM 1/2	0.06	0.11	0.11
16 QAM 3/4	0.07	0.10	0.08
64 QAM 2/3	0.60	0.85	0.90
64 QAM 3/4	0.85	0.97	0.99

apreciaba un peor rendimiento del HackRF One en 5 GHz. Teniendo en cuenta este hecho y para evitar interferencias del AP vecino, los experimentos se realizaron a $2,3$ GHz.

C. Liderazgo estudiantil

Como se mencionó anteriormente, se espera que los estudiantes asuman progresivamente el liderazgo del trabajo. Esta parte de la metodología ABP duró la segunda y tercera sesiones de laboratorio, dependiendo del grado de autonomía de cada grupo de estudiantes.

La actividad propuesta a los alumnos en esta etapa consistió en leer un archivo y enviarlo a través del canal inalámbrico para evaluar el porcentaje de errores de transmisión. El texto elegido fue el libro de caballería *"El ingenioso hidalgo don Quijote de la Mancha"*, escrito por Miguel de Cervantes en 1605. La comparación entre los textos enviados y recibidos para distintas configuraciones del transmisor mostró que las diferencias se debían a las pérdidas de paquetes. Por lo tanto, para automatizar el análisis, el bloque *wireshark connector* disponible en el proyecto *gr-ieee802-11* se utilizó para crear dos archivos *pcap*, con paquetes enviados y recibidos, que fueron procesados posteriormente con un script Python para calcular la tasa de error.

Las tasas de error logradas con cada modulación se muestran en Tabla I. Hay que mencionar que al utilizar las modulaciones 16 QAM y 64 QAM, se requería habilitar el amplificador RF del HackRF One utilizado para la transmisión, mientras que en las demás modulaciones no se requería. Se consideraron tres tamaños de paquetes diferentes para evaluar su contribución a la tasa de error. En general, la tasa de error aumentaba con el tamaño del paquete. El alto aumento de la tasa de error entre 16 QAM y 64 QAM sugiere que la calidad del canal no era lo suficientemente buena para trabajar con modulaciones 64 QAM, ya que la mayoría de los paquetes se perdieron en estos casos. Contrariamente a lo que cabría esperar, se obtuvieron tasas de error más bajas con las modulaciones QPSK que las observadas con las modulaciones BPSK y 16 QAM. Una revisión de los experimentos condujo a resultados similares, lo que indica que será necesaria una comprensión más profunda de la aplicación para determinar la causa de una observación tan inesperada.

IV. TRASLADO DE LA EXPERIENCIA A LOS LABORATORIOS DE GIT

La experimentación de ABP permite a los estudiantes ser los actores principales de su proceso de aprendizaje. Además, la experiencia proporciona al profesor información valiosa relacionada con el rendimiento del sistema y también con los problemas a los que se enfrenta. Toda

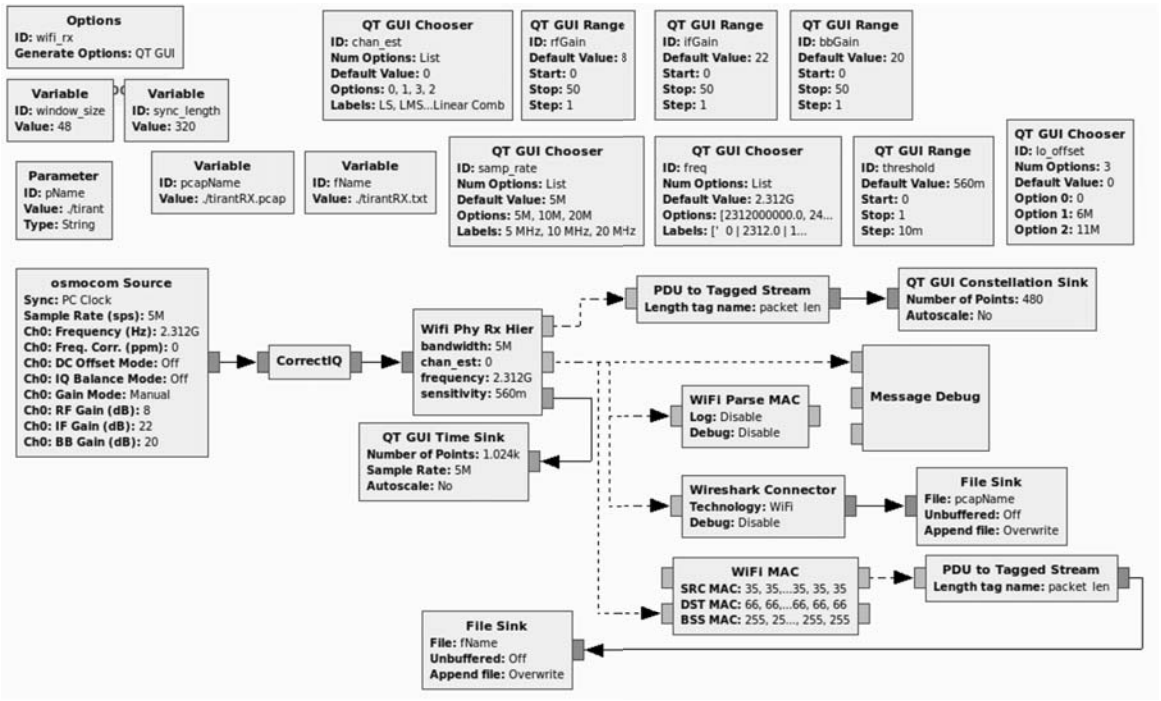


Fig. 4. Diagrama de bloques del receptor propuesto a los estudiantes de GIT.

esta información ha sido recogida y tenida en cuenta para diseñar una experiencia más guiada que realizan los estudiantes de GIT, es decir los estudiantes de máster retroalimentan en cierta forma a los de grado. Por ejemplo, uno de los comentarios de los estudiantes de máster fue que los recursos de computación requeridos por el receptor eran más altos que los del transmisor. Algunos de los estudiantes recomendaron en sus informes usar diferentes computadoras para el transmisor y el receptor. Para facilitar esta tarea, se construyó un bloque específico (ver Fig. 3) que contenía la implementación de la capa física del receptor, mientras que el bloque proporcionado en el proyecto *gr-ieee802-11* todavía se usaba para el transmisor.

A los estudiantes de GIT ya se les proporciona el diseño de los diagramas de bloques del receptor (Fig. 4) y del transmisor, que es similar al diagrama de la Fig. 4. Ambos diagramas incluyen el bloque *wireshark connector* para crear los archivos *pcap*. El script Python utilizado para analizar estos archivos *pcap* también se proporcionará a los estudiantes de GIT para que puedan centrarse en realizar la evaluación del sistema en lugar de diseñar el banco de pruebas. La implementación del receptor que se muestra en la Fig. 3 también estará disponible para los estudiantes, para que puedan identificar las funcionalidades requeridas en el detector.

V. RESULTADOS

Los autores ya evaluaron el impacto de la utilización de plataformas SDR en la capacidad de aprendizaje de los alumnos en los laboratorios docentes de GIT en un trabajo previo [6]. En ese trabajo se comprobó que las prácticas de laboratorio basadas en SDR contribuyen a mejorar ligeramente las calificaciones de los alumnos. Además, se observó que proporcionaban un mayor beneficio a aquellos alumnos que tenían más dificultades para seguir la asignatura. Lo que sugería que contribuye a aumentar la implicación de los estudiantes en su proceso de aprendizaje.

En esta ocasión se pretende cuantificar el impacto del uso de plataformas SDR en el compromiso de los estudiantes con su formación. Para ello se ha adaptado una encuesta que evalúa el compromiso en el trabajo [8], [9]:

- Q1. En la universidad me siento lleno de energía.
- Q2. Pienso que las actividades son relevantes y significativas.
- Q3. El tiempo me parece que ”vuela” cuando estoy estudiando.
- Q4. Me siento con gran fuerza y vigor mientras estudio.
- Q5. Me entusiasma lo que estudio.
- Q6. Cuando estoy trabajando o estudiando me olvido de todas las cosas que pasan a mi alrededor.
- Q7. El trabajo en la universidad me resulta ilusionante.
- Q8. Tengo ganas de ir a clase cuando me levanto por la mañana.
- Q9. Me satisface trabajar con intensidad en mi carrera.

Los estudiantes valoraron cada una de las preguntas en una escala entre 1 (nunca) y 5 (siempre). En ella se evalúan 3 aspectos relacionados con el compromiso:

- **Energía:** es un aspecto relacionado con la capacidad de resiliencia del alumno en la resolución de problemas. (Q1, Q4, Q8)

Tabla II
VARIACIÓN DE LOS VALORES PROMEDIO DE LA ENERGÍA, ABSORCIÓN E IMPLICACIÓN DE LOS ESTUDIANTES.

	Energía	Absorción	Implicación
Previo	2.50	2.82	3.17
Posterior	2.60	3.00	3.31
Diferencia	0.1	0.18	0.14

- **Absorción:** evalúa la capacidad del alumno para concentrarse en aquellas tareas que está realizando. (Q3, Q6, Q9)
- **Implicación:** evalúa la percepción del estudiante sobre la relevancia de las actividades que realiza. (Q2, Q5, Q7)

El laboratorio de FST en el que se llevó a cabo la experiencia está formado por dos partes. La primera parte se basa en simulaciones y cálculos de carácter teórico realizados en Matlab, mientras que en la segunda parte se plantean experiencias más aplicadas basadas en plataformas SDR. Para evaluar la contribución del uso de plataformas SDR en la motivación de los estudiantes se realizó la anterior encuesta en 2 ocasiones a lo largo del laboratorio. La primera al final del bloque de prácticas teóricas basadas en Matlab, y la segunda al final de curso tras haber realizado todas las prácticas basadas en SDR. El tiempo transcurrido entre ambas encuestas es de 2 meses, se ha considerado suficiente como para que en la realización de las encuestas por segunda vez los estudiantes no recuerden lo que contestaron en la primera ocasión.

Las actividades de laboratorio basadas en SDR tienen una contribución del 10 % de la nota final de la asignatura de FST, de 2º de GIT, por lo que la asistencia y realización de las mismas tiene carácter obligatorio. Sin embargo, la realización de las encuestas ha sido totalmente voluntaria, si bien es cierto que todos los 36 estudiantes asistentes al laboratorio respondieron las encuestas. Para cuantificar cada uno de los aspectos relacionados con la motivación académica se han agregado las respuestas a todas las preguntas relacionadas con cada uno de ellos. En la Tabla II se presenta la valoración de los estudiantes en cada uno de los aspectos en la encuesta previa y posterior, así como la diferencia en la valoración. Se aprecia una ligera mejora en todos los aspectos desde la primera a la segunda encuesta. El aspecto en el que se aprecia una mayor mejora es en la capacidad del estudiante en centrarse en la realización de las tareas que está llevando a cabo.

Además de presentar los valores promedio de cada uno de los aspectos, en la Fig. 5 se presenta el porcentaje de veces que los estudiantes asignaron cada valor en cada uno de los aspectos. Además de aumentar la valoración, en promedio, en la segunda encuesta se aprecia cómo en las respuestas relacionadas con la energía y la absorción las valoraciones se concentran más en valores intermedios. Mientras que en el caso de la implicación se aprecia un ligero incremento de las valoraciones más extremas (nunca y siempre).

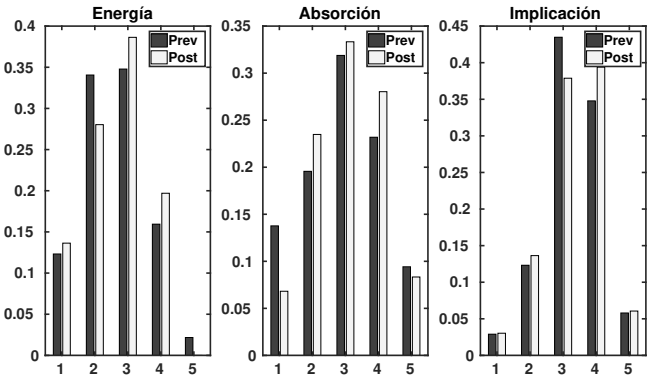


Fig. 5. Porcentaje de respuestas en cada uno de los aspectos. Energía (izquierda), absorción (centro), implicación (derecha). En azul se muestran los resultados de la encuesta previa y en amarillo los de la realizada al final del semestre.

VI. CONCLUSIONES

En este trabajo se ha presentado la experiencia llevada a cabo con los estudiantes de máster (MITUV) y grado (GIT). Aprovechando el menor número de estudiantes y su mayor nivel de formación y autonomía se ha propuesto a los estudiantes de máster una estrategia basada en ABP, en el que se les proponía la implementación de un banco de pruebas para un sistema de comunicación inalámbrica basada en el protocolo 802.11. A partir de la experiencia llevada a cabo con los estudiantes de máster se ha preparado un material más guiado, adaptado para los estudiantes de la asignatura FST del grado GIT.

El hecho de que los estudiantes de máster hayan trabajado previamente el material, ha contribuido a adaptar mejor las guías de prácticas presentadas a los estudiantes de grado. Además, se ha completado un estudio previo en el que se comprobó que la realización de prácticas más aplicadas basadas en plataformas SDR contribuía a mantener la motivación de los estudiantes a lo largo del semestre. En esta ocasión se ha empleado una encuesta para cuantificar los tres aspectos (energía, absorción e implicación) relacionados con la motivación profesional. El resultado de la encuesta ha probado que la realización de prácticas aplicadas ha contribuido principalmente a mejorar la capacidad de los estudiantes para centrarse en la tarea que están realizando. También ha contribuido a mejorar la percepción que los estudiantes tienen sobre la relevancia de las tareas que llevan a cabo en el laboratorio. En menor medida, pero también ha contribuido a mejorar la resiliencia de los estudiantes en la resolución de problemas.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía a través del proyecto BIA2016-76957-C3-1-R y por el "Servei de Formació Permanent i Innovació Educativa" de la Universitat de València a través del proyecto UV-SFPIE-RMD18-841566.

REFERENCIAS

[1] M. El-Hajjar, Q. A. Nguyen, R. G. Maunder, and S. X. Ng, "Demonstrating the practical challenges of wireless communications

using usrp," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 194–201, 2014.

[2] V. P. G. Jimenez, A. L. Serrano, B. G. Guzman, and A. G. Armada, "Learning mobile communications standards through flexible software defined radio base stations," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 116–123, 2017.

[3] R. W. Stewart, L. Crockett, D. Atkinson, K. Barlee, D. Crawford, I. Chalmers, M. Mclernon, and E. Sozer, "A low-cost desktop software defined radio design environment using matlab, simulink, and the rtl-sdr," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 64–71, 2015.

[4] R. Ettus, "Universal software radio peripheral (usrp)," https://www.ettus.com, 2015, (Visited on: 27/02/2019).

[5] Y. Linn, "An ultra low cost wireless communications laboratory for education and research," *IEEE Transactions on Education*, vol. 55, no. 2, pp. 169–179, 2012.

[6] J. Segura-García, A. Soriano-Asensi, C. Botella-Mascarell, S. Felici-Castell, and M. García Pineda, "Uso de software defined radio en la enseñanza de sistemas de telecomunicaciones," in *Actas de las XIII Jornadas de Ingeniería Telemática (Jitel 2017)*. Sept, Valencia (Spain), 2017, pp. 370–378.

[7] B. Bloessl and C. Sommer, "Performance assessment of ieee 802.11p with an open source sdr-based prototype," *IEEE Transactions on Mobile Computing*, vol. 17, pp. 1162–1175, 5 2018.

[8] R. García-Ros, F. Pérez-González, J. M. Tomás, and I. Fernández, "The schoolwork engagement inventory: Factorial structure, measurement invariance by gender and educational level, and convergent validity in secondary education (12-18) years," *Journal of Psychoeducational Assessment*, vol. 36, pp. 588–603, 2017.

[9] W. Schaufeli and A. Bakker, "Uwes: Utrecht work engagement scale," Utrecht University, Tech. Rep., 2004.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Herramienta web portable para la creación, distribución y reproducción de vídeos DASH. MediaDASH Tool

Rafael Fayos-Jordán, Daniel García-Costa, Miguel García-Pineda, Santiago Felici-Castell, Jaume Segura-García

Departamento de Informática. Escuela Técnica Superior de Ingeniería
Universidad de Valencia
Avinguda de la Universitat s/n 46100 Burjassot, València (SPAIN).
rafael.fayos@uv.es, daniel.garcia@uv.es, miguel.garcia-pineda@uv.es, santiago.felici@uv.es, jaume.segura@uv.es

***Resumen*—Actualmente el tráfico en internet de vídeo supone más de un 75 % del total y se prevé que en 2022 esta cifra supere el 80 %. Estas cifras ponen de manifiesto la importancia del desarrollo de tecnologías relacionadas con la transmisión y optimización de vídeo que permitan al usuario una experiencia adecuada. Por ello, el estudio y desarrollo de los formatos adaptativos DASH están, cada vez más, tomando importancia en el marco del streaming de vídeo, sobre todo con la llegada de las plataformas de contenidos a la carta como Netflix, Amazon Prime Video, HBO, etc. En el presente articulo se presenta la herramienta MediaDashTool, concebida como una aplicación web portable destinada a permitir que cualquier usuario, de una forma intuitiva y fácil pueda comprimir y preparar vídeos para su difusión en streaming adaptativo y visualizarlos desde la propia aplicación. Esta herramienta puede ser de gran utilidad para usuarios que pretendan montar su propia plataforma DASH, a investigadores interesados en testar contenidos multimedia empleando esta tecnología así como a docentes que pretendan enseñar a los alumnos los distintos tipos de codificación y sus resultados.**

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

I. INTRODUCCIÓN

Según un estudio de previsiones realizado por Cisco en 2017 [1], el consumo de vídeo se ha medido en 56 Exabytes mensuales para ese año y se prevé un incremento hasta 2022 hasta llegar a los 240 Exabytes al mes que supondrá el 80 % del total de tráfico en internet. Además, el 20,3 % de este se realizará en formato UHD (Ultra High Definition) como consecuencia del incremento de resoluciones de los dispositivos conectados.

Estas altas tasas de tráfico requieren una serie de actuaciones en diversos campos [2] que permitan satisfacer las demandas de los usuarios finales. El primero de ellos deriva directamente del mercado de dispositivos como las Smart TVs que incrementan las resoluciones cada cierto tiempo. Como muestra de ello, hoy en día ya existen en el mercado dispositivos con resoluciones de 8K UHDTV (Ultra High Definition TeleVision, 4320p), lo que propicia que se creen nuevos contenidos audiovisuales adaptados a estas resoluciones. Íntimamente relacionado a este punto están los códecs y contenedores de vídeo. H264, H265, VP8 y VP9 tienen sus ventajas e inconvenientes según el caso [3]. En este trabajo, nos centraremos en el protocolo Dynamic Adaptative Streaming over HTTP (DASH) [4], cuyo objetivo es proporcionar en cada momento la mejor calidad posible según el contexto particular de cada cliente. Este es el protocolo estandarizado que mayor compatibilidad posee siendo utilizado por los principales servicios de streaming de vídeo en Internet como son YouTube, Netflix, Amazon Prime Video, etc. La transmisión de vídeo de forma adaptativa DASH se está expandiendo y consolidando durante los últimos años, siendo actualmente uno de los más empleados en la red.

En este artículo presentamos la herramienta MediaDASH Tool (MDT), una utilidad web portable, capaz de comprimir, preparar y visualizar vídeos para su difusión en DASH y Streaming sobre HTTP de manera muy intuitiva.Es fácilmente desplegable en sistemas operativos en base Unix sin necesidad de instalación de software alguno gracias a que integra un "build in web server" de PHP7 conservando la posibilidad de ser utilizada en todo tipo de servidores dedicados. Para ello, se abordarán los siguientes puntos:

- Desarrollar una aplicación web que permita, especificando diferentes parámetros como resolución, tamaño, tasa binaria, etc. Comprimir vídeos con diferentes códecs y contenedores.

- Preparar y reproducir vídeos en WebM, MP4, MPEG-DASH [5] y WebM-DASH [6] a través de la interfaz web.
- Añadir la capacidad de portabilidad, pudiendo ejecutarse en cualquier equipo Linux o Mac sin necesidad de instalar nada, conservando la capacidad de desplegar en un servidor Web dedicado
- Llevar a cabo la evaluación de la herramienta analizando el comportamiento de la misma empleando diversos métodos de streaming.

La herramienta presentada en artículo está disponible en el repositorio GIT [7] para todo aquel que quiera utilizarla y/o mejorarla. Además en este repositorio se irán aportando mejoras y nuevas características en las que se está trabajando actualmente.

El resto del artículo está divido en las siguientes secciones. La sección 2 muestra algunos de los trabajos relacionados con plataformas de streaming. En la sección 3 se habla de la herramienta especificando su arquitectura y el funcionamiento de la misma. Por último, en la sección 4 se concluye el trabajo y se muestran los trabajos futuros.

II. TRABAJOS RELACIONADOS

Hoy en día, existen diversas plataformas y utilidades de streaming de vídeo adaptativo que implementan soluciones similares a la presentada en este artículo, pero con sus respectivas diferencias:

- Youtube [8]: Se trata de una plataforma web que permite realizar la codificación, compresión y difusión de vídeos de sus usuarios. En 2015 dejó de emplear un reproductor Flash en pos de uno HTML5, momento en que empezó a distribuir sus contenidos en DASH. La principal diferencia con MDT es que ésta plataforma no permite al usuario seleccionar la codificación de los streams ni tener un control de la plataforma.
- Netflix [9]: Esta plataforma de vídeo en streaming bajo demanda también se caracteriza por transmitir sus contenidos en DASH, sin embargo, los usuarios no pueden subir sus propios vídeos.
- Hulu [10]: Esta plataforma, muy similar a Netflix, empleaba streaming basado en FLV (Flash Video) sobre RTMP (Real Time Messaging Protocol) o HLS (HTTP Live Streaming) pero también migró su aplicación para que funcionase bajo streaming adaptativo mediante DASH. Tal y como ocurre con la plataforma anterior, el usuario no puede publicar sus propios contenidos .
- Video Tester [11]: Se trata de un trabajo desarrollado en 2012 que evalúa la transmisión de vídeo entre un servidor y un cliente generando una serie de estadísticas de transmisión. No obstante, esta aplicación está focalizada en el live streaming en tiempo real realizado sobre RTP/RTCP (Real-time Transport Protocol/Real-time Transport Control Protocol).
- End-to-End DASH Platform [12]: como el anterior, está desarrollado mediante el framework Gstreamer y es una plataforma que permite la configuración de los pasos necesarios de cliente y servidor para el streaming DASH. Esta herramienta solo soporta la codificación H264.
- MediaDASH Tool v1 [13]: En este artículo se mostró una primera versión de la herramienta. Las principales mejoras introducidas son: *a)* la portabilidad sin la necesidad de instalar ningún software y *b)* la inclusión de una herramienta core para la gestión de la propia herramienta MediaDASH Tool.

III. HERRAMIENTA

MediaDASH Tool es una herramienta con interfaz web que permite de manera muy intuitiva poder comprimir y preparar vídeos 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. Incluso puede servir para docentes que requieran de este tipo de herramientas para explicar conceptos de streaming de vídeo en sus asignaturas. Se trata de una aplicación portable y multiplataforma, que funciona bajo sistemas GNU-Linux y MacOS y se encuentra en proceso de adaptación para funcionar bajo sistemas Windows. Dado que todas sus dependencias están embebidas en la propia aplicación, no requiere de ningún tipo de instalación ni configuración para ser ejecutada tras ser descargada del siguiente repositorio [7].

A. Arquitectura y Modos

La aplicación MDT dispone de dos modos de funcionamiento, un modo CLI que permite codificar y preparar vídeos DASH desde un terminal y un modo GUI que lanza una interfaz web que permite realizar las mismas operaciones de manera mas intuitiva. Además, toda la información de los vídeos que se procesan se va almacenando en una base de datos SQLite que permite visualizar dicha información accediendo al vídeo en cuestión. Todos los vídeos procesados se almacenan en la aplicación, pudiendo ser visualizados mediante un reproductor javascript con soporte para contenido DASH. En ambos modos, la herramienta utiliza FFprobe para extraer la información necesaria de los archivos de vídeo, FFmpeg para realizar las diferentes codificaciones disponibles y MP4Box para generar los ficheros MPD (Media Presentation Description) necesarios para reproducir DASH (ver Figura 1).

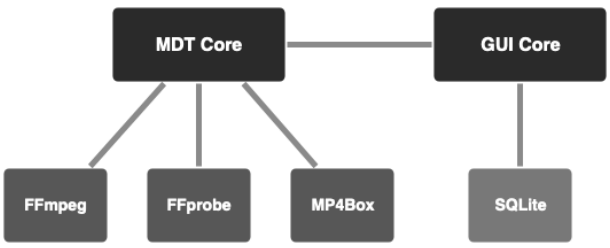


Figura 1. Arquitectura General de la Aplicación

El modo GUI necesita ser lanzado desde un terminal a través del comando `./mediadashtool.sh`. Al hacerlo, si no se especifican parámetros, la aplicación se inicia con los parámetros por defecto. La interfaz es entonces accesible desde la dirección de loopback (127.0.0.1) en el puerto 8080 y se ejecutará en el entorno de ejecución del terminal en el que se haya lanzado. Este comportamiento puede ser modificado pasando diferentes parámetros al lanzar la aplicación. La tabla I muestra los diferentes parámetros admitidos y sus respectivos significados.

Tabla I
PARÁMETROS QUE ACEPTA LA APLICACIÓN AL LANZARSE.

Parámetro	Descripción
-c, –config	Editar el fichero de configuración persistente
-a, –addr <listening address>	IP de escucha de la aplicación (0.0.0.0 para todas las disponibles)
-h, –help	Muestra el mensaje de ayuda
-p, –port <listening port>	Puerto de escucha (requiere privilegios root para menos de 1024)
-e, –encode <params>	Lanzar la codificación en modo MDT Core
-b, –background	Ejecuta la aplicación en segundo plano (log en mdt/log.out)
-k, –kill	Cierra la aplicación si se esta ejecutando en segundo plano

B. GUI Core

El GUI Core de la aplicación está escrito íntegramente en PHP7 y se sirve como aplicación web a través del Build-in web server del intérprete de PHP. Para que este fuera totalmente portable, la versión integrada es una compilación propia del interprete que integra todas las librerías necesarias con ligeras modificaciones en el código para evitar rutas absolutas. Además, dispone de un fichero de configuración del interprete, que permite modificar los parámetros propios de PHP, así como activar o desactivar los módulos integrados.

Esto convierte MediaDASH Tool en un entorno de desarrollo y pruebas para la codificación de vídeo y streaming en DASH. La arquitectura seguida es una modificación del patrón Modelo-Vista-Controlador, que simplifica enormemente la estructura, añadiendo un sistema de routing que permite escribir URLs más amigables y a la vez, prescindir de extensiones de archivo en las mismas. La Figura 2 resume la arquitectura adoptada.

La estructura de la aplicación se divide en 3 niveles (ver Figura 3). El nivel raíz es el que se muestra por defecto cuando se accede a la aplicación. Desde éste se puede acceder a los elementos del segundo nivel, que se compone de los listados de vídeos en sus diferentes estados y las opciones de subida de vídeos. El último nivel está compuesto por las acciones de codificación de vídeos y por el reproductor embebido.

Con la interfaz lanzada y accediendo a ella a través del navegador, se abre por defecto el listado de vídeos procesados por la aplicación (sección “Encoded Videos”). Este listado muestra todos los vídeos que hayan sido codificados utilizando la herramienta en modo GUI. Si

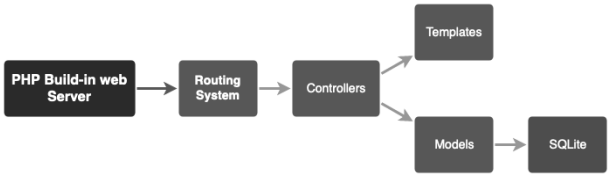


Figura 2. Arquitectura del GUI Core.

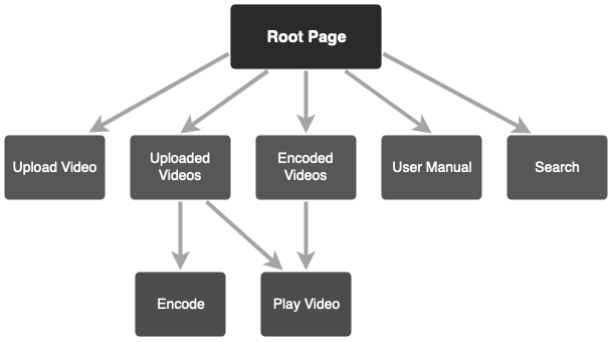


Figura 3. Jerarquía de niveles del GUI Core.

se accede a uno de estos vídeos se abre el reproductor integrado, a la vez que permite ver toda la información del vídeo seleccionado (opción Info) o eliminarlo (opción Delete). Si se trata de un vídeo DASH (archivo con extensión mpd), éste se reproduce haciendo uso del reproductor Dashif, a la vez que aporta información visual a través de una gráfica en tiempo real que indica el bitrate del vídeo (ver Figura 4).

Para codificar un vídeo, primero es necesario subirlo a la aplicación (botón “Upload Video”). Este proceso simplemente copia el vídeo seleccionado en la estructura de directorios de la herramienta y lo inserta en la base de datos, a la vez que extrae información acerca de este.

Una vez cargado el vídeo en la herramienta, éste aparecerá en el listado de vídeos subidos (sección “Uploaded Videos”). Accediendo a alguno de los vídeos subidos, éste se abrirá en el reproductor, aunque en este caso aparece disponible un botón adicional que permite codificar el vídeo (botón “Encode”). Al acceder a codificar un vídeo, existen diferentes opciones disponibles, en función del contenedor de destino que se desee utilizar.

En el caso de seleccionar una codificación DASH, será necesario la resolución y el bitrate para cada uno de los chunks que se quiera crear. En caso contrario, solo es

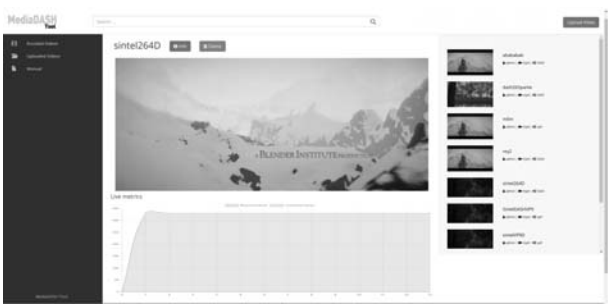


Figura 4. Reproductor integrado de la aplicación.

necesario seleccionar una resolución y el bitrate máximo de destino. Por último, en cualquiera de los casos es necesario indicar el GOP (distancia entre fotogramas tipo I) y los FPS (fotogramas por segundo).

Cuando un vídeo es codificado, sea en el formato que sea, éste aparecerá en el listado de vídeos codificados (sección "Encoded Videos"). Estos archivos no pueden volver a ser codificados y es necesario hacerlo desde el archivo original, que se mantiene en el listado de vídeos subidos (sección "Uploaded Videos") para su posterior uso en futuras codificaciones. En cualquier caso, todos los vídeos pueden ser eliminados mediante el botón Delete. Cuando se trate de un vídeo DASH, se eliminan tanto el archivo MPD como todos los chunks asociados a éste y en cualquier caso, tanto la miniatura, como todas las entradas en la base de datos referentes a ese vídeo serán también eliminadas.

C. MDT Core y línea de comandos

El MDT Core es la parte de la herramienta que se encarga de realizar las diferentes operaciones sobre los vídeos y está escrito íntegramente en BASH. Ese módulo es llamado desde el GUI Core cuando se lanza una codificación y se ejecuta en segundo plano para evitar bloquear las acciones del usuario hasta que la codificación finalice. Tiene la particularidad de poder ser llamado por línea de comandos. Por tanto, puede utilizarse sin necesidad de lanzar la interfaz de la aplicación. Para la extracción de información acerca de los vídeos, se utiliza FFprobe, que devuelve un amplio conjunto de información asociada al vídeo y audio del archivo. En el caso de la codificación, se utilizan FFmpeg y MP4Box. El primero de estos es el utilizado para la codificación en MP4 (H264, H265), WebM (VP8, VP9) y para generar la codificación WebM-DASH. Para la codificación MPEG-DASH se utiliza MP4Box. En ambos casos cuando se trata de DASH, el resultado final es un fichero MPD que contiene la información de los diferentes chunks a diferentes resoluciones, asociados a sus respectivos bitrates, para realizar la adaptación en el lado del cliente.

Independientemente de cómo se invoque al MDT Core con el fin de lanzar una codificación, es necesario indicar una serie de parámetros. El significado de los diferentes parámetros puede verse en la Tabla II.

En el caso de la codificación en DASH, se indican múltiples bitrates con múltiples resoluciones para cada uno de los chunks, simplemente añadiendo tantos parámetros "-b" y "-res" como chunks se desee codificar, por ejemplo:

Tabla II
DESCRIPCIÓN DE LOS PARÁMETROS DEL MDT CORE.

Parámetro	Descripción
-cod	Codec a utilizar [H264, H265, VP8, VP9]
-b	Bitrate
-res	Resolucion (AxB)
-fps	Fotogramas por segundo
-gop	Distancia entre fotogramas tipo I
-in	Fichero de entrada
-out	Fichero de salida

```
./mediadashtool.sh -e -c=vp9 -b=100 -res=160x90 -
    b=200 -res=320x180 -fps=20 -gop=24 -in=
    source/video/original.y4m -out=DASHdestino
```

IV. CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se ha presentado la herramienta MediaDASH Tool [7], una aplicación web portable destinada a permitir que cualquier usuario, de una forma intuitiva y fácil pueda comprimir y preparar vídeos para su difusión en streaming adaptativo y visualizarlos desde la propia aplicación. Se trata de una aplicación en constante evolución, cuyas características están siendo mejoradas. Se pretende en breve añadir nuevas funcionalidades como la recopilación de parámetros en la reproducción de los vídeos DASH, así como la posibilidad de codificar vídeo en AV1 [14]. Por! último, invitamos a la comunidad científica a aportar mejoras para crear una aplicación consistente y sólida en el ámbito del streaming DASH.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado 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] B. K. Wiederhold, G. Riva, and G. Graffigna, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022," *Annual Review of CyberTherapy and Telemedicine*, 2019.

[2] B. Bing, *Next-Generation Video Coding and Streaming*, 2015.

[3] J. Bienik, M. Uhrina, M. Kuba, and M. Vaculik, "Performance of H.264, H.265, VP8 and VP9 Compression Standards for High Resolutions," in *NBiS 2016 - 19th International Conference on Network-Based Information Systems*, 2016.

[4] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys and Tutorials*, 2015.

[5] ISO/IEC, "Dynamic adaptive streaming over HTTP (DASH): Media presentation description and segment formats," 2014.

[6] The WebM Project, "WebM Dash Specification," 2010. [Online]. Available: http://wiki.webmproject.org/adaptive-streaming/webm-dash-specification

[7] U. de València, "MediaDashTool Repositorio GIT." [Online]. Available: https://inmaculados.uv.es/MediaDashTool/develop.git

[8] D. K. Krishnappa, D. Bhat, and M. Zink, "DASHing YouTube: An analysis of using DASH in YouTube video service," in *Proceedings - Conference on Local Computer Networks, LCN*, 2013.

[9] J. Martin, Y. Fu, N. Wourms, and T. Shaw, "Characterizing Netflix bandwidth consumption," in *2013 IEEE 10th Consumer Communications and Networking Conference, CCNC 2013*, 2013.

[10] N. Weil, "Hulu's Move to DASH," 2015.

[11] I. Ucar, J. Navarro-Ortiz, P. Ameigeiras, and J. M. Lopez-Soler, "Video tester - A multiple-metric framework for video quality assessment over IP networks," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, 2012.

[12] D. Gómez, F. Boronat, M. Montagud, and C. Luzón, "End-to-end DASH platform including a network-based and client-based adaptive quality switching module," 2016.

[13] M. García-Pineda, D. García-Costa, J. Hannecke-Esteve, S. Felici-Castell, and J. Segura-Garcia, "Mediadash tool: Plataforma web para la codificación, difusión y recepción de videos dash," *XIII Jornadas de Ingeniería telemática (JITEL 2017). Libro de actas*, pp. 7–14, 2018.

[14] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi *et al.*, "An overview of core coding tools in the av1 video codec," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 41–45.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Herramienta web portable para la creación, distribución y reproducción de vídeos DASH. MediaDASH Tool

Rafael Fayos-Jordán, Daniel García-Costa, Miguel García-Pineda, Santiago Felici-Castell, Jaume Segura-García

Departamento de Informática. Escuela Técnica Superior de Ingeniería
Universidad de Valencia
Avinguda de la Universitat s/n 46100 Burjassot, València (SPAIN).
rafael.fayos@uv.es, daniel.garcia@uv.es, miguel.garcia-pineda@uv.es, santiago.felici@uv.es, jaume.segura@uv.es

***Resumen*—Actualmente el tráfico en internet de vídeo supone más de un 75 % del total y se prevé que en 2022 esta cifra supere el 80 %. Estas cifras ponen de manifiesto la importancia del desarrollo de tecnologías relacionadas con la transmisión y optimización de vídeo que permitan al usuario una experiencia adecuada. Por ello, el estudio y desarrollo de los formatos adaptativos DASH están, cada vez más, tomando importancia en el marco del streaming de vídeo, sobre todo con la llegada de las plataformas de contenidos a la carta como Netflix, Amazon Prime Video, HBO, etc. En el presente articulo se presenta la herramienta MediaDashTool, concebida como una aplicación web portable destinada a permitir que cualquier usuario, de una forma intuitiva y fácil pueda comprimir y preparar vídeos para su difusión en streaming adaptativo y visualizarlos desde la propia aplicación. Esta herramienta puede ser de gran utilidad para usuarios que pretendan montar su propia plataforma DASH, a investigadores interesados en testar contenidos multimedia empleando esta tecnología así como a docentes que pretendan enseñar a los alumnos los distintos tipos de codificación y sus resultados.**

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

I. INTRODUCCIÓN

Según un estudio de previsiones realizado por Cisco en 2017 [1], el consumo de vídeo se ha medido en 56 Exabytes mensuales para ese año y se prevé un incremento hasta 2022 hasta llegar a los 240 Exabytes al mes que supondrá el 80 % del total de tráfico en internet. Además, el 20,3 % de este se realizará en formato UHD (Ultra High Definition) como consecuencia del incremento de resoluciones de los dispositivos conectados.

Estas altas tasas de tráfico requieren una serie de actuaciones en diversos campos [2] que permitan satisfacer las demandas de los usuarios finales. El primero de ellos deriva directamente del mercado de dispositivos como las Smart TVs que incrementan las resoluciones cada cierto tiempo. Como muestra de ello, hoy en día ya existen en el mercado dispositivos con resoluciones de 8K UHDTV (Ultra High Definition TeleVision, 4320p), lo que propicia que se creen nuevos contenidos audiovisuales adaptados a estas resoluciones. Íntimamente relacionado a este punto están los códecs y contenedores de vídeo. H264, H265, VP8 y VP9 tienen sus ventajas e inconvenientes según el caso [3]. En este trabajo, nos centraremos en el protocolo Dynamic Adaptative Streaming over HTTP (DASH) [4], cuyo objetivo es proporcionar en cada momento la mejor calidad posible según el contexto particular de cada cliente. Este es el protocolo estandarizado que mayor compatibilidad posee siendo utilizado por los principales servicios de streaming de vídeo en Internet como son YouTube, Netflix, Amazon Prime Video, etc. La transmisión de vídeo de forma adaptativa DASH se está expandiendo y consolidando durante los últimos años, siendo actualmente uno de los más empleados en la red.

En este artículo presentamos la herramienta MediaDASH Tool (MDT), una utilidad web portable, capaz de comprimir, preparar y visualizar vídeos para su difusión en DASH y Streaming sobre HTTP de manera muy intuitiva.Es fácilmente desplegable en sistemas operativos en base Unix sin necesidad de instalación de software alguno gracias a que integra un "build in web server" de PHP7 conservando la posibilidad de ser utilizada en todo tipo de servidores dedicados. Para ello, se abordarán los siguientes puntos:

- Desarrollar una aplicación web que permita, especificando diferentes parámetros como resolución, tamaño, tasa binaria, etc. Comprimir vídeos con diferentes códecs y contenedores.

- Preparar y reproducir vídeos en WebM, MP4, MPEG-DASH [5] y WebM-DASH [6] a través de la interfaz web.
- Añadir la capacidad de portabilidad, pudiendo ejecutarse en cualquier equipo Linux o Mac sin necesidad de instalar nada, conservando la capacidad de desplegar en un servidor Web dedicado
- Llevar a cabo la evaluación de la herramienta analizando el comportamiento de la misma empleando diversos métodos de streaming.

La herramienta presentada en artículo está disponible en el repositorio GIT [7] para todo aquel que quiera utilizarla y/o mejorarla. Además en este repositorio se irán aportando mejoras y nuevas características en las que se está trabajando actualmente.

El resto del artículo está divido en las siguientes secciones. La sección 2 muestra algunos de los trabajos relacionados con plataformas de streaming. En la sección 3 se habla de la herramienta especificando su arquitectura y el funcionamiento de la misma. Por último, en la sección 4 se concluye el trabajo y se muestran los trabajos futuros.

II. TRABAJOS RELACIONADOS

Hoy en día, existen diversas plataformas y utilidades de streaming de vídeo adaptativo que implementan soluciones similares a la presentada en este artículo, pero con sus respectivas diferencias:

- Youtube [8]: Se trata de una plataforma web que permite realizar la codificación, compresión y difusión de vídeos de sus usuarios. En 2015 dejó de emplear un reproductor Flash en pos de uno HTML5, momento en que empezó a distribuir sus contenidos en DASH. La principal diferencia con MDT es que ésta plataforma no permite al usuario seleccionar la codificación de los streams ni tener un control de la plataforma.
- Netflix [9]: Esta plataforma de vídeo en streaming bajo demanda también se caracteriza por transmitir sus contenidos en DASH, sin embargo, los usuarios no pueden subir sus propios vídeos.
- Hulu [10]: Esta plataforma, muy similar a Netflix, empleaba streaming basado en FLV (Flash Video) sobre RTMP (Real Time Messaging Protocol) o HLS (HTTP Live Streaming) pero también migró su aplicación para que funcionase bajo streaming adaptativo mediante DASH. Tal y como ocurre con la plataforma anterior, el usuario no puede publicar sus propios contenidos .
- Video Tester [11]: Se trata de un trabajo desarrollado en 2012 que evalúa la transmisión de vídeo entre un servidor y un cliente generando una serie de estadísticas de transmisión. No obstante, esta aplicación está focalizada en el live streaming en tiempo real realizado sobre RTP/RTCP (Real-time Transport Protocol/Real-time Transport Control Protocol).
- End-to-End DASH Platform [12]: como el anterior, está desarrollado mediante el framework Gstreamer y es una plataforma que permite la configuración de los pasos necesarios de cliente y servidor para el streaming DASH. Esta herramienta solo soporta la codificación H264.
- MediaDASH Tool v1 [13]: En este artículo se mostró una primera versión de la herramienta. Las principales mejoras introducidas son: *a)* la portabilidad sin la necesidad de instalar ningún software y *b)* la inclusión de una herramienta core para la gestión de la propia herramienta MediaDASH Tool.

III. HERRAMIENTA

MediaDASH Tool es una herramienta con interfaz web que permite de manera muy intuitiva poder comprimir y preparar vídeos 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. Incluso puede servir para docentes que requieran de este tipo de herramientas para explicar conceptos de streaming de vídeo en sus asignaturas. Se trata de una aplicación portable y multiplataforma, que funciona bajo sistemas GNU-Linux y MacOS y se encuentra en proceso de adaptación para funcionar bajo sistemas Windows. Dado que todas sus dependencias están embebidas en la propia aplicación, no requiere de ningún tipo de instalación ni configuración para ser ejecutada tras ser descargada del siguiente repositorio [7].

A. Arquitectura y Modos

La aplicación MDT dispone de dos modos de funcionamiento, un modo CLI que permite codificar y preparar vídeos DASH desde un terminal y un modo GUI que lanza una interfaz web que permite realizar las mismas operaciones de manera mas intuitiva. Además, toda la información de los vídeos que se procesan se va almacenando en una base de datos SQLite que permite visualizar dicha información accediendo al vídeo en cuestión. Todos los vídeos procesados se almacenan en la aplicación, pudiendo ser visualizados mediante un reproductor javascript con soporte para contenido DASH. En ambos modos, la herramienta utiliza FFprobe para extraer la información necesaria de los archivos de vídeo, FFmpeg para realizar las diferentes codificaciones disponibles y MP4Box para generar los ficheros MPD (Media Presentation Description) necesarios para reproducir DASH (ver Figura 1).

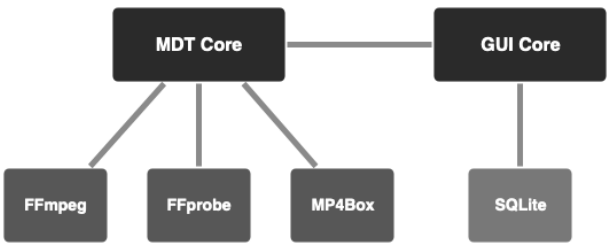


Figura 1. Arquitectura General de la Aplicación

El modo GUI necesita ser lanzado desde un terminal a través del comando `./mediadashtool.sh`. Al hacerlo, si no se especifican parámetros, la aplicación se inicia con los parámetros por defecto. La interfaz es entonces accesible desde la dirección de loopback (127.0.0.1) en el puerto 8080 y se ejecutará en el entorno de ejecución del terminal en el que se haya lanzado. Este comportamiento puede ser modificado pasando diferentes parámetros al lanzar la aplicación. La tabla I muestra los diferentes parámetros admitidos y sus respectivos significados.

Tabla I
PARÁMETROS QUE ACEPTA LA APLICACIÓN AL LANZARSE.

Parámetro	Descripción
-c, –config	Editar el fichero de configuración persistente
-a, –addr <listening address>	IP de escucha de la aplicación (0.0.0.0 para todas las disponibles)
-h, –help	Muestra el mensaje de ayuda
-p, –port <listening port>	Puerto de escucha (requiere privilegios root para menos de 1024)
-e, –encode <params>	Lanzar la codificación en modo MDT Core
-b, –background	Ejecuta la aplicación en segundo plano (log en mdt/log.out)
-k, –kill	Cierra la aplicación si se esta ejecutando en segundo plano

B. GUI Core

El GUI Core de la aplicación está escrito íntegramente en PHP7 y se sirve como aplicación web a través del Build-in web server del intérprete de PHP. Para que este fuera totalmente portable, la versión integrada es una compilación propia del interprete que integra todas las librerías necesarias con ligeras modificaciones en el código para evitar rutas absolutas. Además, dispone de un fichero de configuración del interprete, que permite modificar los parámetros propios de PHP, así como activar o desactivar los módulos integrados.

Esto convierte MediaDASH Tool en un entorno de desarrollo y pruebas para la codificación de vídeo y streaming en DASH. La arquitectura seguida es una modificación del patrón Modelo-Vista-Controlador, que simplifica enormemente la estructura, añadiendo un sistema de routing que permite escribir URLs más amigables y a la vez, prescindir de extensiones de archivo en las mismas. La Figura 2 resume la arquitectura adoptada.

La estructura de la aplicación se divide en 3 niveles (ver Figura 3). El nivel raíz es el que se muestra por defecto cuando se accede a la aplicación. Desde éste se puede acceder a los elementos del segundo nivel, que se compone de los listados de vídeos en sus diferentes estados y las opciones de subida de vídeos. El último nivel está compuesto por las acciones de codificación de vídeos y por el reproductor embebido.

Con la interfaz lanzada y accediendo a ella a través del navegador, se abre por defecto el listado de vídeos procesados por la aplicación (sección “Encoded Videos”). Este listado muestra todos los vídeos que hayan sido codificados utilizando la herramienta en modo GUI. Si

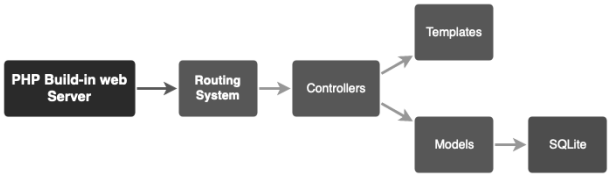


Figura 2. Arquitectura del GUI Core.

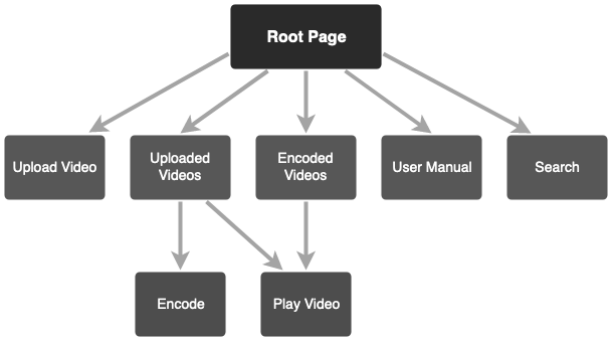


Figura 3. Jerarquía de niveles del GUI Core.

se accede a uno de estos vídeos se abre el reproductor integrado, a la vez que permite ver toda la información del vídeo seleccionado (opción Info) o eliminarlo (opción Delete). Si se trata de un vídeo DASH (archivo con extensión mpd), éste se reproduce haciendo uso del reproductor Dashif, a la vez que aporta información visual a través de una gráfica en tiempo real que indica el bitrate del vídeo (ver Figura 4).

Para codificar un vídeo, primero es necesario subirlo a la aplicación (botón “Upload Video”). Este proceso simplemente copia el vídeo seleccionado en la estructura de directorios de la herramienta y lo inserta en la base de datos, a la vez que extrae información acerca de este.

Una vez cargado el vídeo en la herramienta, éste aparecerá en el listado de vídeos subidos (sección “Uploaded Videos”). Accediendo a alguno de los vídeos subidos, éste se abrirá en el reproductor, aunque en este caso aparece disponible un botón adicional que permite codificar el vídeo (botón “Encode”). Al acceder a codificar un vídeo, existen diferentes opciones disponibles, en función del contenedor de destino que se desee utilizar.

En el caso de seleccionar una codificación DASH, será necesario la resolución y el bitrate para cada uno de los chunks que se quiera crear. En caso contrario, solo es

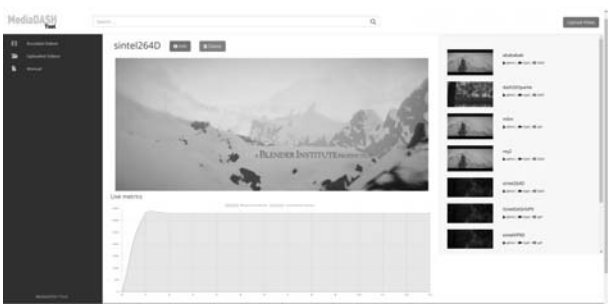


Figura 4. Reproductor integrado de la aplicación.

necesario seleccionar una resolución y el bitrate máximo de destino. Por último, en cualquiera de los casos es necesario indicar el GOP (distancia entre fotogramas tipo I) y los FPS (fotogramas por segundo).

Cuando un vídeo es codificado, sea en el formato que sea, éste aparecerá en el listado de vídeos codificados (sección "Encoded Videos"). Estos archivos no pueden volver a ser codificados y es necesario hacerlo desde el archivo original, que se mantiene en el listado de vídeos subidos (sección "Uploaded Videos") para su posterior uso en futuras codificaciones. En cualquier caso, todos los vídeos pueden ser eliminados mediante el botón Delete. Cuando se trate de un vídeo DASH, se eliminan tanto el archivo MPD como todos los chunks asociados a éste y en cualquier caso, tanto la miniatura, como todas las entradas en la base de datos referentes a ese vídeo serán también eliminadas.

C. MDT Core y línea de comandos

El MDT Core es la parte de la herramienta que se encarga de realizar las diferentes operaciones sobre los vídeos y está escrito íntegramente en BASH. Ese módulo es llamado desde el GUI Core cuando se lanza una codificación y se ejecuta en segundo plano para evitar bloquear las acciones del usuario hasta que la codificación finalice. Tiene la particularidad de poder ser llamado por línea de comandos. Por tanto, puede utilizarse sin necesidad de lanzar la interfaz de la aplicación. Para la extracción de información acerca de los vídeos, se utiliza FFprobe, que devuelve un amplio conjunto de información asociada al vídeo y audio del archivo. En el caso de la codificación, se utilizan FFmpeg y MP4Box. El primero de estos es el utilizado para la codificación en MP4 (H264, H265), WebM (VP8, VP9) y para generar la codificación WebM-DASH. Para la codificación MPEG-DASH se utiliza MP4Box. En ambos casos cuando se trata de DASH, el resultado final es un fichero MPD que contiene la información de los diferentes chunks a diferentes resoluciones, asociados a sus respectivos bitrates, para realizar la adaptación en el lado del cliente.

Independientemente de cómo se invoque al MDT Core con el fin de lanzar una codificación, es necesario indicar una serie de parámetros. El significado de los diferentes parámetros puede verse en la Tabla II.

En el caso de la codificación en DASH, se indican múltiples bitrates con múltiples resoluciones para cada uno de los chunks, simplemente añadiendo tantos parámetros "-b" y "-res" como chunks se desee codificar, por ejemplo:

Tabla II
DESCRIPCIÓN DE LOS PARÁMETROS DEL MDT CORE.

Parámetro	Descripción
-cod	Codec a utilizar [H264, H265, VP8, VP9]
-b	Bitrate
-res	Resolucion (AxB)
-fps	Fotogramas por segundo
-gop	Distancia entre fotogramas tipo I
-in	Fichero de entrada
-out	Fichero de salida

```
./mediadashtool.sh -e -c=vp9 -b=100 -res=160x90 -
    b=200 -res=320x180 -fps=20 -gop=24 -in=
    source/video/original.y4m -out=DASHdestino
```

IV. CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se ha presentado la herramienta MediaDASH Tool [7], una aplicación web portable destinada a permitir que cualquier usuario, de una forma intuitiva y fácil pueda comprimir y preparar vídeos para su difusión en streaming adaptativo y visualizarlos desde la propia aplicación. Se trata de una aplicación en constante evolución, cuyas características están siendo mejoradas. Se pretende en breve añadir nuevas funcionalidades como la recopilación de parámetros en la reproducción de los vídeos DASH, así como la posibilidad de codificar vídeo en AV1 [14]. Por! último, invitamos a la comunidad científica a aportar mejoras para crear una aplicación consistente y sólida en el ámbito del streaming DASH.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado 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] B. K. Wiederhold, G. Riva, and G. Graffigna, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022," *Annual Review of CyberTherapy and Telemedicine*, 2019.

[2] B. Bing, *Next-Generation Video Coding and Streaming*, 2015.

[3] J. Bienik, M. Uhrina, M. Kuba, and M. Vaculik, "Performance of H.264, H.265, VP8 and VP9 Compression Standards for High Resolutions," in *NBiS 2016 - 19th International Conference on Network-Based Information Systems*, 2016.

[4] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys and Tutorials*, 2015.

[5] ISO/IEC, "Dynamic adaptive streaming over HTTP (DASH): Media presentation description and segment formats," 2014.

[6] The WebM Project, "WebM Dash Specification," 2010. [Online]. Available: http://wiki.webmproject.org/adaptive-streaming/webm-dash-specification

[7] U. de València, "MediaDashTool Repositorio GIT." [Online]. Available: https://inmaculados.uv.es/MediaDashTool/develop.git

[8] D. K. Krishnappa, D. Bhat, and M. Zink, "DASHing YouTube: An analysis of using DASH in YouTube video service," in *Proceedings - Conference on Local Computer Networks, LCN*, 2013.

[9] J. Martin, Y. Fu, N. Wourms, and T. Shaw, "Characterizing Netflix bandwidth consumption," in *2013 IEEE 10th Consumer Communications and Networking Conference, CCNC 2013*, 2013.

[10] N. Weil, "Hulu's Move to DASH," 2015.

[11] I. Ucar, J. Navarro-Ortiz, P. Ameigeiras, and J. M. Lopez-Soler, "Video tester - A multiple-metric framework for video quality assessment over IP networks," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, 2012.

[12] D. Gómez, F. Boronat, M. Montagud, and C. Luzón, "End-to-end DASH platform including a network-based and client-based adaptive quality switching module," 2016.

[13] M. García-Pineda, D. García-Costa, J. Hannecke-Esteve, S. Felici-Castell, and J. Segura-Garcia, "Mediadash tool: Plataforma web para la codificación, difusión y recepción de videos dash," *XIII Jornadas de Ingeniería telemática (JITEL 2017). Libro de actas*, pp. 7–14, 2018.

[14] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi *et al.*, "An overview of core coding tools in the av1 video codec," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 41–45.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Despliegue de funciones de red virtualizadas en el cloud mediante ONAP

Adrià Martí, Albert Toro, Leonardo Ochoa-Aday, Adriana Fernández-Fernández, Toni Oller, Cristina Cervelló-Pastor, Jesús Alcober
Departamento de Ingeniería Telemática

Universitat Politècnica de Catalunya (UPC. BarcelonaTECH)

Esteve Terrades, 7, 08860 Castelldefels, Barcelona.

{adria.marti.luque, albert.toro.marin}@estudiant.upc.edu, {leonardo.ochoa, adriana.fernandez}@entel.upc.edu, antoni.oller@upc.edu, cristina@entel.upc.edu, jesus.alcober@upc.edu

***Resumen*- En este documento se explica la experiencia del despliegue de funciones de red virtualizadas en una red cloud privada mediante ONAP, junto con la arquitectura empleada, las máquinas utilizadas y el software necesario. ONAP permite diseñar, orquestar y gestionar todos los elementos relacionados con los servicios de red virtualizados, ofreciendo así una forma eficiente de automatización de redes. El objetivo final de este trabajo es conseguir un entorno en un cloud privado con funciones de red virtualizadas. Uno de los principales retos es la exigencia de ONAP desde el punto de vista de recursos necesarios. Todo lo comentado en este documento puede ser replicado en entornos de mayor envergadura que los definidos en este documento, pudiendo incluso ser utilizado en producción por operadoras de telefonía.**

***Palabras Clave*- jitel, telemática, sdn, nfv, cloud, virtualización, onap**

I. INTRODUCCIÓN

Una de los términos más usados hoy en día en el mundo de la telemática es el de virtualización, sin embargo, ha existido desde la década de los 60, cuando se entendía como separar los recursos de un servidor entre diferentes aplicaciones gracias al uso de máquinas virtuales. Con los años, el significado del término se ha ampliado sustancialmente [1].

Actualmente, hay muchos tipos de virtualización, pero en este documento nos centraremos en la virtualización a nivel de sistema operativo (o contenerización) y la virtualización de funciones de red.

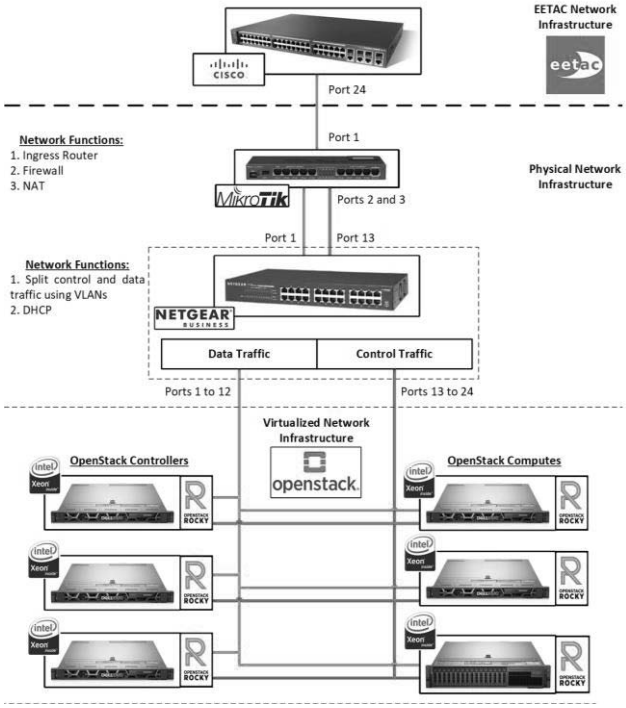
ONAP (Open Networking Automation Platform) es un proyecto de código abierto de la Fundación Linux, y proporciona una plataforma para la orquestación y automatización en tiempo real de funciones de red físicas y virtuales que permiten a los proveedores y desarrolladores de software, redes, IT y nube automatizar rápidamente los nuevos servicios y dar soporte a la gestión completa de su ciclo de vida. Su arquitectura es muy innovadora porque está dividida en dos entornos, un framework usado durante el tiempo dedicado al diseño de la infraestructura (servicios de nuevos tipos y funciones de red virtual (VNF) y funciones de red física (PNF)), y otro usado durante el tiempo de ejecución (creación y administración de instancias de servicios, VNF, PNF), convirtiendo a ONAP en una de las opciones preferentes de los diferentes proveedores de red [2][3].

A continuación, se describe, en primer lugar, la arquitectura diseñada para ofrecer estas funciones de red virtualizadas; en segundo lugar, se presenta el método utilizado para generar una nube privada mediante OpenStack y el mecanismo de orquestación basado en Kubernetes, para la automatización del despliegue, ajuste de escala y manejo de aplicaciones en contenedores, y Rancher, que permite gestionar clusters de Kubernetes. En tercer lugar, se explica cómo se han desplegado estas funciones de red en la plataforma de ONAP y, finalmente, se presentan los resultados y las conclusiones obtenidas.

II. ARQUITECTURA

Para poder realizar este proyecto, ha sido necesario el despliegue de un cloud privado empleando equipamiento e instalaciones de la Universitat

Politècnica de Catalunya (UPC), en el Campus del Baix Llobregat. El diagrama que detalla la interconexión de



los dispositivos físicos que componen este cloud se muestra en la Fig. 1

Fig. 1 Arquitectura del cloud privada.

La red interna que interconecta los servidores sobre los que se desplegó ONAP está en una red privada y está formada por un enrutador de la UPC, al cual está directamente conectado un enrutador bajo nuestra gestión con una configuración personalizada. Este enrutador utiliza la IP pública proporcionada por la UPC para permitir a los usuarios conectarse de forma remota al cloud, actuando como interconexión entre la red interna y la externa.

Conectado a este enrutador con reglas personalizadas, se encuentra un conmutador, con la función principal de filtrar el tráfico generado por nuestro cloud privado.

ONAP es muy exigente en recursos. Para poder realizar su despliegue, son necesarias 14 máquinas virtuales (1 Rancher y 13 nodos de Kubernetes) con 8 vCPUs por máquina, 112 en total, 16 GB de RAM por máquina virtual, 224 GB en total y 160 GB de almacenamiento por máquina virtual [4].

Por consiguiente, para la instalación se han utilizado 5 servidores Dell Rack Server R440 y 1 servidor Dell Rack Server R740.

De estos servidores R440, 3 de ellos actúan como controladores de OpenStack y están conectados al conmutador y a las máquinas virtuales creadas por OpenStack al mismo tiempo. Estos servidores son los que generarán el tráfico de control.

El resto de servidores forman la infraestructura de datos de OpenStack, que está conectada a los controladores a través del conmutador. Estos servidores son los que generarán tráfico de datos cuando las diferentes máquinas virtuales sean creadas.

Como son necesarias altas velocidades de interconexión entre los diferentes servidores, se utilizan cables Ethernet de categoría 6 a 10 Gbps para interconectar los diferentes elementos de red.

El tráfico de datos y de control de OpenStack está separado mediante 2 VLANs, una dedicada únicamente a control (cable rojo en la Fig. 1, etiquetada como 101) y la otra únicamente a datos (cable verde en la Fig. 1, con etiqueta 102).

Finalmente, este tráfico es dividido por el conmutador, el cual también realizará tareas DHCP.

Cada VLAN tiene su propio rango de IP, 192.168.100.0/24 para datos y 172.16.100.0/24 para control.

Las direcciones IP 192.168.100.1 y 172.16.100.1 están reservadas para el enrutador. El resto de direcciones IP se asignan a los servidores en un orden incremental, empezando por 192.168.100.10 y 172.16.100.10 respectivamente.

III. Creación de la nube privada

Una vez configurado y establecido el entorno a nivel físico y de red, se procedió al despliegue de la nube privada para poder alojar ONAP. Para ello, ha sido necesario realizar un conjunto de actuaciones, tal y como se describe en el siguiente apartado.

A. OpenStack con Terraform

Todos los servidores utilizados en este proyecto, tanto los R440 como el R740, tienen instalado Ubuntu 18.04 como sistema operativo.

Sobre este sistema operativo se ha desplegado OpenStack, una plataforma de código abierto la cual nos permite iniciar y gestionar servicios en la nube [5].

OpenStack es un elemento clave para este proyecto, ya que ofrece la automatización de las máquinas virtuales.

Para poder desplegar de forma eficiente y automática las máquinas virtuales en OpenStack, se ha usado Terraform, una herramienta para construir, cambiar y versionar la infraestructura virtual [6].

Usando OpenStack junto con Terraform, se ha creado 14 máquinas virtuales necesarias para el despliegue de ONAP.

Gracias a las características ofrecidas por Terraform para poder gestionar OpenStack, estas máquinas pueden ser desplegadas, escaladas y eliminadas de forma semiautomática utilizando sencillos ficheros de configuración, ahorrando así horas de trabajo manual.

B. Kubernetes

Para poder iniciar los servicios de ONAP son necesarios una gran cantidad de contenedores, convirtiendo el despliegue manual en inviable. Por consiguiente, todas las máquinas virtuales disponen de

Kubernetes, un sistema portátil y extensible de orquestación de contenedores de código abierto [7].

ONAP emplea Kubernetes para poder desplegar y gestionar de forma factible los contenedores en las diferentes máquinas virtuales, los cuales contienen todos los módulos y servicios necesarios para poder arrancar la plataforma.

Los servicios, al ser iniciados en contenedores, se pueden modificar o incluso eliminar sin afectar a la totalidad de la red virtual gracias al uso de comandos, convirtiendo Kubernetes en un elemento necesario para poder realizar susodicho despliegue.

C. Rancher

Kubernetes, aun siendo un orquestador de contenedores muy potente, no permite gestionar un gran número de contenedores de forma simple en diferentes máquinas virtuales que tengan relación entre ellas, por lo tanto, para solucionar este inconveniente, en este proyecto se utiliza Rancher.

Rancher es una plataforma de software de código abierto que permite ejecutar y administrar contenedores mediante una interfaz gráfica [8].

El uso de Rancher para gestionar los clústeres de Kubernetes de ONAP resuelve los problemas operativos y de seguridad de éstos, proporcionando así un plano de control unificado, muy necesario para poder administrar de forma controlada la gran cantidad de contenedores que se van a desplegar en las máquinas virtuales del cloud [9].

IV. ONAP

ONAP puede ser gestionado utilizando OpenStack o con Kubernetes. La instalación de ONAP con OpenStack está orientada para hacer pequeños despliegues de prueba, no válidos para un escenario real. Como en este proyecto se ha querido realizar un despliegue basado completamente en contenedores y pensado para replicar en entornos de producción con casos de uso reales, se ha decidido utilizar la alternativa de Kubernetes para iniciar el ONAP Operations Manager (OOM) [10].

OOM es la opción preferida a largo plazo ya que las necesidades de vCPU, la huella de memoria y el tiempo de arranque se reducen de 5 a 6 veces debido al nivel de virtualización [11].

Una de las características más importantes de ONAP es que está basado en artefactos, modelos independientes que se juntan para crear servicios de red que no requieren de intervención manual de personal altamente cualificado [12].

El portal de ONAP proporciona un panel unificado para el diseño y gestión de estos artefactos. También es extensible mediante la creación de aplicaciones de terceros [13].

Las aplicaciones de portal se prevén en siete áreas: diseño, planificación de operaciones, planificación de capacidad, gestión de tecnología, inserción de tecnología, gestión de rendimiento y gestión de plataforma [14].

V. DEMOSTRACIÓN DE ONAP

Una vez desplegado el OOM e instalado el portal, la forma más directa de comprobar el correcto funcionamiento de ONAP es mediante la demostración del firewall virtual (vFW), uno de los casos de uso más útil y completo que asegura el correcto funcionamiento de ONAP [15].

El servicio de red completo consta de un firewall virtual. En este caso de uso se usa una métrica para recopilar el número de paquetes que pasan a través del firewall en un período de 10 segundos. Existe una

700 o menor que 300. Si es así, un flujo de trabajo de APP-C configura el generador de paquetes para que genere 500 paquetes por 10 segundos. De esta manera, todos los aspectos de ONAP son demostrados y ejercitados [15].

La demostración pasa por más de 40 pasos, que incluyen incorporación de VNF, servicio y creación de políticas, servicio y distribución de pólizas, despliegue de servicios usando scripts VID y Robot, generación de tráfico y automatización en circuito cerrado [15].

VI. CONCLUSIONES

Este proyecto, aunque todavía en desarrollo, demuestra las potenciales ventajas que tendría ONAP en el mercado de los servicios de las telecomunicaciones, donde el uso de funciones de red virtualizadas ayudaría en gran medida a las empresas a innovar rápidamente e implementar las últimas tecnologías para responder a las últimas tendencias, como son 5G e IoT. La implementación descrita en este artículo ha servido como una prueba de concepto inicial para validar las funcionalidades de esta herramienta de orquestación en un entorno de virtualización de funciones de red. Como líneas de continuidad de este trabajo se pretende emplear este despliegue de ONAP para el diseño y puesta en funcionamiento de nuevas funciones virtualizadas para diversos casos de usos.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad del Gobierno de España a través del proyecto TEC2016-76795-C6-1-R. Los autores agradecen a Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels (EETAC), al Departamento de Ingeniería Telemática (ENTEL) y al grupo de investigación BAMPLA de la UPC por su apoyo en este trabajo.

REFERENCIAS[1]

[1] Charles David Graziano. (2011). A performance analysis of Xen and KVM hypervisors for hosting the Xen Worlds Project,,https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=3243&context=etd
[2] The Linux Foundation Projects. (2018), ONAP Platform , , https://www.onap.org/platform-2
[3] Cloudify. (2017). What is ONAP and what does it mean for you? https://cloudify.co/onap/what-is-onap/
[4] ONAP. (2018). Setting up ONAP, https://onap.readthedocs.io/en/casablanca/guides/onap-developer/settingup/index.html#requirements
[5] OpenStack. (2019) . Software, https://www.openstack.org/software/
[6] Terraform. (2019). Introduction to Terraform, https://www.terraform.io/intro/index.html
[7] Serdar Yegulalp. (2019). What is Kubernetes? Container orchestration explained, https://www.infoworld.com/article/3268073/what-is-kubernetes-container-orchestration-explained.html [Accedido: 20/05/2019]
[8] Rancher. (2018). Rancher Documentation V1.6, https://rancher.com/docs/rancher/v1.6/en/
[9] Marksei. (2018). What is Rancher? Containers in the age of Cattle, https://www.marksei.com/rancher/
[10] ONAP. (2018). ONAP on Kubernetes with Rancher, https://onap.readthedocs.io/en/casablanca/submodules/oom.git/docs/oom_setup_kubernetes_rancher.html#onap-on-kubernetes-with-rancher
[11] ONAP. (2017). ONAP Operations Manager , https://wiki.onap.org/pages/viewpage.action?pageId=3246809
[12] ONAP. (2018). Service Design, https://wiki.onap.org/display/DW/Service+Design
[13] ONAP. (2018). ONAP PORTAL, https://wiki.onap.org/display/DW/ONAP+Portal
[14] ONAP. (2018). SDC Artifact List, https://wiki.onap.org/display/DW/SDC+Artifacts+List
[15] ONAP. (2018). vFW CDS Casablanca https://wiki.onap.org/display/DW/vFW+CDS+Casablanca

[1] Referencias online: [Accedidas el 16/05/2019]

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Experiencia de trabajo en equipo multidisciplinar para la creación de un proyecto de ingeniería en créditos optativos en la EETAC

Jesus Alcober, Antoni Oller, David Remondo
Departamento de Ingeniería Telemática
Universitat Politècnica de Catalunya (UPC. BarcelonaTECH)
Esteve Terrades, 7, 08860 Castelldefels, Barcelona.
jesus.alcober@upc.edu, antoni.oller@upc.edu, remondo@entel.upc.edu

Resumen- Este trabajo presenta la experiencia de una asignatura de proyectos de ingeniería relativamente nueva; este año en su cuarta edición ya tenemos datos suficientes para extraer conclusiones sobre la consecución de sus objetivos y la aceptación que ha recibido por parte de los estudiantes. La novedad de esta asignatura optativa es, por un lado, la heterogeneidad del perfil de los estudiantes que la cursan, que pueden ser de aeronáutica, telemática o sistemas de telecomunicación; y, por otra parte, la visión desde tres ópticas totalmente distintas de un proyecto de ingeniería, como es el desarrollo de un proyecto concreto, la propuesta de un proyecto europeo, y una visión de negocio de los proyectos de ingeniería.

Palabras Clave: **assignatura proyectos, ERP, telemática, aeronáutica, proyectos europeos,**

I. INTRODUCCIÓN

Dentro de los estudios adaptados al Espacio Europeo Superior (EEES) de la Escola d'Enginyeria de Telecomunicacions i Aeroespacials de Castelldefels (EETAC) de la Universitat Politècnica de Catalunya (UPC. BarcelonaTech) se han introducido aspectos relacionados al ciclo de vida de un proyecto de ingeniería a una asignatura optativa de 6 créditos ECTS denominada "Projectes d'Enginyeria" (PE). La motivación inicial fue aportar la experiencia del profesorado con amplia experiencia en la redacción y evaluación de propuestas, así como en proyectos de emprendimiento.

El objetivo de este trabajo es el de presentar nuestra experiencia en una asignatura optativa que se basa en un aprendizaje basado en proyectos (PBL), en el que la EETAC ya tiene una experiencia de casi 20 años [1][3] a lo largo de diferentes asignaturas en la escuela [4][5][6]. El proyecto es el foco principal de la asignatura que obliga como requerimiento que éste pueda ser viable en el mercado.

A continuación, se describe en primer lugar la organización de la asignatura. En segundo lugar, se presentan resultados de encuestas de las últimas ediciones y por último las conclusiones obtenidas.

II. PROYECTOS DE INGENIERÍA

"Projectes d'enginyeria" (PE) es una asignatura optativa que se imparte en el cuarto curso del grado de Ingeniería Telemática, grado de Ingeniería en Sistemas de Telecomunicación, grado de Ingeniería de Sistemas Aeroespaciales (mención aeropuertos y aeronavegación), así como en la doble titulación de grado en Ingeniería de Sistemas Aeroespaciales y Ingeniería de Sistemas de Telecomunicación o Ingeniería Telemática y la doble titulación del grado de Ingeniería de Sistemas de Telecomunicación y el grado de Ingeniería Telemática.

Los objetivos de la asignatura son los siguientes:

- Adquirir conocimientos avanzados de gestión de proyectos de ingeniería.
- Analizar y construir propuestas de proyectos tanto de financiación privada como pública.
- Diseñar una estructura empresarial que aborde un proyecto propuesto.
- Seguir el ciclo de vida de un proyecto: oportunidad, propuesta, proyecto.
- Analizar el resultado de un proyecto (coste/beneficio).
- Explicar y defender sus soluciones en presentaciones y memoria.

La asignatura ofrece competencias genéricas como el aprendizaje autónomo, la toma de decisiones y especialmente el trabajo en equipo.

Una de las principales riquezas de la asignatura es el carácter multidisciplinar de los alumnos matriculados ya que pueden tener diferentes perfiles al provenir de los grados indicados anteriormente. Los perfiles con los que la asignatura

trabaja son tres: perfil *aero*, perfil de *sistemas*, perfil de *telemática*. Un alumno con un perfil *aero* proporciona al equipo conocimientos de necesidades y/o problemas por resolver en el sector aeroespacial. Un alumno con un perfil *sistemas* proporciona al equipo conocimientos de electrónica y comunicaciones. Un alumno con un perfil *telemática* proporciona al equipo conocimiento de comunicaciones, protocolos de comunicaciones y desarrollo de aplicaciones y servicios. Cuando se forman los grupos (4 o 5 alumnos) se intenta mantener esta dispersión y se evitan grupos con un único perfil. La potencialidad de los equipos de trabajo con estos perfiles hace que los proyectos que se puedan llevar a cabo sean altamente motivadores y originales ya que son los propios alumnos que proponen el proyecto a trabajar durante el curso. Estos proyectos pasarán por el filtro del profesorado de la asignatura.

La asignatura se viene impartiendo desde el cuatrimestre de primavera del 2017 y únicamente una vez al año. Durante estos cursos en que se ha impartido la asignatura se ha trabajado en proyectos de diversa índole que han sido propuestos por los alumnos con la guía del profesorado. Por ejemplo, en un proyecto para optimizar el seguimiento del equipaje de los viajeros en los aeropuertos y evitar sus pérdidas. Otro proyecto destacado, fue un dispositivo físico que detecta colisiones en vehículos móviles, que podrían ser autónomos, y establece una comunicación hacia un sistema de control de emergencias. En el presente curso se están trabajando en tres nuevos proyectos. Un primer proyecto propone la construcción de una biosfera para hacer experimentación científica. Un segundo proyecto, un dispositivo para la carga de vehículos eléctricos en marcha y una aplicación para poner en común usuarios que realizan trayectos comunes. El último proyecto propone la construcción de un piloto automático para el funcionamiento para aviones en tierra.

PE se organiza en tres partes que tienen un peso equivalente. En una primera parte se trabaja en la construcción de propuestas de proyectos de financiación pública y privada. En una segunda parte se trabaja en la gestión de proyectos de ingeniería desde un punto de vista de negocio y la tercera parte de la asignatura se simula la ejecución de un proyecto de ingeniería, su seguimiento y control.

A. Construcción de propuestas de proyectos de financiación pública y privada

Esta primera parte de la asignatura de “Proyectos de ingeniería” trata de la presentación y evaluación de propuestas de proyectos en convocatorias competitivas de entidades públicas. El motivo de incluir esta temática viene de la constatación del hecho de que muchos ingenieros, en su vida profesional, no solo habrán de participar en proyectos ya definidos, sino que en muchos casos podrán tener un papel –muchas veces fundamental- en la presentación de propuestas para su financiación. Tanto en el caso de empresas innovadoras en el sector privado, como en el de universidades o centros de investigación públicos, estas convocatorias provienen de entidades públicas en la mayoría de los casos. En nuestro entorno, y para el caso de proyectos de investigación o innovación, una buena parte de la financiación está basada en convocatorias competitivas de los programas de la Comisión europea. Por tanto, en esta parte de “Proyectos de ingeniería” se pretende que el estudiante adquiera familiaridad, a través de un ejercicio práctico, con el procedimiento de presentación de propuestas para este tipo de convocatorias, con el fin de que pueda servirle en su futuro trabajo para tener mayores probabilidades de éxito. También se considera que dicha experiencia pueda ser muy útil para la presentación de propuestas en otros contextos, como por ejemplo en las propuestas nacionales.

En el desarrollo de las actividades, los estudiantes trabajan en grupos para proponer un proyecto en una convocatoria real de la Comisión europea. La convocatoria que se les propone es del programa FET (Tecnologías Futuras y Emergentes), ya que es totalmente abierto en cuanto a temática, requiere un grado importante de innovación y posee un carácter multidisciplinar. La libertad temática facilita que trabajen estudiantes de diferentes grados en grupo, promoviendo también la creatividad e iniciativa personal del estudiante; por otro lado, el carácter multidisciplinar facilita la construcción de la propuesta en la mayor parte de los casos. En la preparación de la propuesta, se contempla que el grupo de estudiantes construya un consorcio de entidades participantes –ficticias o reales- que sean relevantes para la ejecución del proyecto.

Esta parte de “Proyectos de ingeniería” no solo contempla la preparación de la propuesta de proyecto, sino que también a cada estudiante se le requiere efectuar una evaluación de una propuesta de acuerdo con los baremos de evaluación de una convocatoria de la Comisión Europea. Además, el estudiante realiza dicha evaluación incluyendo las dos fases en que se realizan las evaluaciones técnicas por los expertos contratados por la Comisión Europea. La primera fase de evaluación es individual: el estudiante propone una serie de calificaciones de acuerdo con los baremos establecidos y adjunta un informe motivando sus calificaciones de forma detallada. En la segunda fase de evaluación, se reúne a los estudiantes que han evaluado una misma propuesta en una sesión de concertación, en la cual se han de poner de acuerdo en una calificación conjunta para cada uno de los aspectos analizados.

B. Gestión de proyectos de Ingeniería

La segunda de las tres partes de la asignatura de “Proyectos de ingeniería” trata de proporcionar una visión global del proyecto desde el punto de vista de negocio, es decir, si el proyecto va a tener beneficios o pérdidas y cuáles son las herramientas para analizarlo. Por tanto, el estudiante debe calcular cuáles son los ingresos del proyecto, y cómo puede conseguir estos ingresos y cuáles son los costes del proyecto.

Para ello se utiliza un software de código abierto llamado Odoo [2]. Es un ERP (Enterprise Resource Planning) muy extendido que cuenta con más de dos millones de usuarios en todo el mundo. Se instalan los siguientes módulos: empleados, proyecto, hojas de trabajo, CRM (Customer Relationship Management), ventas, compras, web y contabilidad.

Se le proporciona a cada estudiante una cuenta de usuario con permisos de administrador, de forma que se pueden crear ellos mismos como empleados y se asignan un coste por hora de 10 €. Previamente, se ha discutido qué significa este precio por hora respecto a la nómina correspondiente a recibir. Y se aprovecha para explicar los distintos conceptos de una nómina

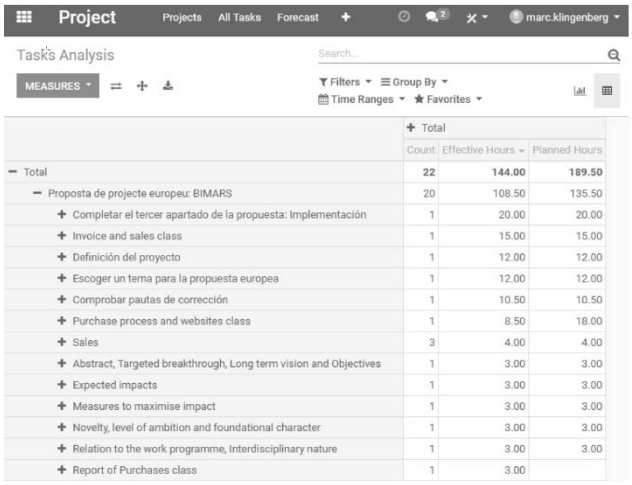
como son el sueldo base, la retención del IRPF y la Seguridad Social pagada por el trabajador y por la empresa.

Posteriormente se pide que creen un proyecto dentro para desarrollar la propuesta del proyecto europeo que se ha explicado previamente, así como las distintas tareas a realizar para llevar a cabo este proyecto. Las características principales de cada tarea son el nombre de la persona responsable y el tiempo estimado de ejecución en horas.

Durante la duración de todo el proyecto se les pide a todos los estudiantes que semanalmente reporten las horas de trabajo invertidas para llevar a cabo el proyecto.

Al cabo de un mes de iniciado el proyecto, podemos analizar el coste desde el punto de vista económico.

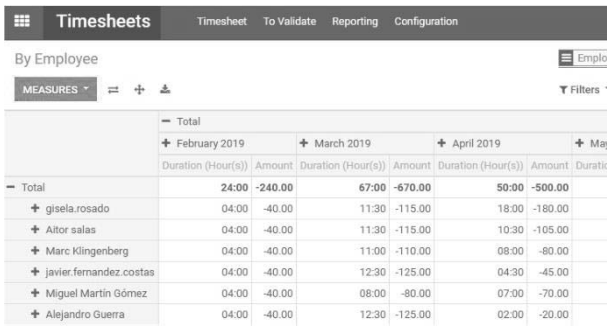
En ese momento también estamos preparados para ver quién es el potencial cliente que podría necesitar nuestros servicios para llevar a cabo esta propuesta de proyecto. Por tanto, hay una sesión en la cual se analiza el cliente potencial, y utilizando el módulo CRM se explica todo el proceso de venta. Se crea una iniciativa que se convierte en oportunidad y la hacen evolucionar por todo el proceso hasta que envían un presupuesto al cliente, que una vez negociado, éste acepta, y se crea una orden de venta, que una vez confirmada ya podemos pasar a facturar, y por último cobrar.



	Total		
	Count	Effective Hours	Planned Hours
Total	22	144.00	189.50
Proposta de projecte europeu: BIMARS	20	108.50	135.50
Completar el tercer apartado de la propuesta: Implementación	1	20.00	20.00
Invoice and sales class	1	15.00	15.00
Definición del proyecto	1	12.00	12.00
Escoger un tema para la propuesta europea	1	12.00	12.00
Comprobar pautas de corrección	1	10.50	10.50
Purchase process and websites class	1	8.50	18.00
Sales	3	4.00	4.00
Abstract, Targeted breakthrough, Long term vision and Objectives	1	3.00	3.00
Expected impacts	1	3.00	3.00
Measures to maximise impact	1	3.00	3.00
Novelty, level of ambition and foundational character	1	3.00	3.00
Relation to the work programme, Interdisciplinary nature	1	3.00	3.00
Report of Purchases class	1	3.00	

Fig. 1. Métricas sobre las tareas más costosas en tiempo. Captura proporcionada por un estudiante como entregable.

Como ejercicio se les pide que facturen mensualmente los trabajos realizados y también que realicen el proceso de cobro de las facturas.



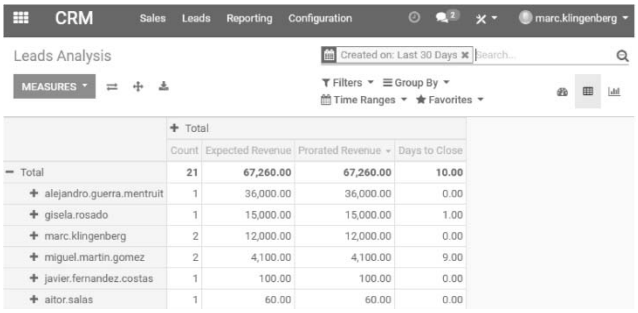
	Total					
	February 2019		March 2019		April 2019	
	Duration (Hour(s))	Amount	Duration (Hour(s))	Amount	Duration (Hour(s))	Amount
Total	24:00	-240.00	67:00	-670.00	50:00	-500.00
gisela.rosado	04:00	-40.00	11:30	-115.00	18:00	-180.00
Aitor salas	04:00	-40.00	11:30	-115.00	10:30	-105.00
Marc Klingenberg	04:00	-40.00	11:00	-110.00	08:00	-80.00
javier.fernandez.costas	04:00	-40.00	12:30	-125.00	04:30	-45.00
Miguel Martín Gómez	04:00	-40.00	08:00	-80.00	07:00	-70.00
Alejandro Guerra	04:00	-40.00	12:30	-125.00	02:00	-20.00

Fig. 2. Métricas sobre las horas trabajadas por empleado. Captura proporcionada por un estudiante como entregable.

Puesto que en la orden de venta ya aparece el ingreso para este proyecto, los estudiantes son capaces de analizar cuál va a ser el beneficio neto.

Fig. 3. Web del grupo 1 [6]

En sesiones posteriores, los estudiantes aprenden el proceso de compras mediante el módulo correspondiente. Así como analizar qué información pueden presentar en una página web, Fig 3, mediante el módulo web. Para ello han tenido que reflexionar sobre su portfolio de servicios. Hay una última sesión en la cual analizan todas las métricas generadas por el ERP, Fig 1, 2 y 4, y que permitiría la toma de decisiones clave para su éxito.



	Total			
	Count	Expected Revenue	Prorated Revenue	Days to Close
Total	21	67,260.00	67,260.00	10.00
alejandro.guerra.mentruit	1	36,000.00	36,000.00	0.00
gisela.rosado	1	15,000.00	15,000.00	1.00
marc.klingenberg	2	12,000.00	12,000.00	0.00
miguel.martin.gomez	2	4,100.00	4,100.00	9.00
javier.fernandez.costas	1	100.00	100.00	0.00
aitor.salas	1	60.00	60.00	0.00

Fig. 4 Estadísticas del CRM. Captura proporcionada por un estudiante como entregable.

C. Ejecución de un Proyecto de Ingeniería

Durante el curso, los alumnos deben definir una propuesta de proyecto que pasará por un estricto proceso de evaluación. Si la propuesta es finalmente aceptada pasará a la fase de ejecución y posteriormente se propone como desafío a los alumnos, preparar la estructura empresarial para poner en marcha un servicio o producto al mercado. El trabajo se focalizará en dos aspectos. En un primer lugar, las necesidades TIC de la compañía (p.ej un ERP) y por otro lado, las necesidades de infraestructura tecnológica para ofrecer el producto y servicio conceptualizado en la propuesta de proyecto.

Para ambas necesidades, se presenta a los alumnos la metodología Devops [8] que define los procesos dentro de una empresa para lanzar productos de forma continuada. Dicha metodología se basa en establecer canales eficaces de

comunicaciones entre los equipos de desarrollo, operaciones y control de calidad. Adicionalmente la metodología devops propone el uso de herramientas que consoliden el objetivo anteriormente descrito: el sistema de control de versiones GIT, integración continua, control de calidad, etc.

Respecto a las necesidades TIC del servicio o producto a desarrollar, cabe destacar que, aunque se deja libertad al grupo de alumnos definir un proyecto que solucione un problema de su elección, siempre se plantea como requisito que exista una cierta computación ya sea en la nube o en equipos en local. Por este motivo, se trabajan diferentes escenarios de computación en la nube y se complementa la metodología devops con el despliegue de servicios basado en contenedores [9]. Respecto a las necesidades TIC de la organización, se proporciona unas guías para el despliegue, basado también en contenedores, de las aplicaciones TIC de la organización: ERP, herramientas de gestión documental, etc.

III. Evaluación de la asignatura

Como parte de la evaluación de la asignatura se realizan un conjunto de trabajos que el profesorado solicita para asimilar los contenidos teóricos y fundamentalmente prácticos enfocados en el proyecto. Al finalizar el curso, los alumnos deben presentar una memoria (la propuesta generada) así como defender su trabajo públicamente delante de un tribunal.

Para concluir la asignatura, después de la entrega y defensa final, se pide que cada alumno valore su propia contribución y la de todos tus compañeros (en una escala de 0 a 5) de acuerdo con los criterios siguientes, añadiendo a la valoración los comentarios que estimen oportunos en un cuestionario SEEQ (Students Evaluation of Educational Quality) [7].

Los resultados de la encuesta de valoraciones de los alumnos pueden verse en el siguiente gráfico (fig 5). La muestra es la matrícula de los alumnos de tres ediciones realizadas hasta el momento (60 alumnos).

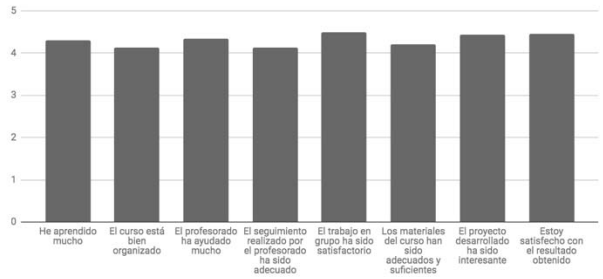


Fig. 5. Resultados de la encuesta de valoración de la asignatura de PE en el curso de primavera 2019

Por otra parte, también se solicita a los alumnos una valoración sobre la asignatura en los siguientes criterios (en una escala de 0 a 5):

- Criterio #1: Ha realizado una parte equitativa del trabajo.
- Criterio #2: Siempre ha estado disponible para ayudar a los demás.
- Criterio #3: Ha tenido siempre una actitud constructiva.
- Criterio #4: Siempre ha hecho el trabajo comprometido y lo ha hecho bien.
- Criterio #5: Sus propuestas e iniciativas han contribuido a mejorar el resultado.

IV. Conclusiones

En este artículo se ha descrito la asignatura optativa Proyectos de Ingeniería, proporcionando detalles de su heterogeneidad de estudiantes matriculados, así como su multidisciplinariedad en los temas tratados. El resultado obtenido por las encuestas de satisfacción dentro de las asignaturas, así como las que la escuela realiza cada curso muestran, en primer lugar, el alto grado de motivación que se genera en los alumnos; esta motivación se traduce en una importante dedicación como demuestran las propias encuestas y los resultados de evaluación. En segundo lugar, esta visión global no sólo técnica, sino práctica, con propuesta europea y visión de negocio, proporciona un resultado que cumple con los objetivos establecidos desde su concepción.

Finalmente, los autores tienen en mente ajustar la asignatura para poder participar en un reto real aprovechando las sinergias que proporciona el parque mediterráneo de la tecnología (PMT) que ubica la incubadora de la agencia europea espacial: ESA Business Incubation Centre Barcelona (ESA BIC Barcelona). De hecho, ya se han producido reuniones para analizar la viabilidad de esta acción.

Agradecimientos

Los autores agradecen a Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels (EETAC) y al Departamento de Ingeniería Telemática (ENTEL) por su apoyo en este trabajo.

Referencias

[1] Alcober, J., Ruiz, S., y Valero, M. (2003). Evaluación de la implantación del aprendizaje basado en proyectos en la EPSC (2001-2003). XI Congreso universitario de innovación educativa en enseñanzas técnicas.

[2] A. Ganesh, K. N. Shanil, C. Sunitha and A. M. Midhundas, "OpenERP/Odoo - An Open Source Concept to ERP Solution," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, 2016, pp. 112-116. doi: 10.1109/IACC.2016.30

[3] S. Machado, R. Messeguer, A. Oller, A. Reyes, D. Rincón, J. Yúfera, "On the impact of PBL-based teaching techniques in an optional course, on distributed applications". International Conference on Engineering and Computer Education (ICECE). Madrid, Nov. 2005. CD p. 1-6.

[4] S. Machado, R. Messeguer, A. Oller, A. Reyes, D. Rincón, J. Yúfera, "Recomendaciones para la implantación del PBL en créditos optativos basadas en la experiencia en la EPSC". XI Jornadas de Enseñanza Universitaria de la Informatica JENUI, Madrid, Julio 2005, p. 21-28. Revista CIDUI 2014 ISSN: 2385-6203 11 www.cidui.org/revistacidui

[5] S. Machado, A. Oller, E. Rodriguez, D. Rincón, J. Yúfera. "Integración de técnicas de e-learning en un bloque docente de enseñanza presencial sobre aplicaciones distribuidas". 6a. Conferencia Iberoamericana en Sistemas, Cibernética e Informática. International Institute of Informatics and Systemics, Orlando (USA) 2007, p. 200-204.

[6] A. Oller, D. Rincon, J.M. Yúfera. "Evaluación del impacto de técnicas docentes modernas en el desarrollo profesional de estudiantes de telemática". IX Jornadas de Ingeniería Telemática JITEL, Universidad de Valladolid, septiembre 2010

[7] H.W. Marsh, "SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching", British Journal of Educational Psychology, 1982, 52, 77-95.

[8] J. Davis, R. Daniels, "Effective DevOps". O'Reilly Media, 2016

[9] A. Mouat "Using Docker". O'Reilly Media, 2015


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Efficient implementation of an IoT deployment for sound-scape monitoring

Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Santiago Felici-Castell, Jaume Segura-Garcia, Rafael Fayos-Jordán, Miguel García-Pineda

Departamento de Informàtica. ETSE
Universitat de València
Avinguda de la Univèrsitat s/n 46100 Burjassot, València (SPAIN).
adolfo.pastor@uv.es, jesus.lopez-ballester@uv.es, felici@uv.es, jsegura@uv.es, rafael.fayos@uv.es, migarpi@uv.es

***Resumen*—Nowadays, the application of Wireless Acoustic Sensor Networks and Internet of Things for noise and subjective annoyance monitoring is a hot-topic in soundscape research. The problems in the evaluation of this annoyance, based on psycho-acoustic parameters, are mainly related to the efficiency of the implementation in the nodes of the network. In this work, an accurate psychoacoustic annoyance model is implemented, by computing the psychoacoustic parameters (namely loudness, sharpness, roughness and fluctuation strength) in the nodes, and later the information is gathered using an IoT system based on FIWARE. In order to implement the calculation of these parameters in real time, we have improved the code as well as we have introduced lightweight virtualization based on containers and their orchestration using a fog computing approach. The information gathered is visualised in a virtual environment using the Unity graphical engine.**

***Palabras Clave*—Dedicated networks, Psychoacoustics, Annoyance Visualization, WASN, IoT, Containers, Orchestration, fog computing**

I. INTRODUCTION

Noise is a problem in urban environments that influences on the health of citizens, ranging from children's cognition, cardiovascular diseases, insomnia, etc. to simple headaches and lack of concentration [1]. Recognizing this as a major problem, the European Commission adopted in 2002 an Environmental Noise Directive (END) 2002/49/EC [2], requiring main cities with more than 250.000 inhabitants, to gather real data on noise exposure in order to produce local action plans and to provide accurate mappings of noise pollution levels.

These measurements are mainly based on the equivalent Sound Pressure Level (SPL), also known as L_{eq}. However, L_{eq} is not enough in terms of Psycho-Acoustic Annoyance (PA) due to the fact that similar values for L_{eq} can lead to different feelings of the noise, perceived by different people, so failing to provide information related to the subjective annoyance [3] and their psychoacoustic properties. This is due to the lack of information from L_{eq} regarding the frequency characteristics. In addition, there are many sources of noise with low levels of L_{eq} that produce a disgusting annoyance and even worse than the ones with high values of L_{eq}, for instance an isolated tone from a mechanic vibration. To define metrics based on the human hearing system, different studies and techniques have been carried out and different methods have been defined in order to estimate the subjective annoyance. Nevertheless, all of them require high computational costs due to the complexity of the analysis and required signal processing. In particular one of the most commonly used and accurate is the Zwicker's model, that provides enhanced indexes, such as Loudness (L), Sharpness (S), Fluctuation Strength (F) and Roughness (R), that allow the estimation of an accurate and precise subjective PA, recently regulated by ISO 12913 [4] [5], instead of just considering L_{eq}.

In this scenario, Wireless Acoustic Sensor Networks (WASN) based on Internet of Things (IoT) are a very interesting tool, for instance to deploy a distributed sensor systems on smart cities for urban noise monitoring, allowing the identification of acoustically problematic areas and critical sound sources in real time, as well as the possibility to react efficiently against such health hazards. But with these systems, the number of samples both in time and space increase, as well as the computational complexity (as mentioned before), making the construction of these mentioned subjective maps a tough and complex task. However, WASNs fail if we want to process the previous parameters in near real time for an accurate soundscape profiling.

Thus, the goal of this paper is focused on an efficient implementation of these indexes or psycho-acoustic parameters to monitor PA and to perform real time mapping of the subjective annoyance or PA, using lightweight virtualization based on containers and their orchestration.

Finally, the amount of data generated with this kind of information systems can be dealt with Big Data systems. A useful environment to work with is RStudio, a framework based on R statistical language that allows the use of multiple statistical tools to any kind of information source. Also, for data visualization, Virtual Reality (VR) is a field that can be very applicable [6]. So, in this paper, in addition we show a VR-based system oriented to visualize PA information collected. In particular we use FIWARE framework [7], an IoT open-source platform, connected to the RStudio environment for spatial statistical data processing. The processed information is sent to the Unity graphical engine to visualize the information as a color map in a transparent plane in the environment.

The rest of the paper is structured as follows. Section II discusses the state of the art. Section III expounds the Zwicker's annoyance model. Section IV describes the implementation and architecture used. Section V exhibits the results obtained. Section VI concludes the paper.

II. RELATED WORK

Psychoacoustic research has been widely studied and several standards for evaluating subjective annoyance and calculating psychoacoustic parameters have been defined in [8] [5]. For instance, regarding PA models, in [9], the Zwicker's annoyance model [10] is used for soundscape categorization to determine how an acoustic environment sounds like, using manually collected noise samples.

Several works have considered the use of WASNs for noise monitoring. In [11] and [12], the authors have evaluated a WASN using Tmote-Sky motes [13] and Tmote Invent (TmI), to monitor traffic noise using the equivalent level, Leq,T and to count the number and type of vehicles. In this deployment, they used a sampling frequency of 8 kHz. In their study, they found that Tmote-Sky had excessive self-noise and TmI (with an integrated microphone) had apparently good audio features. In the references, the authors do not provide a specific calibration. In [14] and [15], a WASN deployment in Ostrobothnia (Finland) is discussed. In these references, the authors report different tests to evaluate the noise impact. They measured the Leq,T with T=125 ms using a sampling frequency of 33 kHz, with 14 calibrated motes (MicaZ from Crossbow - now this company is MEMSIC- with an ad hoc acquisition circuitry to allow a dynamic range of 60 dB), globally synchronized during 96 h with good results.

Other works such as [16] and [17] have used mobile phones for environmental noise monitoring. Although the results are interesting, in our opinion there is a lack of information about the recording conditions which avoids getting accurate noise measurements. When evaluating noise parameters, the location of the measuring devices should follow specific rules based on norms [2].

In terms of the traditional study of the environmental noise, the presentation of maps with SPL has been tackled in [18], where authors measured SPLs with Raspberry Pi (Rpi) and represented the information with spatial statistics, measuring 5 minutes SPL equivalent values with a WASN and also evaluated spatial cross-validation using the kriging method in a small city. Here the Rpi with a microphone is used as a sound-meter measurement system connected to a information gathering system based in RStudio. In [19], the authors also evaluated the environmental noise with a mobile application and display the results in a map using the kriging method.

In [20], authors explored the options offered by WASN for subjective annoyance computation in terms of the psycho-acoustic parameters and SPL. They found that common nodes devoted to SPL, such as TmI, could not afford the psycho-acoustic parameter computation and the audio sampling was quite poor for the audio reconstruction, from recorded audio chunks. Nevertheless, the tendency for the relationship between the psycho-acoustic parameters L, R, S and F) computed from the audio signal gathered with these TmI had good correlation with the SPL in terms of subjective annoyance.

In [21], the authors implemented a edge-computing system by using different Rpi nodes in order to perform an evaluation of performance when computing the binaural loudness directly on the Rpi nodes. An improvement of this work was done in [22], where authors added the computation of binaural sharpness to the set of binaural psycho-acoustic parameters. Also, a simplified version of the Zwicker's psycho-acoustic annoyance model was evaluated (assuming specific conditions), oriented to assess the spatial distribution of the subjective nuisance in indoor and outdoor environments.

Although big efforts have been done, we cannot find many complete solutions and available alternatives for sound scape monitoring. Thus this paper will contribute to improve it.

III. SOUND SCAPE MONITORING AND PYSCHO ACOUSTIC PARAMETERS

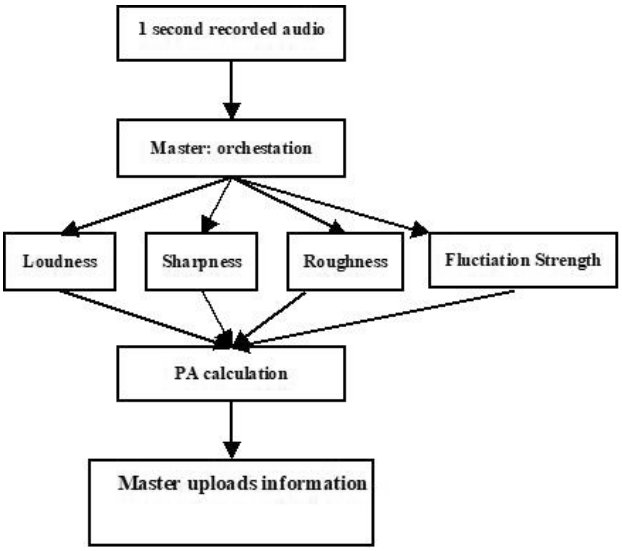


Fig. 1. Flow chart to calculate Psycho-acoustic Annoyance (PA)

Psycho-acoustic metrics are an alternative to express people's feelings by subjective measures. In this section,

we describe how the Zwicker's annoyance model [10] is implemented in WASN for a general purposes by measuring PA, based on L, S, F and R. In other words, PA is defined based on L, S, F and R and we need all of them before calculating PA. This is shown in Figure 1. Due to the complexity of their implementation, we cannot perform this monitoring process with conventional nodes (such as Rpi) and less in real time. We must highlight that we refer to real time when the time required to process an audio chunk is shorter that the chunk itself, by default 1 second. Similar deployments related with this topic but with a different approach and using different techniques can be found in [20] and [21]. A detailed explanation of the algorithms can be found in [23].

The computational cost of these parameters are shown in Table I. Also, this table shows a performance comparison in terms of computational time per each second of audio recording for the different psycho-acoustic parameters, between Matlab and C++/Python running on the computer and the C++/Python implementation running on RPi models 3B and 3B+ [24]. It must be noticed that with RPi, the main program is based on Python and we use C++ in order to implement a Python library that performs all the tough processing from each psycho-acoustic parameter, in an efficient way by using the linear algebra library called Armadillo [25]. This code has not been programmed using threads. The best performance computing times are shown obviously by the computer, first the C++/Python implementation, followed by the Matlab implementation, both in real time. However in the RPis, we cannot achieve real time. We mean real time when one second of recorded audio is processed in less than one second. In order to obtain the most realistic results, 100 random samples of daily sounds of one second of duration have been used.

Tabla I
TIME COMPARISON (IN SECONDS) BETWEEN DIFFERENT DEVICES AND PROGRAMMING LANGUAGES

	L	S	R & F	R	F	Total
Matlab	0.058	0.000	0.638	0.288	0.404	0.699
C++/Python	0.003	0.000	0.235	0.128	0.235	0.238
RPi3B	0.018	0.000	1.462	0.849	0.742	1.479
RPi3B+	0.017	0.000	1.389	0.794	0.694	1.406

Thus, as we can see from Table I in Rpi, these parameters take a long time to be calculated and our proposal is to rely on a fog computing architecture, using virtualization with the acoustic parameters embedded in Linux containers to distribute them easily in the fog.

IV. PROPOSED ARCHITECTURE AND METHODS

The proposed architecture is based on a WASN, composed by Rpi model 3 B+ [24] as shown in Figure 2. The nodes record 1 second of audio and will send it to the master of a cluster of RPi (working in a fog computing scheme). The cluster is based on several RPis and the master is in charge of distributing the workload. Later, the master sends the information to the FIWARE framework, using MQTT protocol, and it is stored in MongoDB, a no-SQL database. The psycho-acoustic information is then used by the RStudio plugin, a statistical framework which was configured to be connected to the FIWARE framework. A statistical script using kriging technique (explained later in this section) process the information as a matrix with the information of the psycho-acoustic annoyance according to the Zwicker's annoyance model. This information matrix is sent to the C# script integrated in the model of the environment in Unity graphical engine via a JSON query. The C# scripts also represents this nuisance matrix as a color map in a transparent surface that can be navigated inside the virtual model of the environment.

A. SENSOR NODE

Figure 3 shows the detail of a sensor node (end user equipment) using Rpi, running Linux containers (based on Docker [26]) to measure the different Acoustic Parameters (AP). Each parameter (L, S, F and R) is embedded in different containers as well as SPL and tonality. It is worth mentioning that we have several options for the fog computing scheme and to orchestrate the resources in the cluster. Some orchestrators are widely used for container-based cluster management, such as Swarm [27] and Kubernetes [28]. While Swarm is the native orchestrator of Docker containers, Kubernetes from Google is considered as the most feature-rich orchestrator. From simplicity and because we are using RPi, we will use Swarm as orchestrator. This orchestrator acting as a master in the cluster will distribute the load (audio chunks) among the slaves in order to calculate PA as explained before. In addition, these nodes within the cluster build a mesh network using their wireless capabilities.

The technical characteristics of the Rpi model 3 include a 1.4 GHz 64-bit quad-core ARM Cortex-A53 CPU, 1 GB RAM, 40 GPIO pins, 4 USB ports, a full HDMI port, an 10/100/1000MB Ethernet port and integrated 802.11ac/n Wireless LAN, and also Bluetooth 4.2 Low Energy (BLE). Also, the USB ports and the GPIO pins are a good solution, providing the Rpi with the possibility to have a range of peripherals available, such as WiFi antennas, ZigBee modules, microphones, cameras and connections with other devices, e.g., Arduino.

B. FIWARE IOT COLLECTION FRAMEWORK

FIWARE [7] is an open-source platform, developed by Telefónica I+D and driven by the European Union (in different EU projects), for the development and global deployment of Future Internet applications. FIWARE aims to provide a fully open, public and free architecture as well as a set of specifications that enable developers, service providers, enterprises and other organisations to develop products that meet their needs, while remaining open and innovative.

The FIWARE IoT architecture deployment of the Service Enablement is usually distributed between a number of devices, several gateways and the backend. As defined, a device is a hardware entity, component or system that

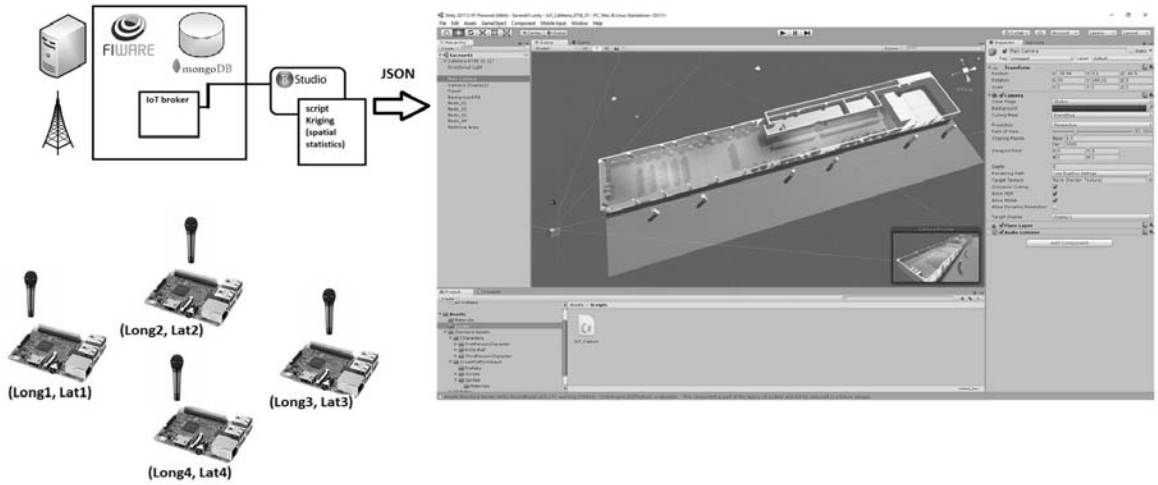


Fig. 2. Schema of the whole system

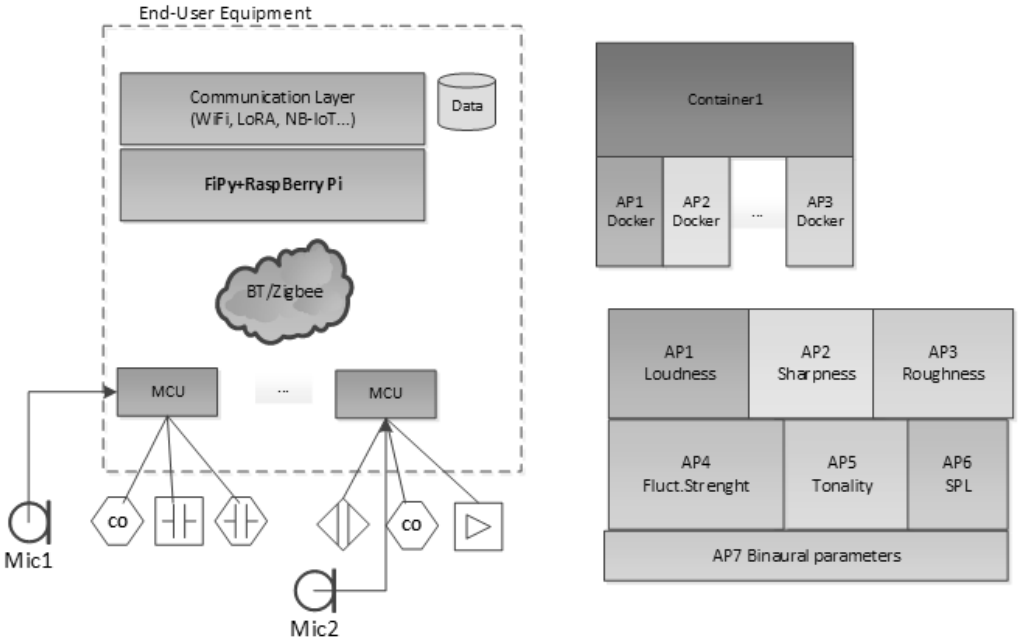


Fig. 3. Sensor node with containers to measure the Acoustic Parameters (AP)

measures properties of a thing/group of things or influences the properties of a thing/group of things or both measures/influences. Sensors and actuators are devices.

IoT Resources are computational elements (software) that provide the technical means to perform sensing and/or actuation on the device. The resource is usually hosted on the device. IoT Generic Enablers (GEs) have been spread in two different domains: Gateway and Backend. While IoT Gateway GE provide inter-networking and protocol conversion functionalities between devices and the IoT Backend GEs, the IoT Backend GEs provide management functionalities for the devices and IoT domain-specific support for the applications.

The use of FIWARE allows to have an autonomous and off-the-shelf IoT service for complex applications. The Fiware IoT architecture is shown in Figure 4.

C. SPATIAL STATISTICS

FIWARE allows the data connection to RStudio plugin. In this statistical framework, the spatial statistical processing can be done. The most common methods in spatial statistics are Inverse Distance Weighted (IDW), spline and kriging. IDW is a simple and intuitive deterministic interpolation method based on the principle that sample values closer to the prediction location have more influence on prediction value than sample values farther apart. The major disadvantage of IDW is "bull's eye" effect and edgy surface. Spline is also a deterministic interpolation method which fits mathematical function through input data to create smooth surface. Kriging is a method based on spatial autocorrelation [29].

The computed PA from the measurements establish a data set considering the locations in terms of GPS coor-

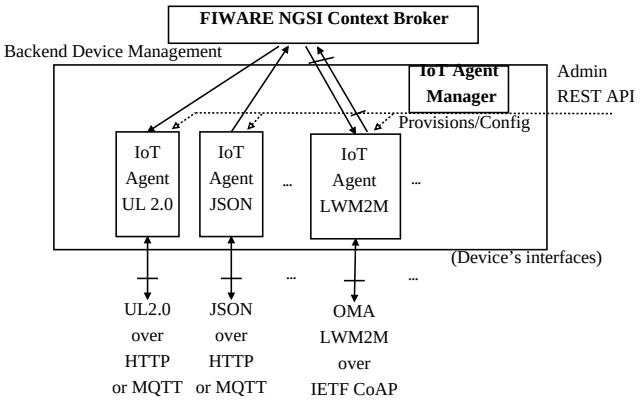


Fig. 4. FIWARE IoT architecture

dinates. By denoting the PA computed with the Zwicker's model at a location x as $Y(x)$, this data set is defined as $\{Y(x),\, x \in \mathcal{D}\}$, where $\mathcal{D}$ are all the locations of the modelling sets, following the kriging technique [30].

In this context, the objective of this proposed model is the prediction of Y(x0) in any location x0, particularly those within the validation set. The annoyance reports contain information of the set of covariables included. Therefore, Y(x) is modeled as a tendency function of the covariables better involved in the process which explains its variability in a large extent plus some random error which is explained by the short term variability, as follows:

$$Y(x) = \mu(x) + \delta(x), \tag{1}$$

where $\mu(x) = E[Y(x)]$ and $\delta(x)$ is a stationary Gaussian process with zero mean, whose spatial dependence characterization is given by the variogram γ [31]:

$$2\gamma(h) = \mathrm{Var}\left[Y(x+h) - Y(x)\right] = \mathrm{Var}\left[\delta(x+h) - \delta(x)\right], \tag{2}$$

where Var denotes the variance and h is an offset. This variogram represents the main function of the kriging method, which presents different procedures such as simple kriging, ordinary kriging, universal kriging, indicator kriging, co-kriging, etc, attending to different statistical aspects considered in the covariable set. Ordinary kriging is the most widely used kriging method. It is used to estimate a value at a point of a region for which a variogram is known, using data in the neighborhood of the estimation location and also can be used to estimate a block value [32].

D. DATA VISUALIZATION

Unity [33] is a cross-platform graphical engine used as a game engine. This tool allow users the ability to create games in 2D and 3D. The engine offers a primary scripting API in C#, for both the Unity editor in the form of plugins, as well as drag and drop functionality. The last release is Unity 2018.

This platform also allows connection with other programs and provides different Software Developer Kits (SDKs), as XR SDK. XR is an acronym that encompassess VR, Augmented Reality (AR), and Mixed Reality (MR). XR SDK allows the use of virtual reality devices directly from Unity, without any external plug-ins in projects. It provides a base API and feature set with compatibility for multiple devices.

V. RESULTS AND DISCUSSION

In this study, we have used the Unity platform to show a model of a premise in our School of Engineering (ETSE) at University of Valencia. We have programmed a C# script to pass the kriging processed matrix computed from the PA information, collected, computed, averaged each 2 minutes by each one of the 6 nodes (one master and 5 slaves in the cluster) in the cafeteria, shown in Figure 5.

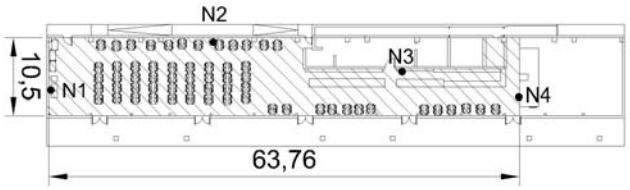


Fig. 5. Map in meters (ETSE cafeteria) with 4 nodes

The PA information, processed by RStudio, is passed to the Unity platform, as a kriging matrix, through a JSON object. This object is collected by the C# script and processed in order to show a color map of the psycho-acoustic annoyance in a transparent plane inside the premise model. Figure 6 shows a perspective view of the Cafeteria model with the color map representing the PA in the environment. In this model, we have configured a first person navigator in order to visualize the annoyance information in each part of the room using a VR Head Mounted Display (HMD) configured with XR SDK.

The graphical rendering of the PA information has been tested and provides good performance to know the nuisance in this environment.

VI. CONCLUSIONS

This paper explains an IoT system for psycho-acoustic annoyance computation and visualization. In order to implement the calculation of the complex acoustic parameters in real time for soundscape monitoring, we have improved their code to be run as well as we have introduced lightweight virtualization based on Linux containers and their orchestration using Docker Swarm. Finally, the visualization is made by using a Unity platform. A C# script has been programmed to receive the kringing matrix from the RStudio plugin, as a JSON object, and to represent it over the model. The navigation is made with a first person controller using a VR viewer.

The navigation has shown its effectiveness in the visualization of the psycho-acoustic annoyance in this environment. Future work will be done for a 3D visualization of the nuisance as a color mist.

ACKNOWLEDGEMENTS

This work has been funded by the Ministry of Innovation and Economy under the project Urbauramon BIA2016-76957-C3-1-R.

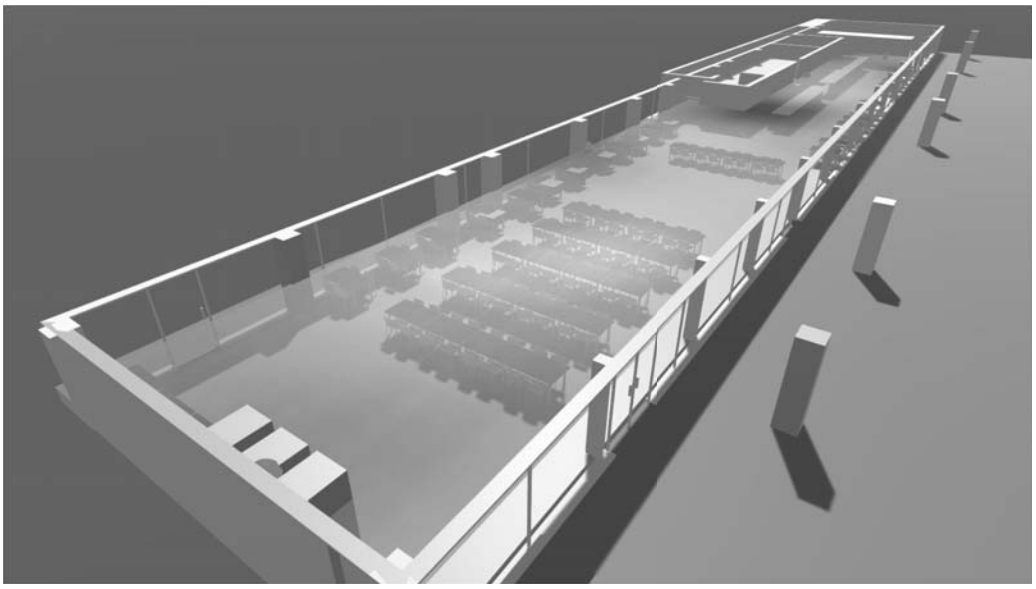


Fig. 6. View of the PA model shown as a color map

REFERENCIAS

[1] K. Levak, M. Horvat, and H. Domitrovic, "Effects of noise on humans," in *2008 50th International Symposium ELMAR*, vol. 1, Sep. 2008, pp. 333–336.

[2] END, "Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the Assessment and Management of Environmental Noise," July 2002.

[3] M. Raimbault and D. Dubois, "Urban soundscapes: Experiences and knowledge," *Cities*, vol. 22, pp. 339–350, 10 2005.

[4] ISO, "Acoustics—Soundscape—Part 2: Data collection and reporting requirements ISO 12913-2:2018," Geneva, Switzerland, 2 2018.

[5] ——, "Acoustics—Soundscape—Part 1: Definition and Conceptual Framework; ISO 12913-1:2014," Geneva, Switzerland, 1 2014.

[6] E. Olshannikova, A. Ometov, Y. Koucheryavy, and T. Olsson, "Visualizing big data with augmented and virtual reality: challenges and research agenda," *Journal of Big Data*, vol. 2, no. 1, p. 22, Oct 2015. [Online]. Available: https://doi.org/10.1186/s40537-015-0031-2

[7] "FIWARE IoT Platform," 2019, accessed: 02/05/2019. [Online]. Available: https://forge.fiware.org/

[8] ISO, "Acoustics—Assessment of Noise Annoyance by Means of Social and Socio-Acoustic Surveys; ISO TS 15666:2003," Geneva, Switzerland, 1 2003.

[9] M. Rychtáriková and G. Vermeir, "Soundscape categorization on the basis of objective acoustical parameters," *Applied Acoustics*, vol. 74, no. 2, pp. 240–247, Feb. 2013.

[10] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, ser. Springer series in information sciences. Springer, 2007.

[11] S. Santini and A. Vitaletti, "Wireless sensor networks for environmental noise monitoring." in *Proceedings of the 6th GI/ITG KuVS Workshop Wireless Sensor Networks, Aachen, Germany*, Jul 16–17, 2007, pp. 98–101.

[12] S. Santini, B. Ostermaier, and A. Vitaletti, "First experiences using wireless sensor networks for noise pollution monitoring." in *Proceedings of the 3rd ACM Workshop Real-World Wireless Sensor Network (REALWSN), Glasgow, Scotland, UK*, Apr 1–4, 2008, pp. 61–65.

[13] J. Polastre, R. Szewczyk, and D. Culler, "Telos: Enabling ultra-low power wireless research." in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks (IPSN), Los Angeles, CA, USA*, April 24–27, 2005, pp. 364–369.

[14] I. Hakala, I. Kivela, J. Ihalainen, J. Luomala, and C. Gao, "Design of low-cost noise measurement sensor network: Sensor function design," in *Proceedings of the IEEE 1st International Conference on Sensor Device Technologies and Applications, Venice, Italy*, Jul 18–25, 2010, pp. 172–179.

[15] I. Kivela, C. Gao, J. Luomala, and I. Hakala, "Design of noise measurement sensor network: Networking and communication part." in *Proceedings of the 5th International Conference on Sensor Technologies and Applications, Côte d'Azur, France*, Aug 21–27, 2011, pp. 280–287.

[16] N. Maisonneuve, M. Stevens, M. Niessen, P. Hanappe, and L. Steels, "Citizen noise pollution monitoring," in *Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government. Digital Government Society of North America, Puebla, Mexico*, May 17–21, 2009, pp. 96–103.

[17] N. Maisonneuve, M. Stevens, M. Niessen, and L. Steels, "Noisetube: Measuring and mapping noise pollution with mobile phones." in *Information Technologies in Environmental Engineering (Environmental Science and Engineering)*, I. Athanasiadis, A. Rizzoli, P. Mitkas, and J. Gómez, Eds. Springer: Verlag/Berlin, Germany, 2009, pp. 215–228.

[18] J. Segura Garcia, J. J. Pérez Solano, M. Cobos Serrano, E. A. Navarro Camba, S. Felici Castell, A. Soriano Asensi, and F. Montes Suay, "Spatial statistical analysis of urban noise data from a wasn gathered by an iot system: Application to a small city," *Applied Sciences*, vol. 6, no. 12, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/12/380

[19] J. Zuo, H. Xia, S. Liu, and Y. Qiao, "Mapping urban environmental noise using smartphones," *Sensors (Basel, Switzerland)*, vol. 16, 2016.

[20] J. Segura-Garcia, S. Felici-Castell, J. J. Perez-Solano, M. Cobos-Serrano, and J. M. Navarro, "Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks," *IEEE Sensors Journal*, vol. 15, pp. 836–844.

[21] J. E. Noriega-Linares, A. Rodriguez-Mayol, M. Cobos-Serrano, J. Segura-Garcia, F.-C. S., and J. M. Navarro, "A wireless acoustic array system for binaural loudness evaluation in cities," *IEEE Sensors Journal*, vol. 17, pp. 7043–7052, 2017.

[22] J. Segura-Garcia, J. M. Navarro-Ruiz, J. J. Perez-Solano, J. Montoya-Belmonte, S. Felici-Castell, M. Cobos-Serrano, and A. M. Torres-Aranda, "Spatio-temporal analysis of urban acoustic environments with binaural psycho-acoustical considerations for iot-based applications," *Sensors*, vol. 18, no. 3, 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/3/690

[23] A. Pastor-Aparicio, J. Lopez-Ballester, J. Segura-Garcia, S. Felici-Castell, M. Cobos, R. Fayos-Jordan, and J. Perez-Solano, "Zwicker's annoyance model implementation in a WASN node," in *Inter Noise 2019*, vol. 1, June. 2019, pp. 1–1.

[24] "Raspberry Pi 3B+," 2018, accessed: 02/01/2019. [Online]. Available: https://www.raspberrypi.org/

[25] C. Sanderson and R. Curtin, "A user-friendly hybrid sparse matrix class in c++," *Lecture Notes in Computer Science (LNCS)*, vol. 10931, no. 1, pp. 422–430, 2018.

[26] "Docker Containers," 2018, accessed: 02/03/2019. [Online]. Available: https://docs.docker.com/engine/examples/

[27] "Docker Swarm," 2017, accessed: 11/11/2018. [Online]. Available: https://docs.docker.com/swarm/
[28] "Kubernetes (K8s)," 2018, accessed: 11/10/2018. [Online]. Available: https://kubernetes.io/
[29] E. Isaaks and R. Srivastava, *An Introduction to Applied Geostatistics*. Oxford University Press: New York, NY, USA, 1989.
[30] N. Cressie, *Statistics for Spatial Data*. John Wiley: New York, NY, USA, 1993.
[31] A. Diblasi and A. Bowman, "On the use of the variogram in checking for independence in spatial data." *Biometrics*, vol. 57, pp. 211–218, 2001.
[32] H. Wackernagel, *Ordinary kriging*. Springer: Berlin, Germany, 2003, pp. 79–88.
[33] "Unity," 2019, accessed: 02/05/2019. [Online]. Available: https://unity3d.com

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Influencia del ciclo de trabajo de los semáforos en una intersección simple en múltiples parámetros ante una densidad de tráfico incremental

Antonio Guillen-Perez, Maria-Dolores Cano
Departamento Tecnologías de la Información y las Comunicaciones
Universidad Politécnica de Cartagena
Campus Muralla del Mar, Edif. Antigones, 30202 Cartagena. Murcia. España.
antonio.guillen@edu.upct.es, mdolores.cano@upct.es

***Resumen*- Los sistemas inteligentes de transporte que controlan intersecciones reguladas por semáforos ofrecen la posibilidad de adaptar el ciclo de trabajo para mejorar el tráfico. Esta mejora se puede hacer en base a multitud de parámetros: tiempo de espera, velocidad media de la vía, disminución de la congestión, etc. Recientemente, estos sistemas están en auge gracias al avance en las Smart Cities y la Internet of Things y muestra de ello es el incremento de contribuciones científicas que abordan esta temática. Una de las metodologías más empleadas para evaluar aportaciones en este ámbito es la simulación por computador, por ejemplo, con un simulador open-source como SUMO. No obstante, la versatilidad que ofrecen estas herramientas se convierte en un arma de doble filo, ya que, si bien permiten modificar un elevado número de parámetros en el sistema, no siempre se configuran con el suficiente rigor. En este trabajo, mostramos las bases de funcionamiento de los sistemas inteligentes de control de tráfico, corroborando las asunciones que se realizan y demostrando que han de tenerse en cuenta determinados límites a la hora de evaluar las prestaciones de una propuesta.**

***Palabras Clave*- control de tráfico, sistemas inteligentes de transporte, Internet of Things, Smart Cities.**

I. INTRODUCCIÓN

Como decía Benjamin Franklin: «Recuerda que el tiempo es dinero», y cada año se pierden en España más de 840 millones de euros debido a los atascos de tráfico, equivalente a 14 días laborales perdidos [1]. Además, si nos fijamos en los gastos indirectos producidos por esos atascos podemos observar que España gasta más de 3.500 millones de euros al año para el tratamiento de enfermedades causadas por la contaminación del tráfico y los atascos [2]. Si bien los atascos se pueden originar de diversas maneras, una de las causas principales es la congestión del tráfico en intersecciones y accesos a ciudades por una gestión ineficiente, es decir, una mala circulación en intersecciones que no están debidamente controladas. Estas intersecciones se pueden regular de diversas formas, pero las principales suelen ser control por semáforos o rotondas. Los primeros presentan una gran ventaja frente a las rotondas cuando existe un pico de flujo de tráfico, ya que el tráfico no pierde la prioridad en los tramos que acceden a la rotonda, y, por consiguiente, no se pierde la jerarquía viaria. Además, las rotondas imponen demoras a todos los usuarios y presentan múltiples problemas relacionados con los vehículos de 2 ruedas (interviniendo en la mitad de los accidentes con víctimas) y los peatones [3].

Utilizar de manera óptima las intersecciones permitiría controlar el tráfico y regular el flujo de vehículos, reduciendo así al mínimo el tiempo de espera, maximizando el flujo de vehículos, reduciendo las emisiones e incluso los accidentes. Sin embargo, esta utilización óptima no es evidente. Los sistemas inteligentes de transporte (*Intelligent Transportation Systems*, ITS) aparecieron con el objetivo de solventar este problema. Los ITS controlan de manera adaptativa las intersecciones en función de diversos parámetros de entrada (como la utilización por carril en tiempo real, la calidad del aire, el tiempo de espera, etc.) con el fin de mejorar el uso de la intersección, teniendo como fundamento principal aumentar el ciclo de los semáforos de una intersección cuando el sistema encuentra un aumento en la densidad de tráfico entrante[4].

Debido a la gran importancia que tienen estos sistemas hoy en día, y a su auge en los próximos años, en este artículo se analizará el fundamento base de estos ITS dedicados al control de tráfico urbano. Es decir, el efecto que tiene la duración del

ciclo de los semáforos para una intersección en función de un rango de flujo de tráfico de entrada incremental para diversas variables objetivo como son el tiempo medio de espera, la velocidad media, el consumo de combustible y las emisiones atmosféricas (todas ellas en relación con los vehículos), con el fin de obtener la duración óptima para cada uno de los flujos vehiculares de entrada y corroborar el funcionamiento base de estos sistemas. Usaremos como caso de estudio una intersección regulada aislada formada por cuatro ramas, con dos carriles de tráfico cada una. La herramienta para la evaluación de prestaciones es SUMO [5], un simulador de tráfico microscópico, ampliamente estudiado y desarrollado por la comunidad científica y utilizado en el estudio de diversos campos como las redes vehiculares, estudio de flujos de tráfico en grandes ciudades y el desarrollo de nuevos ITS. Con este caso de estudio demostraremos que, antes de que la intersección se encuentre en una región de sobresaturación (en el escenario seleccionado para este trabajo viene definido por un flujo vehicular menor de 800 veh/h), existe una región donde los parámetros estudiados convergen en un valor óptimo entre el rango de duraciones de ciclos estudiados. Es interesante señalar que son numerosos los trabajos científicos basados en simulaciones por ordenador que no tienen en cuenta ese valor de sobresaturación o no lo indican, lo que podría comprometer la validez de algunos de los escenarios de tráfico evaluados.

El artículo se organiza como sigue. En la sección II presentaremos una breve introducción al estado del arte en ITS. En la sección III se describen los detalles del caso de estudio que se ha realizado. Los resultados obtenidos se detallan en la sección IV. Finalmente, las conclusiones extraídas se concretan en la sección V.

II. Estado del arte

Existen un gran número de ITS capaces de controlar el tráfico de manera adaptativa [6], dentro de los cuales se pueden clasificar por niveles de capacidad desde los sistemas más básicos que tan solo controlan el *timing* en divisiones temporales [7], [8], pasando por sistemas ampliamente implementados como son SCATS [9] y SCOOT [10], que ajustan los parámetros del esquema de *timing* de la señal (período de la señal, relación de señal verde/rojo y diferencia de fase), hasta sistemas aislados totalmente capaces de autoaprender las características del entorno y que poseen una alta eficiencia computacional [11]–[14]. Existen diversas formas de recolectar información del tráfico, dentro de las cuales destacan el *induction loop detector*, cámaras, etc.; pero gracias a los avances en vehículos autónomos, redes vehiculares (*Vehicle to Vehicle*, V2V), las redes vehículo a infraestructura (*Vehicle to Infrastructure*, V2I) y el auge de la Internet de las Cosas (*Internet of Things*, IoT) y las ciudades inteligentes (*Smart Cities*), nuevas formas de captura de datos están siendo posibles, permitiendo la interconexión entre los vehículos y los sistemas de control de tráfico [15], [16].

El fundamento principal de estos sistemas es aumentar el tiempo de ciclo cuando se detecta un aumento de la densidad de vehículos entrantes [17]. Nótese que el tiempo de ciclo es la suma de todas las fases por las que pasa un semáforo (tiempo en rojo, tiempo en verde, tiempo en ámbar y tiempo en rojo de vaciado de la intersección). Sin embargo, la afirmación anterior sólo se puede admitir siempre y cuando la intersección no esté sobresaturada, es decir, que el número de vehículos que entra a una intersección pueda ser manejado por la misma, ya que al sobrepasar la densidad crítica, el comportamiento de los vehículos y la intersección es impredecible [4]. La Fig. 1 muestra la relación entre la densidad de vehículos (*rate of traffic over distance*) y cuyas unidades son veh/m) y el flujo de vehículos (veh/h).

En la literatura científica, podemos encontrar numerosos trabajos que abordan el control de tráfico inteligente en intersecciones reguladas (es decir, que hacen uso de semáforos como herramienta de control) y que emplean simuladores de tráfico como SUMO. Cuando en este simulador se usan valores de flujo o de densidad de vehículos que están por encima del valor de saturación, la documentación [18] indica que el comportamiento del simulador es el siguiente. Cuando una intersección satura y los brazos que entran a la intersección se llenan de vehículos, los nuevos vehículos que se desean insertar en la intersección no tienen espacio físico para ser insertados y por tanto éstos pasan a almacenarse en una cola de memoria del software para su posterior inserción. En este punto, SUMO puede seguir diversas opciones para insertar estos vehículos que no han cumplido las condiciones de inserción en su momento. En general, lo que hace es que realmente no se esté simulando el flujo de vehículos correcto, ya que, si por ejemplo se está corriendo una simulación con el parámetro establecido de 1000 veh/h en una simulación de 1 h y la intersección simulada sólo puede manejar 500 veh/h, lo que ocurrirá es que la simulación realmente durará 2 horas, aunque en la configuración de la simulación sólo se haya indicado que la duración debía ser de 1 hora con un flujo de 1000 veh/h. La primera hora de simulación SUMO intentará introducir los 1000 vehículos, pero sólo podrá introducir 500, y los 500 restantes que no han podido introducirse se irán introduciendo conforme el simulador pueda, aumentando la duración de la simulación a las 2 horas (es decir que en verdad ha trabajado con un flujo de 500 veh/h). El principal problema radica en que el simulador en sí no genera ningún tipo de aviso o alarma ante esta situación pudiendo justificar el aumento en la duración de las simulaciones al tipo de hardware o software del dispositivo sobre el que corre la simulación. Por lo tanto, pudiendo llevar a resultados erróneos o no rigurosos. Una simple búsqueda en la literatura científica especializada muestra que mientras muchos trabajos mencionan este valor de saturación [19], efectivamente existen trabajos en los que o bien no se ha tenido en cuenta dicho límite de saturación (para corroborar que las simulaciones se encuentran dentro del rango de valores apropiado) o no se menciona [20]–[23].

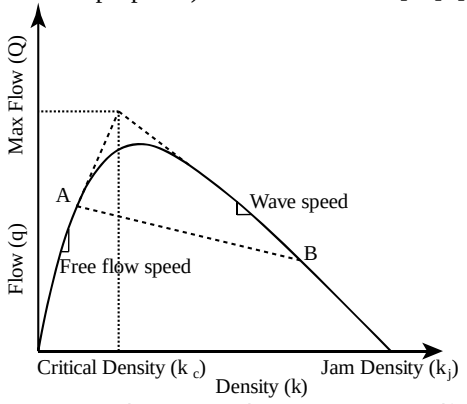


Fig. 1. Representación de la relación entre densidad de tráfico y flujo.

III. Diseño experimental

Para el caso de estudio, realizaremos simulaciones con SUMO [5] en el escenario mostrado en la Fig. 2. En cada brazo de la intersección se define un flujo de entrada por carril que variará entre 300 veh/h y 1000 veh/h. Por simplicidad, sólo se permite a los vehículos atravesar en línea recta la intersección o girar a la derecha. La Tabla I muestra tanto los parámetros de entrada del simulador como los datos de salida que serán analizados. El parque de vehículos simulados se basó en los datos de parque de vehículos ofrecido por la Dirección General de Tráfico para Madrid (España). Respecto a los parámetros de contaminación atmosférica y consumo de combustible de los vehículos, se utilizó la normativa europea de emisiones contaminantes en vehículos EURO 5. Los parámetros escogidos para cada uno de los tipos de vehículos por peso y combustible, así como la distribución de vehículos se pueden ver en la Tabla II.

IV. Resultados

Una vez llevadas a cabo las simulaciones anteriores, en este apartado veremos los resultados obtenidos. Recordemos que la evaluación de prestaciones se realiza midiendo las siguientes métricas del conjunto de vehículos: tiempo de espera medio en la intersección, duración media de viaje, velocidad media, emisiones de CO y consumo medio.

En este primer grupo de soluciones podemos ver por separado el comportamiento de cada una de las métricas estudiadas para todo el rango de tiempos en verde simulados (tiempos en verde entre 15 s y 100 s) y para todos los flujos vehiculares simulados (de 300 veh/h a 1000 veh/h) por carril.

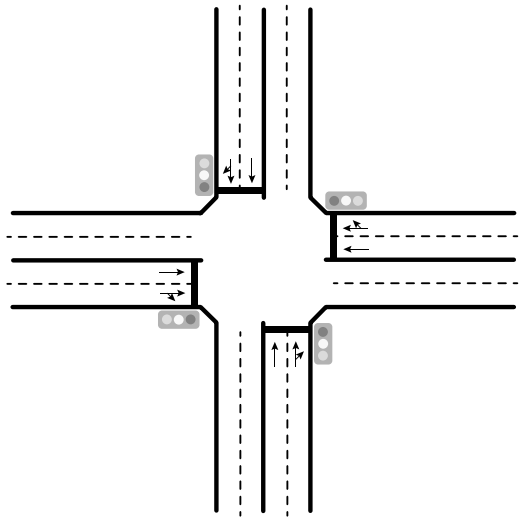


Fig. 2. Representación de la intersección simulada (longitud de cada brazo de la intersección 200 m).

En las Fig. 3-7 podemos ver que el comportamiento en términos de métricas evaluadas es diferente en función del flujo de tráfico, pero se puede apreciar que para flujos pequeños (menores de 500 veh/h) y flujos muy grandes (mayores de 850 veh/h) apenas importa el tiempo en verde escogido y las métricas de prestaciones se mantienen. Sin embargo, para el rango de flujos entre 500 veh/h y 850 veh/h por carril existe una gran diferencia entre unos valores y otros de tiempos en verde.

Tabla I
Parámetros estudiados

Entrada/ Salida	Parámetros	Rango
Entrada	Flujo vehicular por carril (simétrico)[1]	{300-1000} veh/h
	Tiempo ámbar	2 s
	Tiempo rojo de vaciado	5 s
	Tiempo rojo	(Tiempo verde)
	Tiempo verde	15-100 s
	Ciclo	(Verde+Rojo+Ámbar+Vaciado) = {37 – 207} s
Salida	Duración de viaje	
	Tiempo de espera	
	Velocidad media	
	CO emitido	
	Consumo de combustible	

[1]Los flujos vehiculares simulados son simétricos, iguales en todos los carriles de todas las ramas de la intersección. El flujo vehicular que entra en cada rama de la intersección es dos veces el flujo indicado, al tener 2 carriles por sentido.

Tabla II
Distribución de vehículos y tipo de combustible utilizado

Vehículo	Porcentaje	Combustible
Coche	30%	Gasolina
Coche	40%	Diésel
Motocicleta	10%	Gasolina
Ciclomotor	10%	Gasolina
Furgoneta	5%	Diésel
Autobús	5%	Media de todos los tipos de combustible

Por ejemplo, podemos observar en la Fig. 3 que aumentar el tiempo en verde para disminuir el tiempo medio de viaje beneficia (en mayor o menor medida) a flujos de tráfico iguales o superiores a 600 veh/h por carril. Sin embargo, en la Fig. 4 se muestra que ese mismo aumento del tiempo en verde perjudica en términos de tiempo medio de espera a flujos por encima de 800 veh/h por carril. Por otro lado, mientras que la velocidad media (Fig. 5) sigue un comportamiento similar a la métrica de tiempo medio de viaje, las emisiones (Fig. 6) y el consumo de combustible (Fig. 7) siguen el patrón del tiempo medio de espera (es decir, limitando la eficiencia a flujos que estén aproximadamente por debajo de los 800 veh/h por carril).

Finalmente, en la Fig 8 podemos ver un resumen de lo obtenido en las gráficas anteriores. Esta gráfica sólo nos muestra el valor de tiempo en verde que optimiza la variable objetivo estudiada (duración de viaje, tiempo de espera, velocidad media, CO emitido, consumo combustible) para el rango de flujo vehicular estudiado. De esta figura se puede desprender que existen dos regiones de flujo bien diferenciadas. Para la región donde el flujo vehicular es menor de 800 veh/h por carril el comportamiento del tiempo óptimo es lo esperado, a mayor flujo vehicular es necesario aumentar el tiempo en verde (aumentar el ciclo) de los semáforos que controlan la intersección con el fin de optimizar las variables objetivo. Pero si dicho flujo vehicular es superior a 800 veh/h por carril, el comportamiento de las variables objetivo ya no es predecible, algunas mejoran al disminuir el tiempo en verde (tiempo de espera, emisiones de CO y consumo de combustible) y otras lo hacen al aumentar (duración de viaje y velocidad media).

V. Conclusiones

Los sistemas inteligentes de transporte están llamados a revolucionar las ciudades del futuro, en gran medida gracias a

la incorporación de las nuevas tecnologías en el ámbito de las telecomunicaciones. En este trabajo, se muestra cómo es posible alterar las prestaciones del tráfico de vehículos en intersecciones reguladas mediante la modificación de los ciclos de tráfico, poniendo un énfasis especial en la necesidad de acotar la región donde realmente existe un beneficio y la importancia de identificar esa zona en los trabajos basados en simulaciones por ordenador para evitar resultados que pudieran ser científicamente no rigurosos. Como trabajo futuro, se propondrá un sistema inteligente de control de tráfico para zonas urbanas teniendo en cuenta estos resultados.

AGRADECIMIENTOS

This work was supported by the AEI/FEDER-UE project grant TEC2016-76465.C2-1-R (AIM) and by the Dirección General de Tráfico, Ministerio del Interior (Spain), project grant SPIP2017-02230 (STREET).

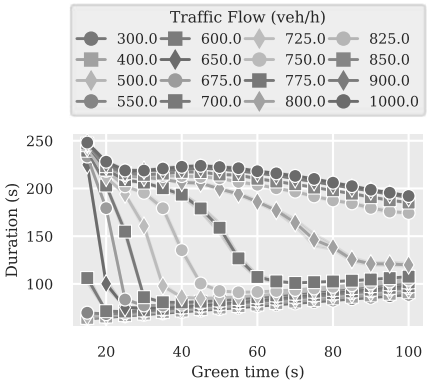


Fig. 3. Duración de viaje (s) frente al tiempo en verde (s) para diferentes valores de flujo de tráfico (veh/h).

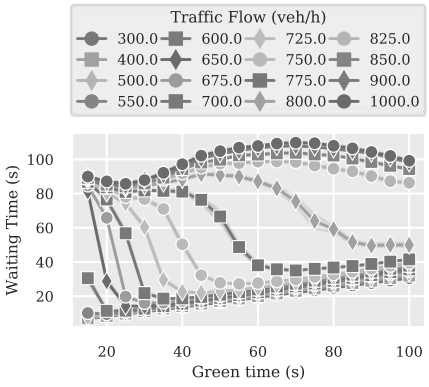


Fig. 4. Tiempo medio de espera (s) frente al tiempo en verde (s) para diferentes valores de flujo de tráfico (veh/h).

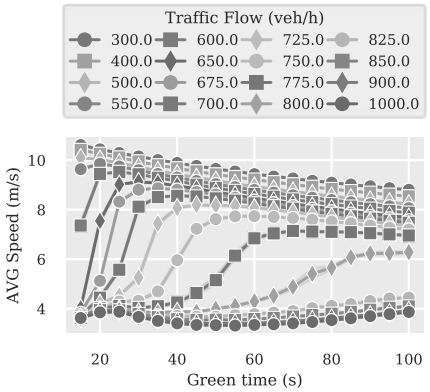


Fig. 5. Velocidad media (m/s) frente al tiempo en verde (s) para diferentes valores de flujo de tráfico (veh/h).

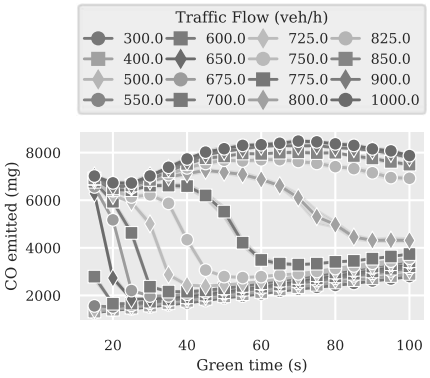


Fig. 6. Emisiones de CO frente al tiempo en verde (s) para diferentes valores de flujo de tráfico (veh/h).

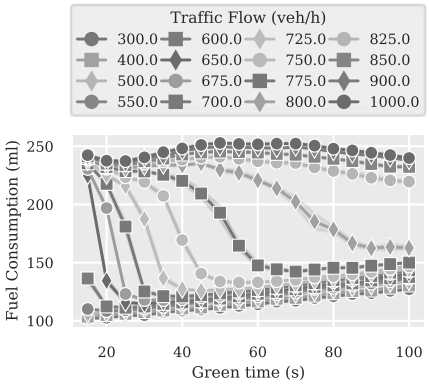


Fig. 7. Consumo (ml) frente al tiempo en verde (s) para diferentes valores de flujo de tráfico (veh/h).

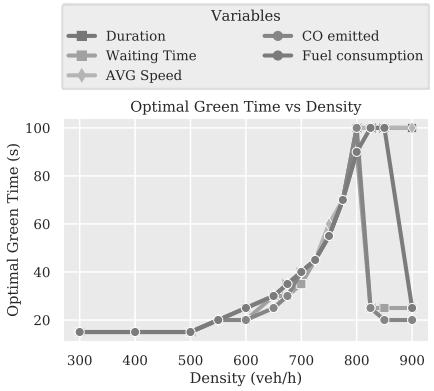


Fig. 8. Valores óptimos de tiempo en verde (s) para diferentes valores de flujo de tráfico (veh/h).

REFERENCIAS

[1] "TomTom Traffic Index." [Online]. Available: www.tomtom.com/en_gb/trafficindex/. [Accessed: 21-May-2019].

[2] "España gasta más de 3.600 millones de euros al año para tratar enfermedades causadas por la contaminación del tráfico." [Online]. Available: https://www.eldiario.es/sociedad/Espana-obligada-millones-enfermedades-contaminacion_0_839916665.html. [Accessed: 21-May-2019].

[3] dgt.es, "Glorietas: Concepto y clases de intersecciones giratorias.," in *Gestión técnica de tráfico*, 1st ed., 2013.

[4] Transportation Research Board (TRB), *Highway Capacity Manual (HCM), Sixth Edition: A Guide for Multimodal Mobility Analysis*, Sixth Edit. Washington, DC: The National, 2016.

[5] P. A. Lopez *et al.*, "Microscopic Traffic Simulation using SUMO," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2575–2582.

[6] Y. Wang, X. Yang, H. Liang, and Y. Liu, "A Review of the Self-Adaptive Traffic Signal Control System Based on Future Traffic Environment," *J. Adv. Transp.*, vol. 2018, pp. 1–12, 2018.

[7] D. I. Robertson, "TRANSYT: A Traffic Network Study Tool," 1969.

[8] J. D. C. Little, M. D. Kelson, and N. M. Gartner, "Maxband: a Program for Setting Signals on Arteries and Triangular Networks," *60th Annu. Meet. Transp. Res. Board*, 1981.

[9] A. G. Sims and K. W. Dobinson, "The Sydney Coordinated Adaptive Traffic (SCAT) System Philosophy and Benefits," *IEEE Trans. Veh. Technol.*, 1980.

[10] D. I. Robertson and R. D. Bretherton, "Optimizing Networks of Traffic Signals in Real Time—The SCOOT Method," *IEEE Trans. Veh. Technol.*, 1991.

[11] J. Hu, M. D. Fontaine, B. B. Park, and J. Ma, "Field Evaluations of an Adaptive Traffic Signal—Using Private-Sector Probe Data," *J. Transp. Eng.*, 2015.

[12] N. Jiang, "Optimal Signal Design for Mixed Equilibrium Networks with Autonomous and Regular Vehicles," *J. Adv. Transp.*, 2017.

[13] R. Sanchez-Iborra and M.-D. Cano, "On the similarities between urban traffic management and communication networks: Application of the random early detection algorithm for self-regulating intersections," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 4, 2017.

[14] R. Sanchez-Iborra, J. F. Ingles-Romero, G. Domenech-Asensi, J. L. Moreno-Cegarra, and M. D. Cano, "Proactive Intelligent System for Optimizing Traffic Signaling," in *Proceedings - 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing, DASC 2016, 2016 IEEE 14th International Conference on Pervasive Intelligence and Computing, PICom 2016, 2016 IEEE 2nd International Conference on Big Data*, 2016.

[15] R. Florin and S. Olariu, "A survey of vehicular communications for traffic signal optimization," *Veh. Commun.*, vol. 2, no. 2, pp. 70–79, 2015.

[16] L. Li, D. Wen, and D. Yao, "A survey of traffic control with vehicular communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp.

425–432, 2014.

[17] A. Pascale, H. T. Lam, and R. Nair, "Characterization of network traffic processes under adaptive traffic control systems," *Transp. Res. Procedia*, vol. 9, pp. 205–224, 2015.

[18] SUMO, "Vehicle insertion in SUMO simulation tool," 2019. [Online]. Available: https://sumo.dlr.de/wiki/Simulation/VehicleInsertion. [Accessed: 30-May-2019].

[19] F. Ahmad, S. A. Mahmud, and F. Z. Yousaf, "Shortest Processing Time Scheduling to Reduce Traffic Congestion in Dense Urban Areas," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 47, no. 5, pp. 838–855, 2017.

[20] J. Garcia-Nieto, A. C. Olivera, and E. Alba, "Optimal cycle program of traffic lights with particle swarm optimization," *IEEE Trans. Evol. Comput.*, 2013.

[21] A. Ahmad, R. Arshad, S. A. Mahmud, G. M. Khan, and H. S. Al-Raweshidy, "Earliest-deadline-based scheduling to reduce urban traffic congestion," *IEEE Trans. Intell. Transp. Syst.*, 2014.

[22] A. Yousef, A. Shatnawi, and M. Latayfeh, "Intelligent traffic light scheduling technique using calendar-based history information," *Futur. Gener. Comput. Syst.*, 2019.

[23] W. Genders and S. Razavi, "Evaluating reinforcement learning state representations for adaptive traffic signal control," in *Procedia Computer Science*, 2018.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Comunicaciones MQTT seguras para el proyecto SmartCampus

Julia Sánchez, Adrián García, Guiomar Corral
Departamento de Ingeniería, Grupo de Investigación en Internet Technologies and Storage (GRITS),
La Salle Campus Barcelona – Universidad Ramon Llull (URL)
Quatre Camins 30, 08022, Barcelona.
j.sanchez@salle.url.edu, adrian.gg@students.salle.url.edu, guiomar.corral@salle.url.edu.

***Resumen*- El grupo de investigación GRITS de la universidad La Salle Campus Barcelona - URL, está llevando a cabo el proyecto SmartCampus que consiste en crear una prueba de concepto de una arquitectura que utilizará tecnologías IoT y Cloud/Fog computing. El proyecto mostrará la viabilidad de su implementación en todo el campus Barcelona para proporcionar diversos beneficios en el día a día de alumnos, profesores y personal de la universidad (mediante los servicios que se implementen). Una arquitectura como esta debe ser escalable y robusta. Soportar varios protocolos de comunicación como MQTT, CoAP, entre otros, y asegurar las transmisiones de los mensajes que se envían mediante estos protocolos. En este documento se presenta una solución basada en MQTT para las comunicaciones entre dispositivos IoT. Este protocolo se adecúa a las necesidades del SmartCampus, pero tiene ciertas limitaciones de seguridad que se deben solventar. Así que se propone una solución de comunicaciones seguras basadas en mecanismos de cifrado AES y se comprueba en un entorno de pruebas la viabilidad de esta solución.**

***Palabras Clave*- IoT, MQTT, Seguridad, Confidencialidad, jitel, telemática**

I. INTRODUCCIÓN

El paradigma IoT (*Internet of Things*) ha ganado importancia en los últimos años, sobre todo, debido a las ventajas que proporciona, ya que puede mejorar nuestra calidad de vida ayudándonos a tomar decisiones difíciles de manera más fácil, o proporcionar inteligencia a los dispositivos que nos rodean para que ejecuten tareas diarias con la mínima intervención humana. IoT convierte los hogares, edificios, hospitales, ciudades, industrias, entre otros, en sistemas inteligentes capaces de obtener el conocimiento del entorno y aplicarlo para su adaptación de acuerdo a las necesidades de los habitantes.

Para obtener estos beneficios, IoT interconecta personas, objetos y dispositivos (vehículos, electrodomésticos, sensores, actuadores, equipos de trabajo, portátiles, smartphones, tablets, etc.) para que se comuniquen inteligentemente los unos con los otros o con las personas. A medida que la cantidad de dispositivos conectados a Internet crece, cada vez es más difícil diseñar infraestructuras TIC que se adapten a las necesidades de los usuarios y faciliten el trabajo de los ingenieros de red y desarrolladores de aplicaciones. Para simplificar estas tareas, surgen nuevas maneras de concebir el IoT, como el *Social IoT* (SIoT). Este nuevo paradigma tiene como objetivo mantener de manera separada el nivel de las personas respecto al nivel de los objetos. Esta división permite a los objetos tener sus propias redes sociales, a la vez que permite a las personas imponer reglas para proteger su privacidad y que solo accedan al resultado de interacciones autónomas entre objetos que ocurren en la red social de dichos objetos. Diseñando la infraestructura con estas premisas, se consigue el soporte de nuevas aplicaciones y servicios de red para IoT de manera más eficiente y efectiva [1].

El trabajo realizado por el grupo de investigación de *Internet Technologies and Storage* (GRITS) en el ámbito de SIoT, apuesta por SIoT como paradigma beneficioso en un entorno como el de la Smart City, por ejemplo, en eficiencia energética (*DSM - Demand Side Management*) [2]. La comunicación entre comunidades de individuos, sensores y actuadores de forma autónoma y dinámica, junto con las tecnologías Cloud/Fog computing (recursos de almacenamiento y computación altamente escalables, elásticos y bajo demanda, para trabajar con datos heterogéneos y en tiempo real), permitirá el diseño de una arquitectura que se adapte a las necesidades del SIoT. Aplicando esta idea, surge el proyecto *SmartCampus*, cuyo objetivo es diseñar una prueba de concepto para (1) almacenar datos heterogéneos en uno (o varios) clouds utilizando el paradigma de SIoT, (2) ofrecer una puerta de acceso única para que los diferentes servicios del *SmartCampus* Barcelona (Smart Lightning, Climatisfacción, Mapa Ruido/Calor, Geolocalización, Asistente Virtual, Escucha Activa, Seguridad de las Personas,

Smart Parking, etc.) puedan acceder a los datos recogidos, y (3) ofrecer una visión holística y centralizada de los eventos que se están monitorizando en el campus. En un futuro, servirá como testbed para un *Facility Manager* del Campus Barcelona. Esta prueba piloto debe sentar las bases para poder extender el proyecto a todo el Campus Barcelona e incluso a otras escuelas o campus Salle.

En arquitecturas de estas características, la gran cantidad de dispositivos de diferentes naturalezas interconectados entre ellos y hacia Internet, junto con el gran volumen de datos que generan, requieren un diseño basado en arquitecturas y tecnologías del entorno IoT, y contemplar la posibilidad de almacenamiento de datos hetereogeneos guardados de manera distribuida en diferentes clouds (para ser tratados posteriormente).

Además, dada la sensibilidad de los datos con los que tratará la *SmartCampus*, esta arquitectura se debe desplegar sobre una infraestructura que asegure la *integridad* y *confidencialidad* de los datos almacenados y en tránsito. Por otra parte, es necesario controlar quién accede a los diferentes servicios y qué acciones se pueden realizar (*control de accesos*).

El proyecto *SmartCampus* está constituido por diferentes fases (evaluación de servicios a desplegar, valoración de las tecnologías IoT y de almacenamiento de datos, diseño de una arquitectura escalable y segura, etc.). Este artículo está focalizado en presentar una *primera aproximación* de la seguridad de las comunicaciones internas entre los dispositivos IoT que forman esta arquitectura *SmartCampus*. En la sección II se presentan las consideraciones que se han tenido en cuenta sobre IoT para la implementación de un entorno seguro. En la sección III, se analizan los diferentes *Smart Gateways* disponibles que se encargarán de la interoperabilidad de los protocolos de comunicación (necesario tener esta visión para un crecimiento futuro de la arquitectura). En la sección IV se presentan las características de MQTT y limitaciones de seguridad, ya que es el protocolo utilizado en las pruebas de confidencialidad realizadas. En la sección V se realiza una propuesta de comunicaciones MQTT seguras aplicando AES y se presentan los resultados de los experimentos de laboratorio llevados a cabo para su comprobación. Finalmente, en la sección VI se muestran las conclusiones y líneas futuras de este trabajo.

II. Consideraciones del ámbito IoT para el entorno SmartCampus

Para plantear una posible solución de la arquitectura *SmartCampus* que refleje las características mencionadas anteriormente, es necesario estar familiarizado con el entorno IoT, entender los componentes que forman su red de información y cómo se relacionan entre ellos.

Como se muestra en la Fig. 1, el ecosistema IoT está compuesto por cinco componentes principales [3]:

- **Dispositivo IoT**. Compuesto por sensores, actuadores, interfaz de comunicación, sistemas operativos, sistemas software, aplicaciones y servicios ligeros (lightweight). Recogen información contextual para después realizar acciones.
- **Coordinador**. Es un gestor de dispositivos que se encarga de monitorizar la salud y las actividades de las *smart things.*
- **Gateway IoT**. Actúa como hub entre la red IoT local y los servicios IoT en el cloud. También interconecta redes IoT locales desiguales (por ejemplo, IP y no IP).
- **Servicios IoT**. Alojados en el cloud para que los usuarios puedan acceder a los objetos IoT en cualquier momento y desde cualquier lugar. Incluye servicios para automatizar procesos, gestionar dispositivos y tomar decisiones.
- **Controladores**. Los dispositivos IoT pueden estar controlados por otros dispositivos como smartphones o tablets.

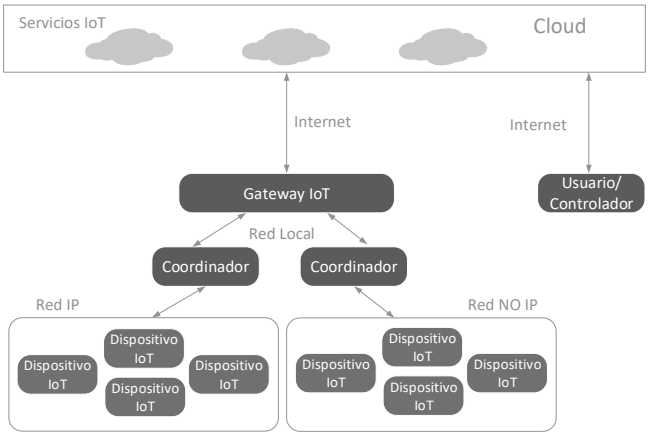


Fig. 1. Interoperabilidad entre dispositivos IoT

La interacción y comunicación entre todos estos componentes para gestionar los datos en IoT, debe cumplir unos requisitos de seguridad. Como en cualquier arquitectura de redes y computadores, los objetivos de conseguir ***confidencialidad***, ***integridad*** y ***disponibilidad*** en las comunicaciones están presentes en estos requerimientos. A parte de estas características básicas, el sistema debe presentar otros requerimientos como *autenticación*, *autorización, control de accesos, anonimidad, no repudio*, *heterogeneidad, anti-replay, soluciones ligeras de seguridad*, entre otros, para proporcionar un entorno seguro [3][4]. Pero, el cumplimiento de estos requerimientos se dificulta debido a que existen ciertas restricciones y limitaciones en los dispositivos y componentes, como recursos de energía y computacionales limitados, que imposibilitan la utilización de protocolos y tecnologías tradicionales de la arquitectura de Internet. Es necesario adoptar otros mecanismos o adaptar los existentes, optimizándolos para que sean capaces de comunicar los dispositivos de manera eficiente y confiable [5].

Dado que las comunicaciones entre los dispositivos IoT son un aspecto clave para la interacción entre los mismos, estos protocolos y tecnologías juegan un rol principal para el despliegue de las aplicaciones. Por ello, surge la necesidad de revisar si estos protocolos y tecnologías utilizadas disponen de los mecanismos adecuados para las comunicaciones seguras. Existe una gran variedad de protocolos y tecnologías que se utilizaran según el entorno IoT específico de trabajo. En la Fig.2 se muestra el *stack* de protocolos de comunicación entre dispositivos IoT. Aunque existen otros, se ha escogido mostrar aquellos más utilizados o consolidados.

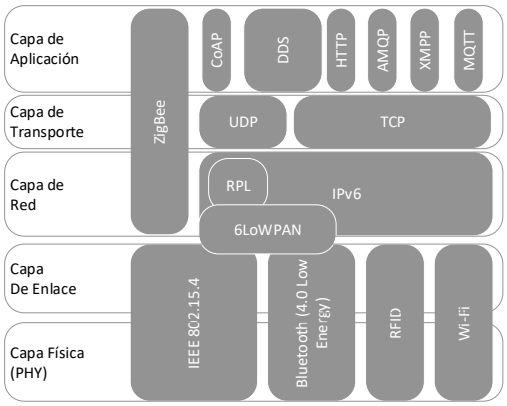


Fig. 2. Stack de protocolos de comunicación IoT

Las implicaciones de seguridad de los protocolos de comunicación de cada capa deben analizarse de manera individualizada para tener una visión completa. El proyecto *SmartCampus* se ha centrado en el análisis en la capa de aplicación y en los protocolos M2M (*machine-to-machine*), ya que se consideran un componente clave para la comunicación entre dispositivos IoT y sus aplicaciones. En la Tabla I se muestra un resumen de las diferentes opciones consideradas por ser los protocolos más analizados en otras investigaciones [6].

Tabla I
CARACTERÍSTICAS PRINCIPALES DE LOS PROTOCOLOS DE LA CAPA DE APLICACIÓN IOT

Protocolo	Transporte	Pub/ Sub	Req/ Res	Security
CoAP	UDP	Si	Si	DTLS
MQTT	TCP	Si	-	SSL/TLS
XMPP	TCP	Si	Si	SSL/TLS
AMQP	TCP	Si	-	SSL/TLS
DDS	TCP/UDP	Si	-	SSL/TLS/DTLS
HTTP	TCP	-	Si	SSL/TLS

De entre ellos, se escoge el protocolo MQTT (*Message Queue Telemetry Transport*) por cumplir con las caracterísitas de ser ligero, flexible y fácil de implementar. Además de ser un protocolo muy extendido en aplicaciones como *healthcare, energía y social networking* (por ejemplo, notificaciones de Facebook) [7]. En la sección IV se detallan las caracteríticas y limitaciones de MQTT.

Por otro lado, para que las comunicaciones sean realmente posibles y como se ha visto en la Fig.1, es necesario el *Gateway IoT*. Este elemento permite adecuar la información transmitida entre dispositivos cuando utilizamos diferentes protocolos o tecnologías para las comunicaciones, tanto entre dispositivos como hacia el exterior (convirtiendo la transmisión entre redes IP y no IP si es necesario). Ésta no es una tarea fácil y por este motivo, otro de los grandes retos de IoT es la **interoperabilidad**. Si se pretende diseñar una infraestructura segura y que además sea escalable y soporte otros protocolos a parte de MQTT, será necesario analizar de cerca este elemento. En la sección III se analizan las diferentes posibilidades de implementación de *Gateways IoT* para decidir la utilizada en la prueba de concepto de *SmartCampus*.

III. SMART GATEWAYS

Los *Gateways IoT* son aquellos dispositivos, normalmente perimetrales, que se encargan de adecuar la información a un entorno compartido y más allá de la red local. Principalmente sirven de convertidores a los protocolos clásicos de Internet como TCP o una transmisión basada en protocolos Web. Esta adecuación de la información es muy necesaria por la naturaleza de los protocolos IoT, los cuales presentan transmisiones diferentes a las tradicionales.

Con la aparición de las tecnologías actuales, el simple concepto de adecuación se queda desfasado convirtiendo a estos dispositivos en *Smart Gateways*. Además de realizar sus funciones principales comentadas anteriormente, se encargan de la orquestación de la red, es decir, actualmente a través de estos dispositivos es posible conocer todo el estado de la red interna, el estado de sus dispositivos y estadísticas individualizadas de cada nodo. Por tanto, se pueden considerar los *Smart Gateways* como dispositivos de administración avanzados. Del mismo modo, la evolución de la tecnología ha permitido que un único dispositivo, *Smart Gateway*, realice las funciones de dos de los componentes principales del entorno IoT (Fig.1), el *Coordinador* y el *Gateway IoT*.

Los dos grandes retos planteados para el diseño de arquitecturas IoT quedan muy reflejados en este tipo de dispositivo. La **interoperabilidad** y la **seguridad** son dos aspectos clave y necesarios a tener en cuenta en el diseño de cualquier red IoT y consecuentemente del *Smart Gateway/s* del sistema.

Para reflejar la dificultad al adecuar los diferentes protocolos entre ellos, se presentan 3 tramas, cada una de un protocolo de comunicación diferente (ver Fig.2), para poder apreciar la gran diferencia tanto de estructura como de información que contiene cada protocolo:

Bit	7	6	5	4	3	2	1	0
Byte 1	MQTT Control Packet type				Flags specific to each MQTT Control Packet type			
Byte 2	Remaining Length							

Fig. 3. Paquete MQTT [8]

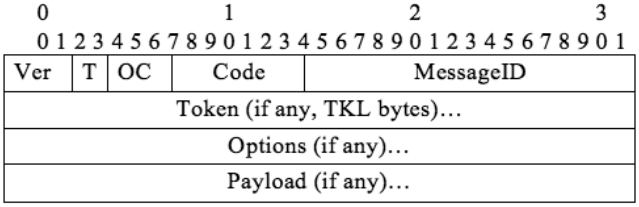


Fig. 4. Paquete CoAP [9]

Transmitted Packet									
Content	Frame Control	MSB	LSB	Identifier	Frame ID	Dest. Address	Options	Payload	Checksum
Size	1 byte	1 byte	1 byte	1 byte	1 byte	2 bytes	1 byte	19 bytes	1 byte

Payload										
Byte	1	2	3	4	5	6	7	8	9	10
Content	Lighting commands		Sync	Spec data[i]		Save	X		Y	
Byte	11	12	13	14	15	16	17	18	19	
Content	Z		V		CCT	Light data	Mode	u'	v'	

Fig. 5. Paquete Zigbee [10]

Durante el paso de los años, tanto las grandes corporaciones como las pequeñas empresas han ido requiriendo diferentes tipos de sensores y actuadores para sus actividades empresariales. Estos productos, ya sea por motivos económicos o bien por la simple evolución del mercado, suelen disponer de diferentes protocolos incorporados, diferentes maneras de comunicarse y, normalmente, es imposible adecuarlos a utilizar uno diferente.

Se presenta un entorno multiprotocolo y con la necesidad de interconectar los dispositivos entre ellos y/o hacia el exterior [11].

La conexión de dichos dispositivos, al igual que los protocolos previamente mencionados, representa un reto de interoperabilidad bastante grande y que a la vez tiene que ser resuelto de una manera u otra. Esta interoperabilidad recae en las *Smart Gateways*, ya que tienen que ser capaces de interconectar o adecuar todas las tramas de estos dispositivos entre ellos. Para solucionar este problema se puede recurrir a varias soluciones, dependiendo de las necesidades y presupuesto. Más adelante se presentan las diferentes opciones de implementación de *Smart Gateway*.

Por lo que respecta a la seguridad, es posible que los Smart Gateways aporten la seguridad del entorno, pero será necesaria la colaboración de los dispositivos IoT, cosa que además de poco probable puede resultar costosa, sobre todo en términos de coste computacional. Por tanto, puede resultar más conveniente securizar a nivel de protocolo o comunicación. Estas soluciones serán presentadas más adelante en la sección V.

Cabe destacar que hacer uso de *Smart Gateways* implica incorporar una capa más de seguridad dentro de la comunicación del sistema; por un lado, con el exterior de la red, o bien con otro Gateway; por otro lado, con el interior de la red. Éste es uno de los grandes puntos a trabajar dentro de las arquitecturas IoT actuales, ya que no se dispone de una estrategia clara para implementar estas medidas de seguridad. En [28] se presenta una propuesta de capa de seguridad añadida como la capa Gateway, aunque de igual manera se podría añadir como parte de la Capa de red tradicional.

A continuación, se presentan las diferentes tipologías de *Smart Gateways* para dar solución en estos entornos multiprotocolo [11].

A. Un Gateway - Varios Protocolos

Es posible pensar que ésta debería ser la solución ideal, ya que se reduce hardware, pero, realmente es la solución menos aplicada, tanto porque tiene un coste bastante elevado como por la complejidad del hardware necesario (con mucha potencia, tanto a nivel de plano de datos como a nivel de plano de control). Esta solución está diseñada principalmente para sistemas en herencia, es decir, sistemas que vengan de modelos clásicos (como puede pasar en entornos ICS, *Industrial Control Systems*) cuyos protocolos sean muy dispersos entre sí y sea necesario un tratamiento de todos ellos de manera conjunta. Estos dispositivos además ofrecen más funciones como el tratamiento de los datos centralizados o la administración de la red.

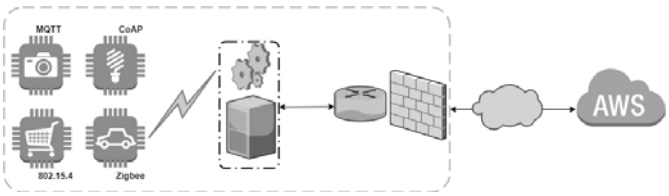


Fig. 6. Smart Gateway multiprotocolo

B. Un Gateway - Un Protocolo

Esta solución, actualmente, es la más aplicada porque en una red estándar, normalmente, no se trabajará con más de 3 - 5 protocolos diferentes. No es necesario disponer de una gran cantidad de Gateways. Este tipo de dispositivos, para la mayoría de protocolos, se puede virtualizar fácilmente reduciendo el espacio y, por tanto, reduciendo el coste del hierro. Un ejemplo sería MOSQUITTO [12] que es un broker/gateway de MQTT Open Source bastante utilizado.

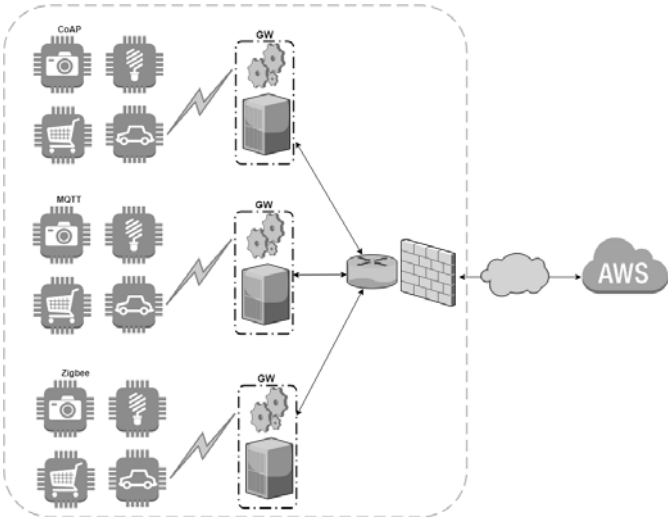


Fig. 7. Smart Gateway por protocolo

C. Fusión de las dos alternativas

La tercera opción se sitúa en un punto intermedio [13]. Consiste en separar la red de sensores/dispositivos IoT en diferentes clusters y que éstos reconduzcan la información hacia el Gateway principal. El Gateway principal puede ser multiprotocolo o también es posible que los propios clusters sean algo más que agrupadores (*Cluster head*) y que hagan la función de Gateway de su protocolo. En este caso, el Gateway superior recibiría todas las tramas con el mismo formato y solo tendría la necesidad de adecuar dicha información hacia el exterior de la red. Este último comportamiento es equivalente a la primera opción de *Un Gateway - Un Protocolo* [14].

Ésta es una solución dinámica y adaptable que podría resultar muy útil en entornos con gran cantidad de sensores, donde sería posible repartir cargas entre los diferentes niveles de Gateways para no sobrecargar al Gateway principal (Fig. 9).

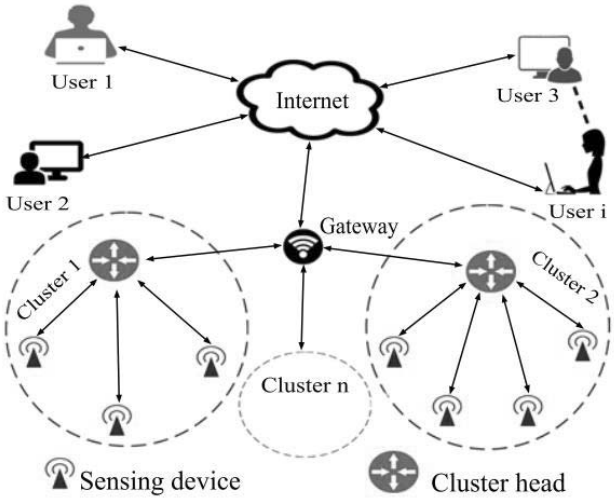


Fig. 8. Smart Gateway con clusters [14]

Conocida la importancia de los *Smart Gateways* en entornos IoT, y vistas las diferentes opciones, es posible decidir qué tipo de gateway utilizar en la prueba de concepto *SmartCampus* (se verá en la sección V).

IV. PROTOCOLO MQTT

MQTT, estandarizado por OASIS (*Organization for the Structured Information Standards*) [15], es uno de los protocolos de comunicación más utilizado y extendido actualmente.

MQTT presenta un modelo muy sencillo y apto para dispositivos con bajas prestaciones, lo cual le ayuda a posicionarse con un gran atractivo dentro del mercado, permitiendo comunicaciones sin prácticamente coste, tanto energético como computacional [16].

A. Características principales de MQTT

MQTT es un protocolo de comunicaciones M2M, basado en la metodología *Publisher-Subscriber*, es decir, los dispositivos basarán su comunicación en una serie de tópicos en los cuales se publicarán diferentes mensajes y los dispositivos interesados se suscribirán a dichos tópicos (Fig. 9).

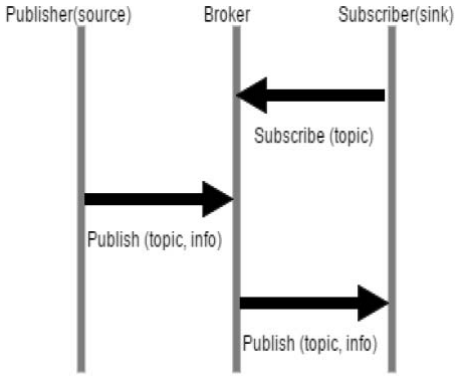


Fig. 9. Proceso Publish/Subscribe de MQTT [16]

Como se aprecia en la Fig. 9, MQTT utiliza un *Broker* como dispositivo central que se encarga de procesar la comunicación entre los clientes y distribuir los mensajes entre ellos basándose en los Topics de interés. Los *Brokers* son el puente entre las aplicaciones y los dispositivos físicos. Existen diferentes tipos, entre ellos, los más populares son *Mosquitto, RSMB, MQTT.js, HiveMQ y VerneMQ*. En [16] se presenta una comparativa de las características y limitaciones entre cada uno de ellos.

MQTT es un protocolo muy ligero, que no consume casi ancho de banda, utiliza TCP (puerto 1883) y tiene tres niveles de QoS.

Los diferentes mensajes del protocolo se diferencian gracias al campo *MQTT Control Packet Type* de 4 bits (Fig. 3). Existen 16 tipos de mensajes diferentes, 14 de los cuales están en uso (1 - 14) y dos reservados para futuros usos (0 y 15), según el estándar 3.1.1, publicado el 24 de octubre de 2014. Estos mensajes permiten establecer tanto la conexión como la desconexión mediante petición y acknowledgement (en caso de la conexión), y gestionar todos los mensajes de publicación y subscripción, que dependiendo del nivel de QoS se tratarán de una manera u otra. Existen tres niveles de QoS: (1) Nivel 0 - un mensaje se entrega como máximo una vez y no requiere acknowledgement, (2) Nivel 1 - cada mensaje se entrega al menos una vez y se requiere acknowledgement, (3) Nivel 2 - se utiliza mecanismo *four-way-handshake* una vez para entregar el mensaje. Cuanto más alto es el nivel de QoS requerido, más mensajes se intercambian con el *Broker* para asegurar la correcta comunicación. Este comportamiento asegura una buena comunicación y por lo tanto una buena difusión de los mensajes de publicación del dispositivo, pero, si se utilizan niveles de QoS altos con frecuencia, se puede provocar sobrecarga en exceso de la red. Es necesario controlar los niveles de QoS según calidad y criticidad de los datos enviados [16].

MQTT incorpora ciertos mecanismos para su protección. Autenticación mediante usuario y contraseña o la posibilidad de usar certificados X.509. Para securizar los datos se puede utilizar SSL/TLS, que permite una comunicación segura cliente-servidor. También ofrece métodos de autorización basados en políticas de seguridad (tipo ACLs (*Access Control Lists*), o políticas RBAC, es decir, políticas de control de acceso basadas, por ejemplo, en un PostgreeSQL o similares). Estas características de seguridad se pueden ofrecer a través del *Broker* [7].

B. Limitaciones de seguridad MQTT

MQTT, es un protocolo que viaja, en su integridad y por defecto, en texto plano por la red. En implementaciones más tempranas, tratando datos sin ningún valor, se podía pensar que no era un problema, pero con un mundo cada día más controlado, hasta un valor de temperatura podría ser crítico [17]. Al viajar sin encriptar es susceptible a ataques MITM (Man-In-The-Middle), o robo de información al vuelo, sobre todo si se consiguen superar las barreras perimetrales de la red.

Aunque ofrece métodos de autenticación mediante usuario y contraseña, al no viajar los datos encriptados, no ofrecen ningún beneficio.

Aunque es posible securizar los datos mediante SSL/TLS, esta tarea solo es viable si, tanto servidor como cliente, tienen la capacidad para ello. Este aspecto supone una gran problemática, ya que la mayoría de dispositivos que se decantan por utilizar este protocolo es, básicamente, porque se trata de un protocolo de bajo coste y por lo tanto es muy probable que no tengan la capacidad necesaria para implementar esta solución (puede suponer mucha sobrecarga de CPU en dispositivos y *Brokers*, sobre todo en la fase de *handshake*) [7].

Por otro lado, las políticas de seguridad ofrecidas para la autorización, tampoco se adaptan a los requerimientos de bajo coste computacional y energético de los dispositivos IoT.

Finalmente, cabe destacar que el 7 de marzo de este año 2019, OASIS sacó una nueva versión de MQTT, la versión 5.0 [18]. Desafortunadamente, no incorpora grandes medidas de seguridad, más allá de la introducción de códigos de respuesta que pueden servir para, como dispositivo final, implementar ciertos métodos de control.

V. SMARTCAMPUS - SEGURIDAD EN LAS COMUNICACIONES MQTT

Como se ha comentado anteriormente, la implementación de una arquitectura adecuada, que sea escalable y robusta, se enfrenta a dos retos importantes; la **interoperabilidad** entre los diferentes protocolos de comunicación presentes en el

entorno IoT y la **seguridad**, condicionada por las limitaciones de los dispositivos IoT que se utilizan.

De este modo, después del análisis presentado y para esta primera aproximación de la solución de comunicaciones seguras en el entorno *SmartCampus*, se ha decidido utilizar MQTT como protocolo de comunicaciones entre los dispositivos (sin perder de vista que esta primera iteración no debe limitar la posibilidad de trabajar, conjuntamente, con otros protocolos como CoAP, o similares, en un futuro). MQTT necesita un *Broker*, elemento central de su funcionamiento. Esta función se centraliza en el *Smart Gateway*. Dado que solo se va a trabajar con un protocolo de comunicaciones por el momento, la implementación del *Smart Gateway* en este caso es *Un Gateway – Un Protocolo*. En posteriores mejoras/incorporaciones del sistema, será necesario evaluar otras posibilidades teniendo en cuenta la cantidad de dispositivos a interconectar y el número de protocolos diferentes que se utilicen.

Aunque MQTT es muy adecuado para la prueba de concepto *SmartCampus* (ligero, flexible, fácil de implementar), los mecanismos de seguridad que incorpora por defecto presentan ciertas limitaciones (vistas en la sección anterior). Las primeras mejoras en seguridad que se introducirán en esta prueba de concepto *SmartCampus*, estarán relacionadas con la *confidencialidad* de los datos transmitidos. Por tanto, es necesario plantear una solución que encripte los datos en transmisión y que sea viable tanto para MQTT como para los dispositivos que lo utilizan.

A. Propuesta

Existen diversos trabajos de investigadores del sector que tratan esta problemática de la confidencialidad en las transmisiones MQTT; [19] [20] [21] investigaciones que realizan comparativas entre diferentes técnicas de cifrado como RSA (Rivest-Shamir-Adleman) y ECC (Elliptic Curve based Cryptography), [22] [23] propuestas de cifrado basadas en Key/Ciphertext Policy-Attribute Based Encryption (KP/CP-ABE) utilizando una versión ligera de ECC e, incluso, [24] alguna propuesta que añade la utilización de Dynamic S-Box Advanced Encryption Algorithm (AES) a KP-ABE sobre una versión ligera de ECC.

Después de analizar la aplicación de estas propuestas en el entorno *SmartCampus*, se ha llegado a la conclusión de que su coste computacional y la complejidad de la infraestructura que lo tendría que soportar, no se adecúan a las necesidades del proyecto *SmartCampus*.

Por otro lado, trabajos como los presentados en [25] muestran que, según el entorno de trabajo, se pueden asegurar las comunicaciones en el sistema mediante la aplicación de AES.

Para la propuesta de sistema de comunicaciones seguras con MQTT en el *SmartCampus* se ha decidido cifrar los datos utilizando cifrado simétrico AES (128 o 256) en modo CBC (Cipher-Block Chaining). La solución se basará en la incorporación de una capa intermedia de seguridad dentro de cada uno de los mensajes PUBLISH de MQTT, concretamente, se asegura solo el payload de cada mensaje (siendo extensible a todo el mensaje si se adapta a los requerimientos del protocolo y el *broker/gateway* utilizado).

La seguridad en la comunicación entre el *Smart Gateway* y el exterior (Internet/Cloud), se realiza mediante protocolos tradicionales de Internet. Y aunque está fuera del alcance de esta propuesta asegurar la comunicación más allá del *Smart Gateway* (conexión hacia el exterior), es interesante destacar que para asegurar mejor el sistema es aconsejable la utilización de dispositivos como IDS/IPS (*Intrusion Detection/Intrusion Prevention System*). En caso de tener alguna brecha de seguridad en el perímetro y, por tanto, con la posibilidad de acceder a las tramas MQTT de la red interna, la confidencialidad de las transmisiones se puede ver comprometida. Un sistema IDS/IPS puede detectar a un intruso, bloquearlo y alertar para que se tomen las medidas necesarias en la parte interna de la red (por ejemplo, forzar la renovación de las claves de cifrado por si han sido extraídas de alguna manera o se han visto comprometidas).

Esta propuesta del cifrado del payload con AES permite obtener una solución de fácil implementación, que no requiere muchos recursos y que además sirve como testeo para aplicar la misma idea a otros protocolos como CoAP. Además, es una solución escalable para entornos con muchos sensores o dispositivos. Pero su escalabilidad dependerá de la manera en que se proporcionan las claves a los usuarios, dado que un gran número de usuarios puede decrementar el nivel de seguridad porque la distribución de las claves a una gran cantidad de usuarios puede comprometer la confidencialidad. Esto implica que es necesario realizar una adecuada gestión de claves [20], aspecto que está fuera del alcance del trabajo presentado en este documento, pero se pretende tratar en futuras investigaciones.

B. Implementación Testbed SmartCampus

Para la implementación del entorno de pruebas de la propuesta definida, existen las siguientes opciones.

- **Implementación mediante *middleware***. Es una solución que consiste en un conjunto de módulos desacoplados, basada en que tanto cliente como *Broker/Gateway* MQTT sean completamente transparentes a la aplicación de los mecanismos de cifrado. Son otros módulos ajenos a la comunicación MQTT (aunque dentro de los dispositivos) los que se encargan de aplicar la seguridad de la transmisión (Fig. 10).

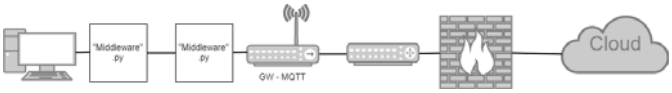


Fig. 10. Solución Middleware

- **Implementación mediante cliente:** Solución que permite tener un comportamiento más basado en el estándar MQTT, ya que serán los propios clientes los encargados de asegurar sus mensajes y otro cliente conectado al *Broker/Gateway*, el encargado del tratamiento de esta seguridad (Fig. 11).

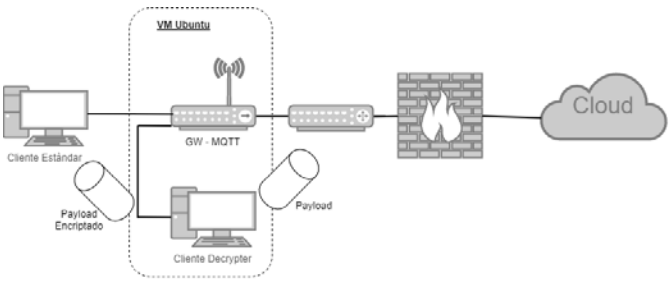


Fig. 11. Solución cliente

Se utilizará la segunda opción, *implementación mediante cliente*, porque respeta los estándares de comunicación y las bases del protocolo MQTT, pero además porque es posible mantener un único programa activo dentro del dispositivo cliente. En la opción de *middleware*, es necesario tener una segunda instancia o programa en el cliente que se encargue de generar la conexión TCP al *middleware* del gateway añadiendo complejidad y más carga (pudiendo llegar a debilitar la batería).

En cambio, la segunda opción, requiere modificar el hilo principal del dispositivo (cliente y *gateway*) cargando el módulo de encriptación y, a parte, en el *gateway* se requiere añadir un nuevo cliente que se encarga del descifrado de los datos.

Las pruebas se han realizado en dos escenarios diferentes que contemplan: (1) la posibilidad de implementar la solución propuesta en diferentes tipos de dispositivos IoT (con diferentes características), (2) la diferenciación entre comunicaciones de la red interna a la externa y comunicaciones dispositivo-a-dispositivo.

- **Pruebas para la comunicación dispositivo-a-dispositivo** (Fig. 12). Se utiliza un cliente basado en C++ implementado en un Arduino MEGA (plataforma muy extendida para la implementación de dispositivos IoT). El cliente en el Arduino MEGA dispone de un único fichero capaz de recibir y tratar la información de sus sensores, cifrarla con AES128 y mandarla mediante MQTT. Para la implementación se ha requerido el uso de las bibliotecas AESLib [26] y PubSubClient [27]. Además, para la simulación de sensores de este dispositivo se ha utilizado un lector RFID MFRC522 con lo que también se ha necesitado la biblioteca MFRC522 [29]. Paralelamente, se han utilizado bibliotecas nativas de Arduino como SPI y Ethernet porque la conexión entre dispositivos se ha simulado cableada en vez de realizarla inalámbrica (lo cual no afecta al hecho de que se llegue a cifrar correctamente la información y se pueda determinar si la solución es viable). Además, el mismo fichero actúa como cliente de descifrado, ya que sencillamente, al disponer de las bibliotecas y las claves tanto de AES128 como el IV (Initialization Vector), se puede activar el *callback* de MQTT para procesar los mensajes a los que está suscrito.

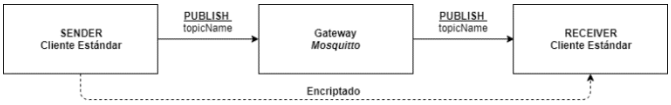


Fig. 12. Diagrama de estados End-to-End

El código fuente de la implementación dispositivo-a-dispositivo está disponible en Github [30].

- **Pruebas para la comunicación de la red interna a la red externa.** Se utiliza un cliente basado en Java SDK 1.8 que podría instalarse fácilmente en diferentes tipos de dispositivos IoT. La solución se implementa en dos programas diferentes, uno simula el cliente y su comportamiento y el otro, implementado en el *Broker/Gateway* para descifrar las tramas previas al envío hacía el exterior. La seguridad más allá del Gateway es una responsabilidad ajena a la propuesta realizada en este documento y, además, podrá realizarse normalmente con tecnologías tradicionales de Internet. La solución ha requerido utilizar la biblioteca Paho MQTT [31] que permite implementar el protocolo MQTT en programas Java. Como se aprecia en la Fig 13, el programa del cliente se encarga de cifrar el mensaje con el tópico determinado por el cliente, de manera completamente transparente al comportamiento del sensor, y enviarlo al Gateway. Por otro lado, el programa en el Gateway que juega el papel de cliente anexado a éste, es el encargado de descifrar la información recibida para poder enviarla al Cloud.

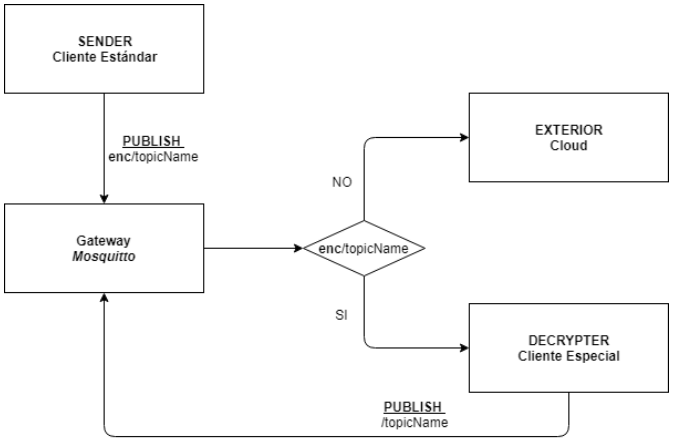


Fig. 13. Diagrama de estados Interior-to-Exterior

El código fuente de la implementación de las comunicaciones de la red interna a la red externa está disponible en Github [32] [33].

Para ambas pruebas, como *Broker*/*Gateway* se ha utilizado Mosquitto 1.4.15 (lanzamiento de febrero 2019).

Después de implementar cada una de las pruebas, ha sido necesario comprobar que los datos se estaban cifrando, llegaban al *Broker*/*Gateway*, y éste era capaz de descifrarlos. Se ha utilizado un capturador de tráfico (Wireshark) para recoger el tráfico entre cliente y *Broker*, y se ha comprobado que efectivamente el *payload* del mensaje MQTT se transmite cifrado.

VI. Conclusiones

El trabajo presentado en este documento, muestra como la implementación de una arquitectura viable para el proyecto *SmartCampus* debe tener en cuenta solucionar dos grandes retos: la interoperabilidad de los diferentes protocolos de comunicación y la seguridad de las comunicaciones entre dispositivos.

Respecto a la interoperabilidad, las tendencias del mercado marcan soluciones de *Smart Gateways* de *Un Gateway - Un Protocolo.* Por el momento, esta solución es la más recomendable para el caso de uso del proyecto *SmartCampus*, dado que no se prevé un gran incremento en la cantidad de protocolos de comunicación que estarán presentes en un futuro. Pero, con la constante evolución del mercado y los dispositivos, es posible que en los próximos años aparezcan soluciones virtualizadas exitosas de Gateways multiprotocolo.

Por otro lado, se ha escogido MQTT para empezar a realizar pruebas de comunicaciones entre dispositivos IoT que simulen las posibles interacciones en la arquitectura *SmartCampus*. Pero, MQTT tiene ciertas limitaciones a nivel de seguridad que hace falta tratar para que la arquitectura

SmartCampus sea segura. Como primera aproximación a un entorno seguro, se ha tratado la confidencialidad de los datos utilizando AES para cifrarlos al transmitirlos. Cabe destacar que se propone esta mejora para empezar a incorporar características de seguridad dentro del proyecto *SmartCampus.* En iteraciones y pruebas posteriores, se seguirán tratando e incorporando otras características principales de seguridad, como autenticación y autorización.

Es muy necesaria la creación de líneas de trabajo y entornos de pruebas como los presentados en este documento para llegar a cubrir los diferentes casos de uso dentro del proyecto *SmartCampus*. Estas nuevas líneas de trabajo pueden contemplar mejoras sobre la seguridad del propio trabajo presentado, como, por ejemplo, la implementación de un buen sistema de gestión de claves (KMS, *Key Management System*) como los presentados en [34] [35]. También sería interesante estudiar la extrapolación de las pruebas realizadas a otros protocolos de comunicación para ver su viabilidad y adecuación.

AGRADECIMIENTOS

Este trabajo se está llevando a cabo gracias a la financiación recibida por la URL (Universitat Ramon Llull) para promover la investigación en sus diferentes facultades.

REFERENCIAS

[1] ATZORI, Luigi, et al. The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization. Computer networks, 2012, vol. 56, no 16, p. 3594-3608.

[2] GELAZANSKAS, Linas; GAMAGE, Kelum AA. Demand side management in smart grid: A review and proposals for future direction. Sustainable Cities and Society, 2014, vol. 11, p. 22-30.

[3] Hossain, M. M., Fotouhi, M., & Hasan, R. (2015, June). Towards an analysis of security issues, challenges, and open problems in the internet of things. In Services (SERVICES), 2015 IEEE World Congress on (pp. 21-28). IEEE.

[4] Yousuf, T., Mahmoud, R., Aloul, F., & Zualkernan, I. (2015). Internet of Things (IoT) Security: Current Status, Challenges and Countermeasures.

[5] Granjal, J., Monteiro, E., & Silva, J. S. (2015). Security for the internet of things: a survey of existing protocols and open research issues. IEEE Communications Surveys & Tutorials, 17(3), 1294-1312.

[6] ESFAHANI, Alireza, et al. A lightweight authentication mechanism for m2m communications in industrial iot environment. IEEE Internet of Things Journal, 2017.

[7] SONI, Dipa; MAKWANA, Ashwin. A survey on MQTT: a protocol of internet of things (IoT). En International Conference On Telecommunication, Power Analysis And Computing Techniques (ICTPACT-2017). 2017.

[8] Victor Pasknel. Mqtt packet format. https://images.app.goo.gl/QFgHWs4WHNtJ9Lr79.

[9] Xi Chen. Coap packet format. https://images.app.goo.gl/wV8qqMydJF44LMqL8.

[10] Ivan Chew. Zigbee packet format. https://images.app.goo.gl/2V83JVMYAvbaqSSH7.

[11] KIM, Jung Tae. Requirement of security for IoT application based on gateway system. communications, 2015, vol. 9, no 10, p. 201-208.

[12] Eclipse. Mosquitto broker. https://mosquitto.org/.

[13] AAZAM, Mohammad; HUH, Eui-Nam. Fog computing and smart gateway based communication for cloud of things. En 2014 International Conference on Future Internet of Things and Cloud. IEEE, 2014. p. 464-470.

[14] Vanga Odelu Neeraj Kumar Mauro Conti Minho Jo Mohammad Wazid, Ashok Kumar Das. Design of secure user authenticated key management protocol for generic iot networks. IEEE Internet of Things Journal, 2017.

[15] OASIS Standard. Mqtt version 3.1.1. http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3. 1.1-os.html.

[16] YASSEIN, Muneer Bani, et al. Internet of Things: Survey and open issues of MQTT protocol. En 2017 International Conference on Engineering & MIS (ICEMIS). IEEE, 2017. p. 1-6.

[17] Eric Auchard Christoph Steitz. German nuclear plant infectedwith computer viruses, operator says. http://bit.do/eShxC.

[18] OASIS Standard. Mqtt version 3.1.1. https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html.

[19] JAYAN, Arvind P.; HARINI, N. A Scheme to Enhance the Security of MQTT Protocol. International Journal of Pure and Applied Mathematics, 2018, vol. 119, no 12, p. 13975-13982.

[20] MEKTOUBI, Abdessamad, et al. New approach for securing communication over MQTT protocol A comparaison between RSA and Elliptic Curve. En 2016 Third International Conference on Systems of Collaboration (SysCo). IEEE, 2016. p. 1-6.

[21] DIRO, Abebe Abeshu; CHILAMKURTI, Naveen; VEERARAGHAVAN, Prakash. Elliptic curve based cybersecurity schemes for publish-subscribe Internet of Things. En International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness. Springer, Cham, 2016. p. 258-268.

[22] SINGH, Meena, et al. Secure mqtt for internet of things (iot). En 2015 Fifth International Conference on Communication Systems and Network Technologies. IEEE, 2015. p. 746-751.

[23] RAHMAN, Abdur, et al. A Lightweight Multi-tier S-MQTT Framework to Secure Communication between low-end IoT Nodes. En 2018 5th International Conference on Networking, Systems and Security (NSysS). IEEE, 2018. p. 1-6.

[24] BISNE, Lochan; PARMAR, Manish. Composite secure MQTT for Internet of Things using ABE and dynamic S-box AES. En 2017 Innovations in Power and Advanced Computing Technologies (i-PACT). IEEE, 2017. p. 1-5.

[25] CHOWDHURY, Fahim Shahriar, et al. An implementation of a lightweight end-to-end secured communication system for patient monitoring system. En 2018 Emerging Trends in Electronic Devices and Computational Techniques (EDCT). IEEE, 2018. p. 1-5.

[26] Davy Landman. Aes lib. https://github.com/DavyLandman/AESLib.

[27] knolleary. Arduino client for mqtt. https:// pubsubclient.knolleary.net/.

[28] Jung Tae Kim. Requirement of security for iot application basedon gateway system. International Journal of Security and ItsApplications Vol.9, No.10 (2015), pages 201–208, 2015.

[29] Miguel Balboa. Arduino rfid library for mfrc522 (spi). https://www.arduinolibraries.info/libraries/mfrc522.

[30] Adrián García Garrido. Arduino project mqtt repository. https://github.com/adrig-geek/MQTT_Arduino.

[31] Eclipse. Paho mqtt java. https://www.eclipse.org/paho/clients/java/.

[32] Adrián García Garrido. Java decrypter project mqtt repository. https://github.com/adrig-geek/MQTT_Java_Decrypter.

[33] Adrián García Garrido. Java encrypter project mqtt repository.https://github.com/adrig-geek/MQTT_Java_ClientEncrypter.

[34] KUNG, Yi-Hsuan; HSIAO, Hsu-Chun. GroupIt: Lightweight Group Key Management for Dynamic IoT Environments. IEEE Internet of Things Journal, 2018, vol. 5, no 6, p. 5155-5165.

[35] ABDMEZIEM, Mohammed Riyadh; CHAROY, François. Fault-tolerant and Scalable Key Management Protocol for IoT-based Collaborative Groups. En Security and Privacy in Communication Networks: SecureComm 2017 International Workshops, ATCS and SePrIoT, Niagara Falls, ON, Canada, October 22–25, 2017, Proceedings 13. Springer International Publishing, 2018. p. 320-338.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Algunos aspectos de modelado en la gestión de traspasos

Vicente Casares-Giner, Daniel Sáez-Domingo
VCG: Instituto ITACA, DSD: Instituto Tecnológico de Informática
Universitat Politècnica de València
Camino de vera s/n. 46022, València
vcasares@upv.es, dsaez@iti.upv.es

***Resumen*—Se describen diversos algoritmos de traspaso de llamadas o sesiones en sistemas celulares inalámbricos. La novedad reside en aportar cotas superiores e inferiores en los parámetros de prestaciones de uno de los algoritmos. La proximidad de las cotas es arbitrariamente cercana, tal que resultan prácticamente iguales a efectos de ingeniería. Se proponen generalizaciones para otros algoritmos cuyas características analíticas son similares al analizado.**

***Palabras Clave*—Traspaso (*handover*), bloqueo, pérdida, demora, terminación forzosa.**

I. INTRODUCCIÓN

En sistemas celulares inalámbricos, cuando una llamada en curso abandona la zona de cobertura en la cual está siendo atendiendo, el sistema debe de garantizar su continuidad asignándole los recursos de radio necesarios en la nueva zona o célula que visita. Los mecanismos de señalización asociados a la gestión de tales recursos de radio han venido identificándose como algoritmos de traspaso siendo éstos de capital importancia por lo que un adecuado modelado y evaluación de sus prestaciones se hace necesario. Aunque tradicionalmente las llamadas de telefonía celular y los algoritmos de traspaso se han asociado a conmutación de circuitos [1], tales algoritmos siguen vigente por cuanto servicios de VoIP ofrecidos por sistemas "All-IP" como LTE pueden no estar implementados. Mientras tanto, las referidas llamadas cabe encaminarlas por las redes de acceso 2G/3G suponiendo la existencia de coberturas solapadas. Tal es el caso del servicio *Circuit Switched FallBack* (CSFB) [2]. En esencia, en CSFB cuando el terminal de usuario móvil (TU) está operando en LTE modo paquete y tiene una llamada entrante, la entidad *Mobility Management Entity* (MME) ordena una búsqueda (*paging*). El TU responde con un mensaje de petición de servicio especial y la red señaliza con el TU indicándole que se conecte a 2G/3G para aceptar la llamada. De forma similar para llamadas salientes, el mismo mensaje de petición de servicio especial se utiliza para encauzar la llamada saliente del TU a través de redes 2G/3G. Aun admitiendo situación "All-IP", sesiones en curso en modo paquete tienen la necesidad de gestionar los traspasos en modo *host based* o *network based* [3] dada la carga de señalización asociada que debe ser evaluada. Razones todas ellas, entre otras, de peso justificativo del presente trabajo.

El trabajo se ha estructurado según sigue. Tras la presente sección introductoria, en la sección II se describen diversos esquemas de tratamiento de llamadas nuevas y en modo traspaso. En la sección III se justifica el uso de herramientas de Markov para modelar los escenarios bajo estudio. En la sección IV se aporta la algorítmica para la obtención de las cotas de los parámetros de relevancia. El estudio de un sistema con varias células se aborda en la sección VI. En la sección VII se justifica el modelo de movilidad elegido. Los resultados ilustrativos se reportan en la sección VIII. El artículo finaliza con las conclusiones y objetivos para la extensión del presente trabajo.

II. DIVERSOS ESQUEMAS DE TRASPASO EN CELULARES

Desde el punto de vista del usuario, es menos deseable la interrupción de una llamada en curso que bloquear el inicio de la misma. Ello supone un trato de favor para las primeras al tratar de darles continuidad cuando cambian de zona de cobertura. De ahí la existencia de esquemas de priorización, uno de los cuales es el *Guard Channel Algorithm* (GCA) [4]. En el GCA[1], tanto las llamadas originadas en la propia célula como las que provenientes de células vecinas que precisan ser traspasadas, se admiten mientras el número de canales libres sobre un total de C, sea superior a C_h, ($C_h < C$). De lo contrario, únicamente se aceptan de inmediato las llamadas de traspaso,

[1]La idea subyacente en el GCA no es de uso exclusivo en sistemas celulares, pues también ha sido propuesto en sistemas *trunking* para la priorización de llamadas de interconexión frente a las propias peticiones de *dispatch*, [5]. En tales entornos, la denominación adoptada ha sido la de *Reserve Margin Algorithm* (RMA) [6].

Tabla I
TRATAMIENTO DE LLAMADAS.

Esquemas	Nuevas, (f)	Traspasadas, (h)	Tamaño colas, Q_*
1	pérdida	pérdida	$(Q_f = 0, Q_h = 0)$
2	pérdida	espera	$(Q_f = 0, Q_h \to \infty)$
3	espera	pérdida	$(Q_f \to \infty, Q_h = 0)$
4	espera	espera	$(Q_f \to \infty, Q_h \to \infty)$

bloqueándose las de origen. Finalmente, tras alcanzar la máxima ocupación de C canales, todas las llamadas se bloquean. Obviamente al particularizar a $C_h = 0$ el tratamiento del traspaso deja de ser prioritario.

Tras la acción de bloquear de una llamada nueva o de traspaso caben diversos tratamientos, ver Tabla I. Un primer esquema es el bloqueo de ambos flujos en modo pérdida o rechazo. Un segundo esquema es rechazar las llamadas de origen y posicionar las de traspaso en una cola de capacidad Q_h quedando a la espera de asignarles algún canal que se libere. La espera vendrá condicionada al tiempo de residencia en el área de traspaso, identificada como el área de solape o de cobertura común entre células geográficamente vecinas, tiempo de residencia que es función de la superficie de solape y de la velocidad del TU, Fig. 1. Tal esquema ha sido analizado por Hong y Rappaport en [7] con algoritmo GCA y disciplina de cola *first-in-first-out* (FIFO) para las llamadas traspasadas. Añadir que una llamada en modo traspaso podría finalizar durante su estancia en la zona de traspaso o abandonar la misma previo a su finalización, experimentando así una terminación forzosa. El primer efecto no ha sido considerado en [7], el cual fue incorporado en [8] y también en nuestro presente estudio. Un tercer esquema, dual del anterior, consiste en rechazar las llamadas en modo traspaso y posicionar las nuevas o de origen en una cola de capacidad Q_f a la espera de asignarles algún canal que se libere. En este esquema, la espera es función del tiempo residual de residencia en el área de cobertura de la célula de origen, área no común con otras células vecinas, Fig. 1. Tal esquema ha sido analizado por Guerin en [4] y

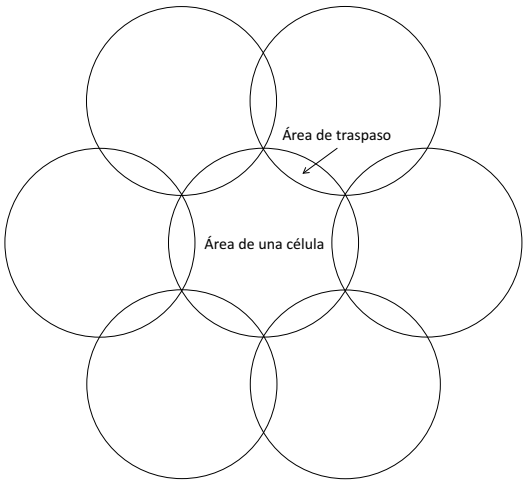


Figura 1. Ilustración del área celular (sin solape) y del área de traspaso.

Tabla II
PRINCIPALES PARÁMETROS PARA EL ESQUEMA 2 DE LA TABLA I

Capacidad (canales)	Definición
C	# de canales en la célula
C_h	# of canales reservados para traspaso
Tasa (v.a.e.)	**Definición**
λ_f	Llamadas nuevas ofrecidas
λ_h	Llamadas traspasadas ofrecidas
$\lambda_t = \lambda_f + \lambda_h$	Tasa total llamadas ofrecida
γ_f	Llamadas nuevas admitidas
γ_h	Llamadas traspasadas admitidas
$\gamma_t = \gamma_f + \gamma_h$	Tasa total llamadas admitidas
μ_M	Duración de la llamada (mensaje)
μ_R	Residencia en el área de una célula
$\mu_H = \mu_M + \mu_R$	Ocupación del canal en una célula
μ_F	Residencia en el área de traspaso
$\mu_Q = \mu_M + \mu_F$	Ocupación del canal en modo traspaso
$q = C\mu_H/\mu_Q$	$q_f = \lfloor q \rfloor, \quad q_c = \lceil q \rceil$
Erlangs	**Tráfico**
$A_{os} = \lambda_f/\mu_M$	Tráfico ofrecido -sesión-
$A_f = \lambda_f/\mu_H$	Tráfico ofrecido -canal-
$A_h = \lambda_h/\mu_H$	Tráfico de traspaso -canal-
$A_t = A_f + A_h$	Tráfico total ofrecido -canal-
$A_q = \lambda_h/\mu_Q$	Tráfico de traspaso en espera
Probabilidades	**Definición**
$P_{sc} = \mu_M/\mu_H$	Fin de llamada en célula
$P_{hd} = \mu_R/\mu_H$	Solicitud de un traspaso
$P_{sh} = \mu_M/\mu_Q$	Fin de llamada en el área de traspaso
P_P	Pérdida de llamadas nuevas
P_{fh}	Traspaso solicitado fallido
P_{FT}	Terminación forzosa
P_{NC}	No finalización de llamada

reconsiderado por Daigle y Jain en [9] al proponer técnicas de análisis alternativas sustentadas en procesos QBD y en los fundamentos de la cola $M/G/1$ [10]. Finalmente, un cuarto esquema es el bloqueo en modo espera de ambos flujos de llamadas, el cual fue tratado en Chang *et. al.* [8], suponiendo colas de capacidad finita, $Q_f < \infty$, $Q_h < \infty$. Un análisis aproximado para cuando $Q_f \to \infty$, $Q_h \to \infty$ se reporta en [11].

III. HIPÓTESIS MARKOVIANA

La propiedad de memoria nula de la variable aleatoria exponencial (v.a.e.) ha sido el principal atractivo en el modelado de sistemas y redes de telecomunicación, razón por la que tal hipótesis se asume en el presente trabajo. Por una parte, la v.a. tiempo entre llegadas consecutivas se suponen exponencial, esto es, ley de Poisson, tanto para llamadas nuevas como para las que solicitan un traspaso, con tasas respectivas λ_f y λ_h. Aunque en el segundo caso la suposición puede cuestionarse, la misma ha sido ampliamente debatida, justificada y aceptada en la literatura, [12], [13]. Igualmente supondremos v.a.e. para la duración de una llamada o sesión y para los tiempos de residencia en el área celular sin solape alguno y en el área de traspaso, con tasa respectivamente μ_M, μ_R, y μ_F. En

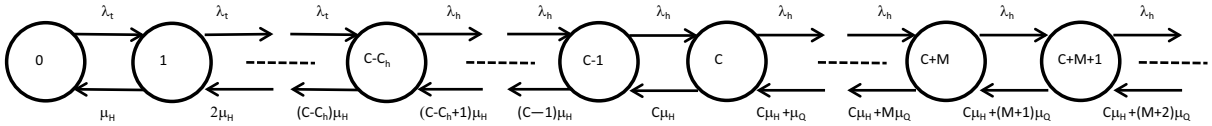


Figura 2. Diagrama de transición de estados del esquema pérdida-espera.

cuanto a la ocupación de un canal, hemos de subrayar la diferencia de si el TU reside en el área celular sin solape o en área de traspaso, Fig. 1. Por la propiedad de memoria nula, las respectivas tasas resultan ser $\mu_H = \mu_M + \mu_R$ y $\mu_Q = \mu_M + \mu_F$. Obviamente $\mu_R < \mu_F$ por lo que $\mu_H < \mu_Q$. Hemos de señalar que en la mayoría de trabajos se ha supuesto $\mu_Q \approx \mu_F$ y según nuestro conocimiento, en [8] se supuso por primera vez, que la llamada puede finalizar en el área de traspaso mientras permanece a la espera de un traspaso. Y añadimos que, según [14], [15], el tiempo medio de residencia en el área de traspaso, $1/\mu_F$, puede ser del orden de 5-10 segundos, siendo ser bastante inferior al tiempo medio de residencia celular $1/\mu_R$.

IV. El modelo pérdida-espera

De los cuatro esquemas reflejados en la Tabla I, tres de ellos son modelos Markovianos con infinito número de estados cuando $Q_f \to \infty$ y/o $Q_h \to \infty$ y que en general no ofrecen expresiones analíticas computables en modo exacto. Nuestra contribución consiste en aportar cotas muy próximas que permitan paliar la anterior limitación en el esquema 2, y que nos puede servir de punto de inicio para una formulación paralela de los esquemas 3 y 4.

El diagrama de estados asociado al esquema 2 se refleja en la Fig. 2. Planteando las ecuaciones de flujo obtenemos las probabilidades en régimen permanente,

$$P_k = \begin{cases} \frac{A_t^k}{k!} P_0; & 0 \le k \le C - C_h \\ \frac{A_t^{C-C_h} A_h^{k-(C-C_h)}}{k!} P_0; & C - C_h \le k \le C \\ \frac{A_t^{C-C_h} A_h^{C_h} A_q^{k-C}}{C! \prod_{n=1}^{k-C} (q+n)} P_0; & C < k \end{cases} \tag{1}$$

con,

$$1/P_0 = P_0^{-1} = P_0^{(-1)} + P_0^{(-2)} + P_0^{(-3)} \tag{2}$$

y,

$$P_0^{(-1)} = \sum_{k=0}^{C-C_h} \frac{A_t^k}{k!} \tag{3}$$

$$P_0^{(-2)} = \frac{A_t^{C-C_h}}{(C-C_h)!} \sum_{k=1}^{C_h} \frac{A_h^k}{\prod_{n=1}^{k} (C - C_h + n)} \tag{4}$$

$$P_0^{(-3)} = \frac{A_t^{C-C_h} A_h^{C_h}}{C!} S_1(A_q, q) \tag{5}$$

habiendo definido A_f, A_h, A_t, A_q y q en la Tabla II. En (5) se ha definido,

$$\begin{aligned} S_1(A_q, q) &= \sum_{k=1}^{M} \frac{A_q^k}{\prod_{n=1}^{k} (q+n)} + S_1(A_q, q, M) \\ S_1(A_q, q, M) &= \sum_{k=M+1}^{\infty} \frac{A_q^k}{\prod_{n=1}^{k} (q+n)} \end{aligned} \tag{6}$$

A. Acotación de P_0

Nuestro objetivo es evaluar la suma con infinitos términos de (5)-(6) para cualquier valor real y positivo de q y consecuentemente los parámetros de prestaciones de interés, Sec. V. Para tal fin, en primera instancia aportamos un cota superior y otra inferior de $P_0^{(-3)}$, ambas función del número entero positivo M introducido al definir $S_1(A_q, q, M)$, Ec. (6). Sean q_f y q_c dos números enteros definidos en la Tabla II. Obviamente, la suma $S_1(A_q, q, M)$ queda acotada por, [16],

$$S_1(A_q, q_c, M) \le S_1(A_q, q, M) \le S_1(A_q, q_f, M) \tag{7}$$

Dado un entero positivo z, p.e., q_f o q_c, $S_1(A_q, z, M)$ se puede evaluar mediante la siguiente expresión cerrada,

$$S_1(A_q, z, M) = \frac{z!}{A_q^z} \left(e^{A_q} - \sum_{k=0}^{M+z} \frac{A_q^k}{k!} \right) \tag{8}$$

Al computar (8) nos hemos encontrado con algunos errores de redondeo significativos, en particular para grandes valores de M y de A_q. Inconveniente que se ha mitigado derivando nuevas cotas para $S_1(A_q, q, M)$, aportación novedosa al comparar con [16]. De hecho, de la Ec. (6) podemos escribir,

$$S_{1,lw}(A_q, q, M) \le S_1(A_q, q, M) \le S_{1,up}(A_q, q, M) \tag{9}$$

con,

$$\begin{aligned} S_{1,lw}(A_q, q, M) &= 0 \\ S_{1,up}(A_q, q, M) &= \frac{A_q^M}{\prod_{n=1}^{M} (q+n)} \frac{A_q}{q + M - A_q} \end{aligned} \tag{10}$$

En (10), la cota inferior $S_{1,lw}(A_q, q, M)$ es obvia. La cota superior $S_{1,up}(A_q, q, M)$ se alcanza al aproximar los infinitos términos de $S_1(A_q, z, M)$ mediante un progresión geométrica, esto es, reemplazando $q+n$ por $q+M$ cuando $n > M$, siempre y cuando se cumpla que $A_q < q + M$.

Insertando $S_{1,lw}(A_q, q, M)$ y $S_{1,up}(A_q, q, M)$ en (6) y el resultado en (5) obtenemos sendas cotas, inferior y superior para $P_0^{(-3)}$, respectivamente denotadas por $P_{0,lw}^{(-3)}$ y $P_{0,up}^{(-3)}$. Esto es,

$$P_{0,lw}^{(-3)} \leq P_0^{(-3)} \leq P_{0,up}^{(-3)}$$

con lo cual,

$$P_{0,lw} \leq P_0 \leq P_{0,up}$$

siendo,

$$\begin{aligned} P_{0,lw} &= \left[P_0^{(-1)} + P_0^{(-2)} + P_{0,up}^{(-3)}\right]^{-1} \\ P_{0,up} &= \left[P_0^{(-1)} + P_0^{(-2)} + P_{0,lw}^{(-3)}\right]^{-1} \end{aligned} \tag{11}$$

V. Parámetros de prestaciones

En la presente sección y para el esquema 2, obtenemos cotas inferior y superior para los principales parámetros analizados; esto es, para la probabilidad de bloqueo -pérdida- de llamadas nuevas, P_P, para la probabilidad de traspaso fallido, P_{fh}, para la probabilidad de terminación forzosa, P_{FT} y para la probabilidad de no finalización, P_{NC}. Todas las cotas vienen dadas en función de las cotas de P_0 obtenidas en la sección IV-A.

A. La probabilidad de bloqueo -pérdida-

La probabilidad de pérdida para llamadas nuevas, P_P, viene dada por ((29) en [7]),

$$\begin{aligned} P_P &= \sum_{k=C-C_h}^{\infty} P_k = \sum_{k=C-C_h}^{C} P_k + \sum_{k=C+1}^{\infty} P_k = \\ &= (P_0^{(-2)} + P_0^{(-3)})P_0 \end{aligned} \tag{12}$$

la cual puede acotarse utilizando los resultados previos,

$$P_{P,lw} \leq P_P \leq P_{P,up} \tag{13}$$

con,

$$\begin{aligned} P_{P,lw} &= [P_0^{(-2)} + P_{0,lw}^{(-3)}(q)]P_{0,lw} = \\ &= \frac{P_0^{(-2)} + P_{0,lw}^{(-3)}}{P_0^{(-1)} + P_0^{(-2)} + P_{0,up}^{(-3)}} \\ P_{P,up} &= [P_0^{(-2)} + P_{0,up}^{(-3)}(q)]P_{0,up} = \\ &= \frac{P_0^{(-2)} + P_{0,up}^{(-3)}}{P_0^{(-1)} + P_0^{(-2)} + P_{0,lw}^{(-3)}} \end{aligned} \tag{14}$$

B. La probabilidad de traspaso fallido

Al solicitar el traspaso de una llamada en curso a una célula vecina, la probabilidad de que el intento resulte fallido, P_{fh}, puede expresarse como ((31) en [7]),

$$P_{fh} = \sum_{k=0}^{\infty} P_{C+k} P_{fh/k} \tag{15}$$

siendo $P_{fh/k}$ la probabilidad de que el intento de traspaso resulte fallido condicionado a que la solicitud del traspaso entrase en su cola de espera en la posición $k+1$, esto es, habiendo encontrado k llamadas en frente de nuestra llamada o sesión en observación -disciplina de servicio FIFO-. $P_{fh/k}$ viene dada por,

$$P_{fh/k} = \frac{(k+1)\mu_Q}{C\mu_H + (k+1)\mu_Q} = \frac{k+1}{q+k+1} \tag{16}$$

Insertando (1), (2) y (16) en (15), P_{fh} resulta ser,

$$\begin{aligned} P_{fh} &= \sum_{k=0}^{\infty} P_{C+k} P_{fh/k} = \sum_{k=1}^{\infty} \frac{k A_q^{k-1}}{\prod_{n=1}^{k}(q+n)} P_C = \\ &= \frac{A_t^{C-C_h} A_h^{C_h}}{C!} P_0 S_2(A_q, q) \end{aligned} \tag{17}$$

En (17) se ha definido,

$$\begin{aligned} S_2(A_q, q) &= \sum_{k=1}^{N} \frac{k A_q^{k-1}}{\prod_{n=1}^{k}(q+n)} + S_2(A_q, q, N) \\ &= \frac{d}{dA_q} S_1(A_q, q) \end{aligned} \tag{18}$$

$$S_2(A_q, q, N) = \sum_{k=N+1}^{\infty} \frac{k A_q^{k-1}}{\prod_{n=1}^{k}(q+n)}$$

Para acotar P_{fh}, Ec. (17), cabe encontrar cotas para $S_2(A_q, q, N)$, Ec. (18). Por ejemplo, resulta inmediato comprobar la siguiente relación entre $S_2(A_q, q, N)$ y $S_1(A_q, q, N)$,

$$\begin{aligned} S_2(A_q, q, N) &= \frac{q}{A_q}\left[S_1(A_q, q-1, N) - S_1(A_q, q, N)\right] \\ &= S_1(A_q, q, N-1) - \frac{q}{A_q} S_1(A_q, q, N) \end{aligned}$$

Obviamente, una cota inferior de $S_2(A_q, q, N)$ es cero, $S_{2,lw}(A_q, q, N) = 0$. Para obtener una cota superior, $S_{2,up}(A_q, q, N)$, podemos hacer uso de la Ec. (10) en la anterior relación. Alternativamente, $S_{2.up}(A_q, q, N)$ puede derivarse directamente a partir de la Ec. (18). En efecto, siempre que $A_q < (q + N)$, apelando a la expresión de la derivada de una progresión geométrica, y operando de manera similar a la obtención de las cotas en la Ec. (10), tras cierta álgebra se alcanza,

$$
\begin{aligned}
&S_{2,lw}(A_q,q,N)=0\\
&S_{2,up}(A_q,q,N)=\\
&=\frac{(N+1)A_q^N}{\prod_{n=1}^{N}(q+n)}\frac{(q+N)^{2N+1}}{\left[(q+N)^{N+1}-A_q^{N+1}\right]^2}
\end{aligned}
\tag{19}
$$

Insertando (19) en Ec. (18) y el resultado en Ec. (17) al tiempo que utilizamos $P_{0,lw}(q)$ y $P_{0,up}(q)$ de Ec. (11), obtenemos las cotas de P_{fh}, esto es,

$$P_{fh,lw}\leq P_{fh}\leq P_{fh,up} \tag{20}$$

con

$$
\begin{aligned}
P_{fh,lw}&=\frac{A_t^{C-C_h}A_h^{C_h}}{C!}P_{0,lw}S_{2,lw}(A_q,q)\\
P_{fh,up}&=\frac{A_t^{C-C_h}A_h^{C_h}}{C!}P_{0,up}S_{2,up}(A_q,q)
\end{aligned}
\tag{21}
$$

En la práctica, eligiendo valores de M y N entre C y $2C$ hallamos cotas muy próximas entre sí, tal que la diferencia es prácticamente despreciable a efectos de ingeniería de dimensionado.

C. La probabilidad de terminación forzosa

Para su evaluación se supone que el sistema celular es homogéneo, equivalente a considerar todas las células de igual tamaño, con el mismo tráfico de origen (λ_f), mismas capacidades (C, C_h), con igual movilidad (μ_R, μ_F) y distribución uniforme de los TUs. Entonces, la probabilidad de terminación forzosa P_{FT} puede evaluarse según la expresión ((36) en [7]),

$$
\begin{aligned}
P_{FT}&=P_{hd}\sum_{k=1}^{\infty}[(1-P_{fh})P_{hd}]^{k-1}P_{fh}=\\
&=\frac{P_{hd}P_{fh}}{1-(1-P_{fh})P_{hd}}
\end{aligned}
\tag{22}
$$

siendo P_{hd} la probabilidad de solicitud de traspaso de una llamada en curso, parámetro que se define en la Tabla II. Claramente P_{FT} puede acotarse según,

$$
\begin{aligned}
P_{FT,lw}&=\frac{P_{hd}P_{fh,lw}}{1-(1-P_{fh,lw})P_{hd}}\\
P_{FT,up}&=\frac{P_{hd}P_{fh,up}}{1-(1-P_{fh,up})P_{hd}}
\end{aligned}
\tag{23}
$$

D. Probabilidad de pérdida más no finalización

Un parámetro importante es la probabilidad de no completarse una petición, P_{NC}. Es la fracción de llamadas que inicialmente se bloquean -pérdida- o que tras haber sido admitidas y con anterioridad a su finalización, quedan interrumpidas debido a algún traspaso fallido. Se expresa como,

$$P_{NC}=P_P+(1-P_P)P_{FT} \tag{24}$$

En paralelo a anteriores derivaciones, las cotas inferior y superior asociadas a P_{NC} resultan ser, respectivamente,

$$
\begin{aligned}
P_{NC,lw}&=P_{P,lw}+(1-P_{P,up})P_{FT,lw}\\
P_{NC,up}&=P_{P,up}+(1-P_{P,lw})P_{FT,up}
\end{aligned}
\tag{25}
$$

VI. Ecuaciones de flujo

Las probabilidades de estado, Ec. (1), se refieren a una célula específica del conjunto que conforman la red celular, en las cuales la tasa de traspasos ofrecida, λ_h, es un parámetro a determinar, célula a célula, y que básicamente depende de la movilidad de los terminales. Al considerar una red celular, [17], resaltamos el acoplamiento entre células vecinas debido al flujo de traspasos, hecho que invalida la suposición de Poisson para tal flujo. No obstante, un buen número de estudios han analizado tal escenario, alcanzando la conclusión de que la suposición de Poisson para el flujo de traspasos puede ser válida a efectos de ingeniería, [12].

Suponiendo una red celular homogénea según se ha descrito y haciendo nuestras las hipótesis de [13] centramos el estudio en una única célula. Planteamos las ecuaciones de flujo, en donde la tasa de llamadas en curso en nuestra célula, $\gamma_{c,in}$ es la suma de dos contribuciones. La primera es la tasa de llamadas nuevas locales que son admitidas, $\gamma_{f,in}=\lambda_{f,in}(1-P_{P_{in}})$. La segunda contribución proviene de las llamadas en curso en células vecinas-colindantes que previo a su finalización solicitan traspaso a nuestra célula y son admitidas, $\gamma_{c,out}P_{hd,in}(1-P_{fh,in})$. Esto es,

$$
\begin{aligned}
&\gamma_{c,in}=\\
&=\lambda_{f,in}(1-P_{P,in})+\gamma_{c,out}P_{hd,in}(1-P_{fh,in})
\end{aligned}
\tag{26}
$$

En equilibrio tenemos que $\gamma_c=\gamma_{c,in}=\gamma_{c,out}$; $\lambda_f=\lambda_{f,in}=\lambda_{f,out}$; $P_{hd}=P_{hd,in}=P_{hd,out}$; $P_P=P_{P,in}=P_{P,out}$ y $P_{fh}=P_{fh,in}=P_{fh,out}$. Resolviendo (26) para la tasa total admitida -en curso-, γ_c, por lo tanto también para la tasa de traspasos ofrecida, λ_h,

$$\lambda_h=\gamma_cP_{hd}=\frac{\lambda_f(1-P_P)}{1-P_{hd}(1-P_{fh})}P_{hd} \tag{27}$$

En Ec. (27), λ_h puede acotarse haciendo uso de las cotas de P_P y de P_{fh} previamente obtenidas, esto es,

$$
\begin{aligned}
\lambda_{h,lw}&=\frac{\lambda_f(1-P_{P,up})}{1-P_{hd}(1-P_{fh,up})}P_{hd}\\
\lambda_{h,up}&=\frac{\lambda_f(1-P_{P,lw})}{1-P_{hd}(1-P_{fh,lw})}P_{hd}
\end{aligned}
\tag{28}
$$

Nótese que λ_h es una incógnita a resolver y junto con las probabilidades en régimen permanente, Ec. (1), conforman un sistema no lineal de ecuaciones. Para su resolución cabe utilizar el método de punto de iteración fija propuesto en [17], [18]. Por otra parte, a efectos de cómputo de los parámetros relevantes, Sec.V, observamos que λ_h influye en el numerador de $P_0^{(-1)}$, $P_0^{(-2)}$ y $P_0^{(-3)}$. Por lo tanto, insertando $\lambda_{h,lw}$ y $\lambda_{h,up}$ en la Ec. (3) y en la Ec. (4) obtenemos cotas inferior y superior de $P_0^{(-1)}$ y

de $P_0^{(-2)}$, respectivamente, $P_{0,lw}^{(-1)}$, $P_{0,lw}^{(-2)}$y $P_{0,up}^{(-1)}$, $P_{0,up}^{(-2)}$. Para $P_0^{(-3)}$ al insertar simultáneamente $S_{1,lw}(A_q, q, M)$ y $\lambda_{h,lw}$, respectivamente $S_{1,up}(A_q, q, M)$ y $\lambda_{h,up}$ en la Ec. (5) se obtienen las nuevas cotas, $P_{0,lw}^{(-3)'}$ y $P_{0,up}^{(-3)'}$. Por lo que las cotas de P_0 en la Ec. (11) resultan modificadas según indicamos,

$$\begin{aligned} P_{0,lw} &= \left[P_{0,up}^{(-1)} + P_{0,up}^{(-2)} + P_{0,up}^{(-3)'}\right]^{-1} \\ P_{0,up} &= \left[P_{0,lw}^{(-1)} + P_{0,lw}^{(-2)} + P_{0,lw}^{(-3)'}\right]^{-1}. \end{aligned} \tag{29}$$

y con estos nuevos valores evaluamos los parámetros de interés, esto es, las cotas de P_P, Ec. (14), las de P_{fh}, Ec. (21) y de nuevo las de λ_h, Ec. (28).

Así pues el procedimiento o método de punto de iteración fija que proponemos contiene los siguientes pasos,

1. Se conjetura un valor inicial para λ_h, p.e. $\lambda_h \approx \lambda_f P_{hd}$ (inicialmente bloqueo nulo en llamadas nuevas).
2. Se evalúan $P_0^{(-1)}$, Ec. (3), $P_0^{(-2)}$, Ec. (4), y $P_{0,lw}^{(-3)}$, $P_{0,up}^{(-3)}$, Ec. (5), (6), (10).
3. Se evalúan las cotas de P_P, Ec. (14) y de P_{fh}, Ec. (21).
4. Se actualiza λ_h evaluando $\lambda_{h,lw}$ y $\lambda_{h,up}$, Ec. (28).
5. Se calculan las cotas $P_{0,lw}^{(-1)}$, $P_{0,lw}^{(-2)}$, $P_{0,lw}^{(-3)'}$ y $P_{0,up}^{(-1)}$, $P_{0,up}^{(-2)}$, $P_{0,up}^{(-3)'}$ según se ha descrito anteriormente, para luego evaluar las cotas $P_{0,lw}$ y $P_{0,up}$ de la Ec. (29).
6. Tras la iteración i, se evalúan los valores absolutos $|\lambda_{h,lw}(i) - \lambda_{h,lw}(i-1)|$ y $|\lambda_{h,up}(i) - \lambda_{h,up}(i-1)|$. Si ambos valores resultan inferiores a cierto umbral dado, $\epsilon = 10^{-8}$, finaliza el proceso iterativo. Caso contrario ir al paso 3) con las nuevas cotas de $P_0^{(-1)}$, $P_0^{(-2)}$ y $P_0^{(-3)}$.

Tabla III
CASOS ANALIZADOS DEL ESQUEMA 2 DE LA TABLA I.

Parámetros	Caso 1	Caso 2	Caso 3	Caso 4
Radio celular (km)	0.1	0.8	3	5
$E(v)$ (km/h)	5-6	20-40	40-60	60-80
Tasa (s^{-1})				
λ_f	0.01-0.1	0.1-0.5	0.02-0.2	0.2-1
μ_M, llamada	1/120	1/120	1/120	1/120
μ_R, residencia	1/100	1/200	1/300	1/400
$\mu_H = \mu_M + \mu_R$	11/600	8/600	7/600	6.5/600
$\mu_F = 5 \times \mu_R$	5/100	5/200	5/300	5/400
$\mu_Q = \mu_M + \mu_F$	7/120	4/120	3/120	2.5/120
$q = C\mu_H/\mu_Q$	$11C/35$	$8C/20$	$7C/15$	$13C/25$
$A_{os} = \lambda_f/\mu_M$	1.2-12	12-60	2.4-24	24-120
Capacidades, C	5-20	20-75	7-35	35-138
C elegido	12	47	21	86
C_h elegido	1,3	1,3	1,3	1,3
Umbrales				
M, Ec. (6)	C	C	C	C
N, Ec. (18)	C	C	C	C

VII. MODELO DE MOVILIDAD

Cuanto menos es cuestionable la caracterización del tiempo de residencia de un móvil en una célula y/o en la zona de traspaso mediante una v.a.e. Al respecto se han aportado diversos estudios sobre la conveniencia de tal caracterización según tamaños y geometrías celulares varias, entre otros [7], [19], [20], [21], [22]. De ellos, una primera conclusión es que la v.a.e. puede ser una buena aproximación para macrocélulas, [7], [19], dudando de su validez a medida que el área de la zona de residencia se reduce, caso de micro y picocélulas, [20], [21], [22]. Decir que tales tiempos de residencia cabría caracterizarlos mediante distribuciones *phase type* (PH), [10], preservando así la ventaja de operar con herramientas de Markov. Si bien los errores de ajuste se reducen incrementando el número de estados de la distribución PH, el agravante es el crecimiento exponencial de la dimensión del proceso de Markov asociado. Por el momento, en el presente estudio y debido a su sencillo manejo, adoptamos la suposición de v.a.e. con valor medio evaluado según el modelo de movilidad de fluidos de G. Moralles y M.Villén, [23]. Por lo tanto, a partir de las Ec. (12)-(13) en [24], estimamos los parámetros μ_R y μ_F. Esto es,

$$\mu_x = \frac{E(v)L_x}{\pi A_x}; \quad x = R, F \tag{30}$$

siendo $E(v)$ la velocidad media de los móviles y, L_x (A_x) el perímetro (el área) de la zona de cobertura en consideración; $x = R$ para la zona celular y $x = F$ para la zona de traspaso. La Tabla III muestra cuatro casos que se consideran más adelante. Los dos primeros se asocian a zonas urbanas con velocidades respectivas, $E(v)$, entre 5-6 y 20-40 km/h, y los casos 3 y 4 a zonas interurbanas y/o rurales con velocidades en torno a 40-60 y 60-80 km/h, respectivamente. Insertando $E(v)$ en la Ec. (30) obtenemos las tasas μ_R habiéndose elegido un valor intermedio y tras practicar un redondeo que facilite la lectura visual y su comprensión. En cuanto a la zona de traspaso, asumimos un mosaico hexagonal con circunferencias circunscritas a cada hexágono, según se muestra en la Fig. 1. El cálculo de los perímetros y las áreas, comunes o de solape (traspaso) para μ_F, y no comunes para μ_R, nos revela que $\mu_F \approx 4{,}78\mu_R$ habiendo redondeado a $\mu_F = 5\mu_R$.

VIII. RESULTADOS

Los cuatro esquemas de la Tabla I fueron presentados y analizados en [16]. En cuanto al esquema 2, objeto de la presente contribución, hemos verificado la bondad o calidad de las nuevas cotas derivadas en la Sec. IV, para diferentes rangos de M y de N. Nuestra contribución conlleva ciertas mejoras con respecto a lo ofrecido en [16]. En particular, identificamos $C = 12$, $C_h = 1$, $A_{os} = 5$ Erlangs siendo los demás parámetros los del Caso 1 de la Tabla III. En la Tabla IV podemos apreciar que con el incremento de $M = N$ las cotas inferiores y superiores, respectivamente crecen y decrecen, estando muy próximas entre sí para valores de $M = N$ sustancialmente inferiores

Tabla IV
COTAS DE λ_h, P_P Y P_{NC}; $C = 12, C_h = 1, A_{os} = 5$.

$M = N$	cota	λ_h	P_P	P_{NC}
	$upper(10^{-3})$	49,9074	2,4121	3,4020
0	$lower(10^{-3})$	49,8311	1,8518	1,8518
# iter=5	$error, \%$	0,1518	30,2557	83,7080
	$upper(10^{-3})$	49,8673	2,2635	2,9077
1	$lower(10^{-3})$	49,8548	2,1883	2,6538
# iter=5	$error, \%$	0,0248	3,4378	9,5686
	$upper(10^{-3})$	49,8579	2,2468	2,8810
2	$lower(10^{-3})$	49,8560	2,2379	2,8411
# iter=5	$error, \%$	0,0039	0,4008	1,4043
	$upper(10^{-3})$	49,8563	2,2450	2,8788
3	$lower(10^{-3})$	49,8561	2,2440	2,8734
# iter=5	$error, \%$	0,0005	0,0424	0,1865
	$upper(10^{-3})$	49,8561	2,2448	2,8786
4	$lower(10^{-3})$	49,8561	2,2444	2,8778
# iter=5	$error, \%$	≈ 0	0,0040	0,0214

al número de canales C. El error relativo se ha definido como $error = (upper - lower)/lower$. Para valores de M, N entre C y $2C$ la diferencia entre cotas no es significativa.

Para el análisis de los cuatro casos de la Tabla III, hemos dado cierto margen de holgura al suponer $M = N = C$. Fijado el número total de canales C y el de priorización C_h, se ha observado el impacto en los parámetros de prestaciones indicados al variar la tasa ofrecida λ_f. A partir de $A_{os} = \lambda_f/\mu_M$ podemos hacer una primera estimación de C mediante el uso de la fórmula de Erlang-B[2]. Por ejemplo, a un valor de $P_P = 10^{-2}$ y para el rango de valores de λ_f indicado, el rango de canales correspondientes es de, respectivamente, 5-20, 20-75, 7-35 y 35-138 para los casos 1, 2, 3 y 4. La fila C en la referida Tabla III muestra el valor intermedio escogido.

La Fig. 3 recoge los resultados del caso 1. En primer lugar vemos que la probabilidad P_P, Ec. (13), de las llamadas nuevas, símbolo "+", ligeramente se incrementa con el término $(1 - P_P)P_{fh}$, llamadas que siendo admitidas sufren un traspaso fallido. La suma de ambos conforman la P_{NC}, símbolo "$\diamond$", siendo P_P mucho más significativo que el segundo, ello consecuencia del algoritmo GCA. En segundo lugar, la probabilidad de traspaso fallido, P_{fh} con símbolo $*$, y la de terminación forzosa, P_{FT} con símbolo "$\triangle$", ambas son menores que las dos anteriores, también consecuencia de la priorización, diferencias que se incrementan a medida que lo hace C_h, ello a costa de empeorar P_{NC}. Por otra parte, un incremento de C_h supone un mejor tratamiento de las llamadas en curso pero a costa de penalizar las de origen, ya que P_P se incrementa. El resultado neto es el empeoramiento de la probabilidad P_{NC} al aumentar C_h. Similares resultados se han obtenido para los demás casos analizados, caso 2, 3 y 4, respectivamente mostrados en las Fig. 4, 5 y 6. En otras palabras, considerando un Grado de Servicio GoS tal que $P_{NC} < 10^{-2}$, una inspección ocular de las gráficas nos dice que el máximo tráfico de origen (en Erlangs) y por célula que cada escenario puede soportar resulta ser, considerando $C_h = (1, 3)$, para el caso 1 con $C = 12$, $A_{os} < (5{,}8, 4{,}4)$; para el caso 2 con $C = 47$, $A_{os} < (37, 35)$; para el caso 3 con $C = 21$, $A_{os} < (14, 12{,}5)$; y para el caso 4 con $C = 86$, $A_{os} < (79, 77)$ respectivamente.

[2]Debido a la movilidad, la tasa original ofrecida a una célula, λ_f viene incrementada por la tasa de traspasos, λ_h, esto es, $\lambda_t = \lambda_f + \lambda_h$. Por otra parte la tasa de ocupación de un canal también se incremente, dado que $\mu_H = \mu_M + \mu_R$. Al inicio se desconoce λ_h por lo que asumimos una compensación mutua provisional entre ambos incrementos tal que $\lambda_t/\mu_H \approx= \lambda_f/\mu_M$. En una nueva iteración, cabe modificar (aumentar o disminuir) el valor inicial obtenido de C si los resultados de los parámetros de diseño, $P_P, P_{FT}, \ldots$, no son satisfactorios.

IX. CONCLUSIONES

Se han revisado algunos algoritmos de traspasos de llamadas/sesiones en sistemas radio celulares. Con formulación Markoviana, para uno de ellos se han aportado cotas de los parámetros más relevantes (probabilidad de bloque, de traspaso fallido, de terminación forzosa y de no finalización). Las cotas son de fácil cálculo resultando arbitrariamente próximas, suponiendo una ayuda en una primera aproximación al dimensionado de los recursos de radio en redes celulares. Como parte del trabajo futuro, se piensa en aplicar las aproximaciones aquí obtenidas en el dimensionado de escenarios heterogéneos. Adicionalmente se piensa abordar la obtención de cotas similares para los esquemas 3 y 4 de la Tabla I.

AGRADECIMIENTOS

El presente trabajo es parte del proyecto PGC2018-094151-B-I00. Los autores agradecen la financiación recibida del *Ministerio de Ciencia, Innovación y Universidades (MCIU)*, de la *Agencia Estatal de Investigación (AEI)*, del *Fondo Europeo de Desarrollo Regional (FEDER)* (MCIU/AEI/FEDER.UE) y del *Instituto Valenciano de la Competitividad Empresarial (IVACE)*.

REFERENCIAS

[1] V. Casares-Giner, J. Martínez-Bauset and X. Ge, "Performance model for two-tier mobile wireless networks with macrocells and small cells," *Wireless Networks*, vol. 24, n. 4, pp. 1327-1342, May 2018.

[2] "Qualcomm Circuit-switched fallback. The first phase of voice evolution for mobile LTE devices," white paper, 2012.

[3] R. A. Khan and A. H. Mir "Performance analysis of host based and network based IP mobility management schemes over IPv6 network," IEEE-ICACCI, pp. 1798-1803, November 2014.

[4] R. Guerin, "Queueing blocking system with two arrival stream and guard channels," *IEEE Trans. on Communications*, vol. 36, n. 2, pp. 153-163, February 1988.

[5] V. Casares-Giner, "Integration of dispatch and interconnect traffic in a land mobile trunking system. Waiting time distributions," *Telecommunication Systems*, vol. 16, n. 3-4, pp. 539-554, 2001.

[6] C. J. Powell and V. C. M. Leung, "Traffic engineering for integrated telephone and dispatch land mobile radio traffic,"in *ICWC'92*, pp. 168-171, June 1992.

[7] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone sysems with prioritized and nonprioritized handoff procedures," *IEEE Trans. on Vehicular Technology*, vol. 35, n. 3, pp. 77-92, 1986. (Also in Technical Report No.773, version 2a, Department of Electrical and Computer Engineering, State University of New York. June, 1999.

[8] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," *IEEE/ACM Trans. Networking*, vol. 2, pp. 166-175, April 1994.

[9] J. N. Daigle and N. Jain, "A queueing system with two arrival streams and reserved servers with application to cellular telephone," in *INFOCOM'92*, pp. 2161-2167, June 1992.

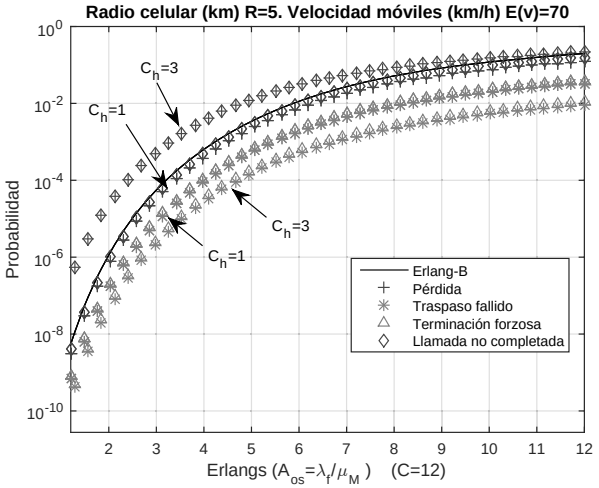


Figura 3. Caso 1 de la Tabla III.

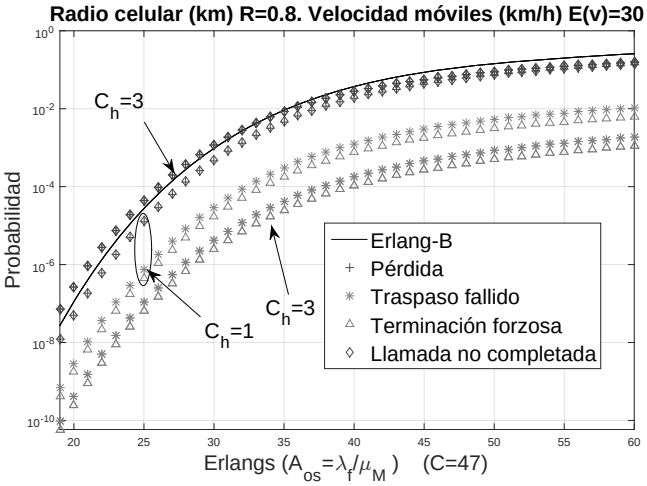


Figura 4. Caso 2 de la Tabla III.

[10] M. F. Neuts, "Matrix geometric solution in stochastic models". The Johns Hopinks University Press, Baltimore, 1981.

[11] V. Casares-Giner and J. Martínez-Bauset, "Approximate analytical model for queued handover requests in cellular networks," in *UBICOMM'19*, September 2019.

[12] E. Chlebus and W. Ludwin, "Is handoff traffic really Poissonian?," *IEEE ICUPC*, pp. 348-353, Nov. 1995.

[13] P. V. Orlik and S. S. Rappaport, "On the handover arrival process in cellular communications," *ACM/Baltzer Wireless Networks*, vol. 7, no. 2, pp. 1477-157, March/April 2001.

[14] C. Jedrzycki and V. C. M. Leung, "Probability distribution of channel holding time in cellular telephony systems," in Proc. *IEEE Veh. Technol. Conf. (VTC'96)*, vol. 1, pp. 247-251, 1996.

[15] F. Barceló and J. Jordan, "Channel holding time distribution in public telephony systems," *IEEE Tras. Veh. Technol.*, vol. 49, pp. 1615-1625, September 2000.

[16] D. Sáez Domingo, "Mecanismos de gestión de recursos en redes de comunicaciones móviles celulares", Proyecto Fin de Carrera. ETSIT-UPV, Septiembre de 2001.

[17] D. McMillan, "Traffic modelling and analysis fo cellular mobile networks," in *ITC-13*, vol. 3 (North Holland), pp. 627-632, June 1991.

[18] F. P. Kelly, "Fixed point models of loss networks," *J. Austral. Math. Soc.* Ser. B, vol. 31, pp. 204-218, 1989.

[19] R. A. Guerin, "Channel occupancy time distribution in a cellular radio system," *IEEE Trans. Veh. Technol.*, vol. 35, pp. 89-99, August 1987.

[20] M. M. Zonoozi and P. Dassanayake, "User mobility modeling and characterization of mobility Patterns," *IEEE JSAC*, vol. 15, n. 7, pp. 1239-1252, September 1997.

[21] V. Pla and V. Casares-Giner, "Analytical-numerical study of the handoff area sojourn time," in Proc. *Globecom'02*, vol. 1, pp. 886-890, December 2002.

[22] M. Schweigel, "The cell residence time in rectangular cells and its exponential approximation," in Proc. *ITC 18th. Teletraffic Science and Engineering*, Vol.5, pp. 761-770, Elsevier. Berlin, August 31-September 5, 2003.

[23] G. Morales-Andres, M. Villen-Altamirano, "An approach to modelling subscriber mobility in cellular radio networks," in Proc. *Forum Telecom 1987*, pp. 185-189, Geneva, Switzerland, December 1987.

[24] V. Casares-Giner, V. Pla and P. Escalle-García, "Mobility models for mobility management," *Springer-Verlag*, D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 716-745, 2011.

[25] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Pers. Commun.*, vol. 3, pp. 4-9, December 1996.

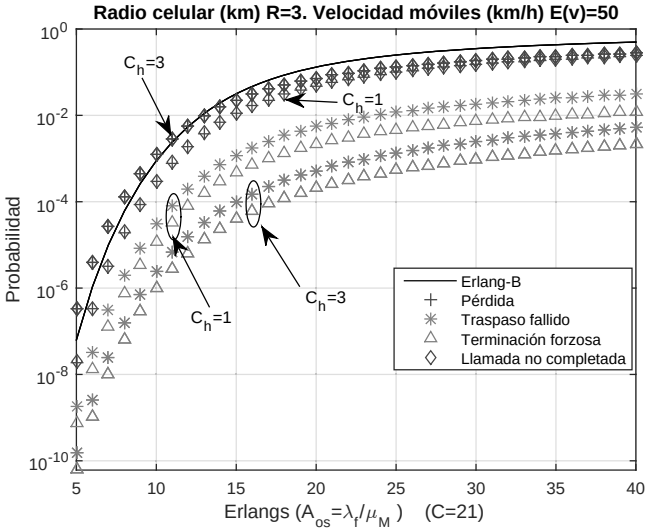


Figura 5. Caso 3 de la Tabla III.

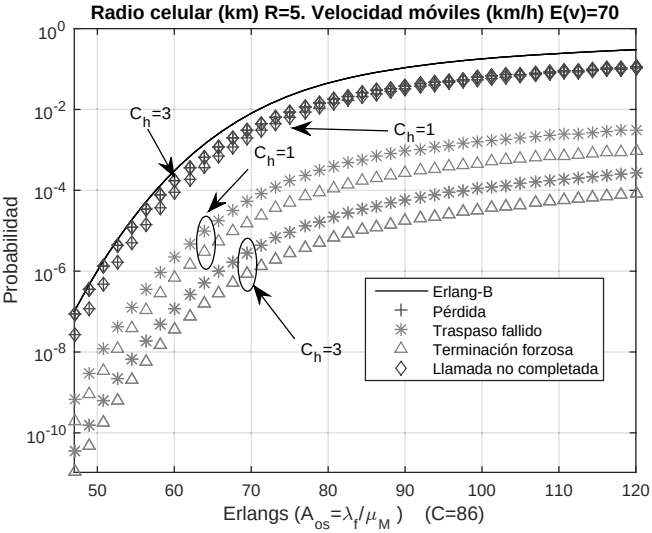


Figura 6. Caso 4 de la Tabla III.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Evaluación automática de la QoE del streaming DASH utilizando el estándar ITU-T P.1203 y Google Puppeteer

Paola Guzmán Castillo, Pau Arce Vila, Juan C. Guerri Cebollada
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*- Este documento presenta un sistema de evaluación de sistemas DASH que permite realizar medidas de prestaciones de forma automatizada y sistemática. Mediante el uso de Puppeteer, la librería en Node.js desarrollada por Google que proporciona una API de alto nivel que permite automatizar acciones sobre Chrome Devtools Protocol, se pueden automatizar acciones como iniciar la reproducción, realizar cambios de ancho de banda y guardar resultados de los procesos de cambios de calidad, instantes de tiempo, etc. A partir de dichos datos se realiza un procesado para permitir la reconstrucción del vídeo visualizado, así como una evaluación subjetiva utilizando el estándar ITU-T P.1203.**

***Palabras Clave*- DASH, Puppeteer, Video Streaming, Streaming adaptativo, QoE.**

I. INTRODUCCIÓN

La distribución de contenidos multimedia, y en particular el *streaming* de vídeo, domina actualmente el tráfico global de Internet, y se prevé que, a nivel mundial, el consumo de tráfico de vídeo por Internet crecerá 4,3 veces de 2017 a 2022, una tasa de crecimiento anual del 34% [1].

Poder ofrecer la mejor calidad posible al usuario en todo momento, maximizando su calidad de experiencia QoE (*Quality of Experience*), ha dado origen al *streaming* adaptativo de vídeo sobre HTTP (HAS), siendo el estándar DASH [2] (*Dynamic Adaptive Streaming over HTTP*) el más representativo.

HAS permite una adaptación flexible de la calidad de vídeo a los recursos de red disponibles y las capacidades del dispositivo cliente. De este modo, permite una mejor gestión del estado del buffer, el control de las interrupciones durante la reproducción y una mejor utilización del ancho de banda, lo que generalmente redunda en una mayor QoE. Para aplicar HAS, los servidores de contenidos ofrecen varias versiones (representaciones) del mismo vídeo, cada una codificada y divida en pequeños segmentos (*chunks*) de unos pocos segundos. Toda la información asociada a los segmentos de vídeo, como la resolución, duración y la tasa de bits promedio, se encuentra especificada en la MPD (*Media Presentation Description*). De esta forma, los clientes descargan los segmentos en un orden secuencial y pueden ir cambiando de representación de un segmento a otro según el ancho de banda actual estimado y/o el estado del buffer, de modo que se eviten las paradas y se pueda utilizar de forma óptima el ancho de banda disponible.

La tecnología HAS ha sido adoptada por una amplia gama de aplicaciones y proveedores de contenido de vídeo, como *YouTube* [3] o *Netflix*. Por tal motivo, en los últimos años han aparecido muchas publicaciones relacionadas con DASH y su impacto en la QoE del usuario en diferentes contextos [4][5][6]. El objetivo en muchos casos, es la optimización del algoritmo de adaptación [7] o su valoración en base a los resultados de la evaluación subjetiva de los usuarios [8].

El desarrollo de este tipo de estudios involucra las siguientes etapas según se muestra en la Fig. 1. En primer lugar, identificamos una etapa de pre-procesado de los datos asociada con la codificación y generación de los segmentos DASH que van a estar disponibles en el servidor. El siguiente bloque corresponde a la emulación de la red y terminales cliente. En este punto se definen las condiciones de ancho de banda, retardos, pérdidas y prestaciones de los dispositivos que van a recibir y reproducir los contenidos de vídeo. El auge del *streaming* adaptativo de vídeo ha llevado al desarrollo de diferentes implementaciones de DASH. En este sentido la selección del reproductor de vídeo teniendo en cuenta aspectos como: formatos soportados, algoritmo de adaptación, código abierto o propietario, etc. es otra etapa importante en este proceso. Una vez realizada la emulación del *streaming* de vídeo, nos encontramos con una etapa de post-procesado

orientada al análisis de la información obtenida y la reconstrucción del vídeo para su posterior evaluación subjetiva, lo que constituye la última etapa del proceso.

Hoy en día, existen una gran cantidad de reproductores DASH y algoritmos de adaptación, y la mayoría de ellos han sido implementados en JavaScript. Sin embargo, se encuentran pocos trabajos orientados a la implementación de un sistema común para la realización de pruebas de desempeño de forma fácil y accesible [9]. Este trabajo tiene como objetivo desarrollar un sistema fácilmente exportable, reproducible y escalable que permita automatizar y sistematizar la realización de pruebas de evaluación de calidad de experiencia (QoE) en un escenario de transmisión adaptativa de vídeo, evaluando diferentes perfiles de variación de ancho de banda y/o latencia.

El sistema propuesto se orienta a la ejecución del reproductor DASH, usando el navegador *Google Chrome*, y en concreto utilizando *Puppeteer*, la nueva librería desarrollado por Google que permite la automatización de pruebas funcionales en entornos web, mediante el acceso a las herramientas de desarrollador ofrecidas por la aplicación.

En la sección II se presenta la arquitectura del sistema de pruebas y los componentes que lo conforman. En la sección III se describen algunos casos de uso en los que puede ser aplicado el sistema y, en la sección IV, se exponen las conclusiones del trabajo.

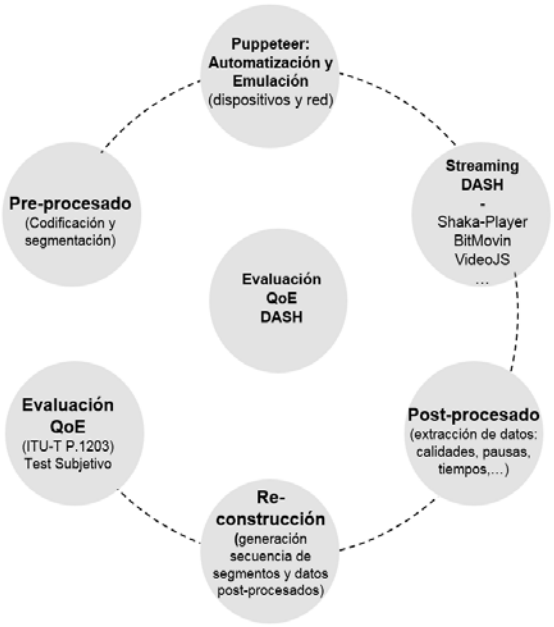


Fig. 1. Etapas del proceso de evaluación automática.

II. Arquitectura del Sistema de Pruebas (Testbed)

El sistema propuesto incluye los siguientes componentes:

- Servidor web basado en HTTP que aloja los contenidos pre-procesados (vídeo codificado y segmentado).
- Herramienta para la emulación de condiciones de red (variaciones de ancho de banda y/o latencia)
- Cliente con el reproductor DASH.
- Herramientas para la obtención y post-procesado de los datos relativos a la transmisión.
- Algoritmo para la evaluación subjetiva de la calidad de experiencia (QoE).

La arquitectura del sistema de pruebas se muestra en la Fig. 2 y los detalles de cada uno de los componentes se describen en las siguientes subsecciones.

Los bloques del sistema propuesto, tanto el pre-procesado como el servidor o el cliente, se han desarrollado en un PC con *Ubuntu* (*versión 18.04.2 LTS*). Los bloques son modulares y pueden ejecutarse en una misma máquina. Así, podría utilizarse un sistema de virtualización o contenedores, como Docker, para poder hacer el despliegue de cada uno de los módulos en un mismo ordenador. El hecho de implementar todo el proceso de *testbed* en un único dispositivo permite que el sistema desarrollado sea fácilmente exportable y replicable por la comunidad científica.

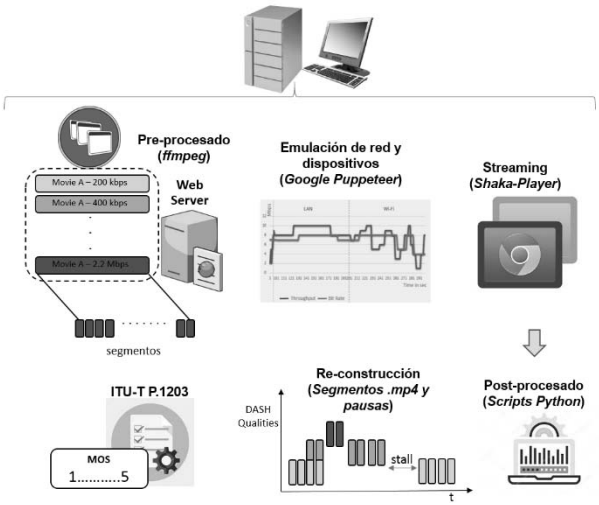


Fig. 2. Arquitectura del sistema.

A. Servidor Web

El servidor web aloja el contenido de vídeo para la transmisión adaptativa a través de HTTP. Como paso previo a la generación de los contenidos DASH se debe realizar un proceso de codificación, que en nuestro caso se realiza empleando la *librería libx264* de la aplicación *ffmpeg* y la tasa de bits máxima como parámetro de codificación. El sistema está abierto a otros codificadores y cualquier otra opción de codificación, incluyendo otros parámetros de codificación, como por ejemplo basado en QP (*Quantization Parameter*) o CRF (*Constant Rate Factor*) en lugar de bitrate. El uso de otros codificadores como (HEVC, VP9, AV1), está sujeto a la compatibilidad con el navegador web, el reproductor DASH y la implementación del estándar para la evaluación subjetiva.

Una vez codificadas las secuencias de vídeo, se procede a la generación de los segmentos DASH y la MPD, que contiene toda la información sobre las diferentes calidades de vídeo usadas y los anchos de banda de cada una. En la actualidad estamos trabajando con MPEG-DASH como formato de entrega, aunque también está abierta la opción de usar *HTTP Live Streaming* (HLS), el protocolo de *streaming* multimedia basado en HTTP implementado por Apple.

B. Emulación de condiciones de red y terminales cliente (Puppeteer)

El *streaming* adaptativo de vídeo se puede dar en entornos heterogéneos, y un solo cambio en las condiciones de contexto

puede tener un gran impacto en el comportamiento del reproductor y, seguramente, en la experiencia de visualización del usuario final. Como herramienta para la automatización de las pruebas extremo a extremo, incluida la emulación de las variaciones de red se empleará *Puppeteer* [10]. *Puppeteer* es la nueva librería desarrollada por Google que ofrece una interfaz basada en node.js que permite ejecutar y controlar Chrome (o Chromium) en modo *headless* a través del protocolo DevTools mediante la ejecución de un *script* desde la línea de comandos. En concreto, para esta etapa se establece una sesión CDP (*Chrome Devtools Protocol*) con la web en la que se encuentra alojada la implementación del *Shaka Player*. Mediante el acceso a los recursos `Network.emulateNetworkConditions` y `Emulation.setCPUThrottlingRate` se proporcionan los parámetros necesarios para la activación del *throttling* y la definición de las condiciones de la CPU del cliente. Dentro de las condiciones de red que se deben definir se encuentran: *downloadThroughput* (bytes/s), *uploadThroughput* (bytes/s) y *latency* (ms), que serán seleccionadas según las condiciones de contexto deseadas (Fig. 3). Cabe resaltar que, la versión actual de *Puppeteer* no permite emular entornos de red complejos. Por tal motivo, esta primera versión del sistema de pruebas, se centra en la evaluación de la calidad a partir de la emulación de variaciones del ancho de banda (cambios rápidos, cambios lentos, escalonados, etc.) y/o la latencia del enlace.

```
const client = await page.target().createCDPSession()
  await client.send('Network.emulateNetworkConditions', {
    'offline': false,
    'downloadThroughput': 6 * 1024 * 1024 / 8,
    'uploadThroughput': 3 * 1024 * 1024 / 8,
    'latency': 5
  });
```

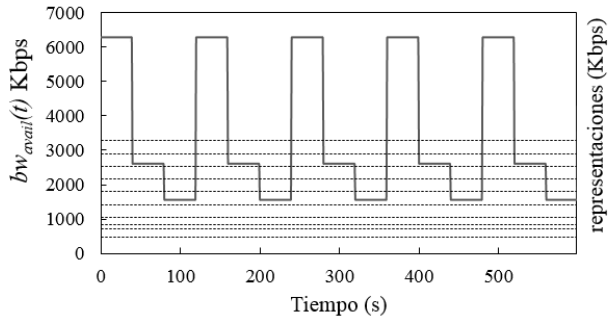


Fig. 3. Ejemplo de la variación de ancho de banda durante la reproducción del vídeo.

C. Cliente reproductor DASH (Shaka Player)

mpatibles con el estándar DASH disponibles actualmente, para el desarrollo de este trabajo hemos seleccionado el reproductor *Shaka Player* [11]. *Shaka Player* es una librería de código abierto de JavaScript que permite la reproducción de contenidos multimedia tanto en formato DASH como HLS en un navegador estándar, sin requerir el uso de *plugins* o Flash. El player debe alojarse en un servidor web, en este caso es el ordenador local. Para esto, se ha desarrollado un script en node.js que, mediante el uso de *Puppeteer*, permite acceder a la web donde está el *Shaka Player*, seleccionar el vídeo, activar el log por consola, donde se recogerán los datos para el post-procesado, e iniciar la reproducción del vídeo haciendo clic programáticamente en el botón correspondiente.

La Fig. 4 muestra una instantánea de lo que el usuario vería al ejecutar el sistema y deshabilitar el modo headless. Se muestra un mensaje en la esquina superior izquierda de la pantalla, que indica que un software de prueba automatizado está controlando Chrome. El marco derecho de la pantalla muestra que el acceso a las DevTools está activo, lo que permite tener acceso al *bandwidth throttling* y las estadísticas de red.

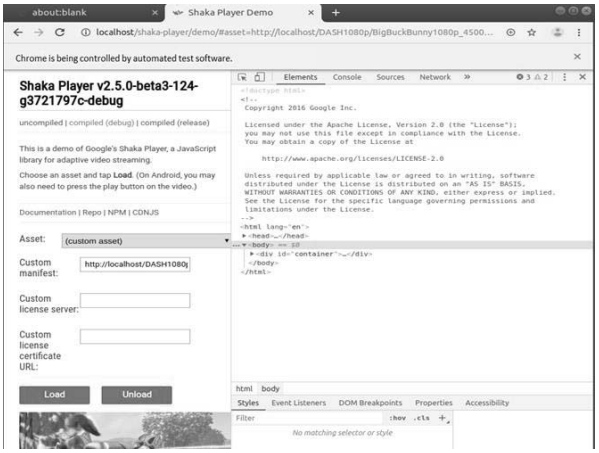


Fig. 4. Navegador Chrome controlado por *Puppeteer* Modo Headless deshabilitado.

D. Post-procesado

La posibilidad de tener acceso mediante *Puppeteer* a todos los recursos o herramientas de desarrollador orientadas a la evaluación del desempeño de los servicios web permite que, una vez iniciada la reproducción del vídeo, se inicie también la captura de métricas y registros relacionados con: estadísticas de la red, estado del buffer, número de paradas, duración de paradas, tiempo de reproducción y representaciones transmitidas, entre otros. Esta información se extrae empleando métodos como `page.setRequestInterception` o `page.evaluate`, que permite definir un *callback* en el que podemos especificar mediante código los elementos de la página que resultan de interés. Estos datos se obtienen mediante su presentación por consola o la generación de un archivo en formato JSON que registra la interacción entre el cliente y el servidor. Estos archivos son procesados posteriormente para extraer los datos requeridos para la reconstrucción del vídeo (segmentos descargados, número y duración de las paradas). La información obtenida por el sistema permite analizar aspectos como: el estado del buffer (Fig. 5), el comportamiento del algoritmo de adaptación frente a los cambios de ancho de banda (Fig. 6), entre otros.

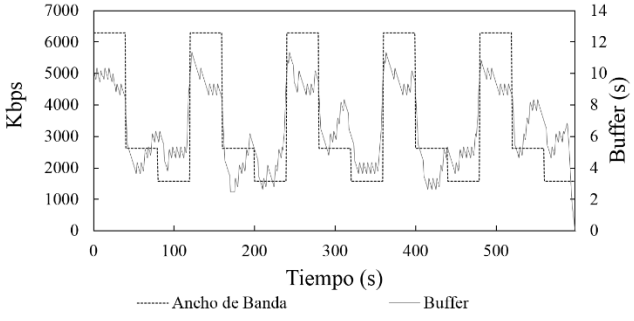


Fig. 5. Ancho de banda disponible y estado del buffer.

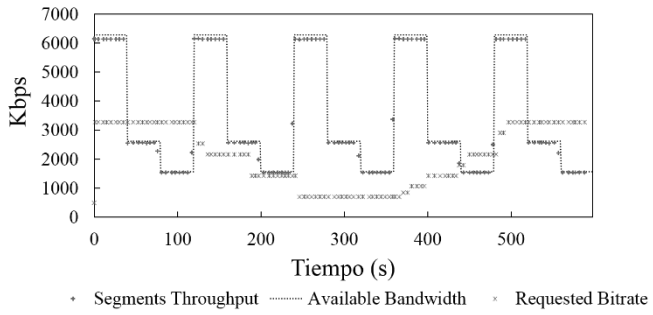


Fig. 6. Ancho de banda disponible, bitrate solicitado y throughput por segmento.

E. Evaluación de la calidad de experiencia (QoE)

Mientras que la capacidad de cambiar alternativamente entre representaciones de un vídeo reduce significativamente el riesgo de paradas, las variaciones de calidad que esto conlleva pueden también resultar molestas para el usuario. Por tal motivo, conocer la QoE percibida por el usuario es una pieza clave para evaluar el desempeño de los algoritmos de *streaming* adaptativo. En esta etapa, usaremos una implementación en Python del estándar ITU-T Rec. P.1203 [12]. Esta recomendación fue publicada en 2017 y se ha convertido en el primer modelo estandarizado para la evaluación de la calidad de los servicios de *streaming* adaptativo de audio y vídeo.

La Fig. 7 presenta un diagrama de la arquitectura general de la recomendación P.1203. La P.1203 consta de tres módulos, uno para la estimación de la calidad del vídeo (P.1203, Pv), otro para la estimación de la calidad del audio (P.1203, Pa) y uno más que proporciona una percepción global de la calidad (P.1203.3, Pq). Asimismo, la herramienta tiene 4 modos de operación, desde el modo 0 hasta el modo 3. Los modos se distinguen según la cantidad de información disponible, que va desde sólo los metadatos (códec, resolución, tasa de bits, tasa de frames, resolución del display, duración del segmento) en modo 0, hasta el acceso al flujo de bits completo en el modo 3.

En este trabajo nos centramos en el modo 3 y en la salida O.46, que corresponde con la evaluación integral del Mean Opinion Scale (MOS). La recomendación P.1203 tiene en cuenta la información sobre las paradas para la predicción de la calidad y proporciona un valor en una escala de 1 a 5 por cada segundo del vídeo. Como limitación tenemos, que la implementación disponible actualmente de la recomendación P.1203, solo soporta vídeos codificados en H264 con una resolución de hasta 1080p.

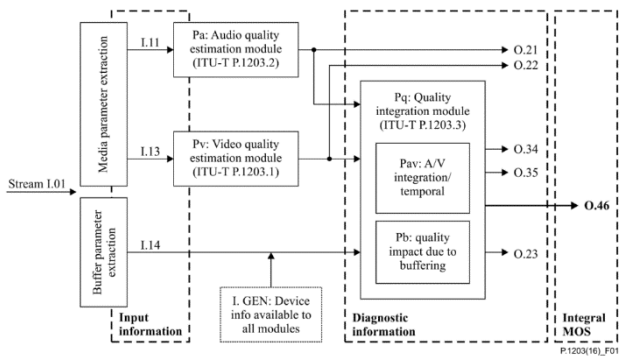


Fig. 7. Diagrama de bloques del modelo ITU-TP.1203.

III. CASOS DE USO

A. Video 2D – Modificación del algoritmo de adaptación del Shaka Player

Los resultados de la evaluación subjetiva de la calidad de experiencia de los usuarios frente a los diferentes escenarios de variación de ancho de banda (cambios rápidos, cambios lentos, cambios bruscos, etc.) permiten evaluar el desempeño del reproductor (*Shaka Player*) y se usan como realimentación del sistema para inferir qué modificaciones en el algoritmo de adaptación serían deseables. El objetivo es validar, entre otros, la respuesta del usuario frente a escenarios donde, con el objetivo de mantener la calidad en un nivel aceptable, se admitan las paradas, o si reacciona mejor a un cambio (degradación) brusco de la calidad para evitar una parada.

B. Vídeo 3D - Evaluación de la QoE de vídeo 3D

El uso de un reproductor web y la utilización de formatos estereoscópicos compatibles con 2D (Frame-compatible o Full-resolution Frame-compatible) hace posible que el sistema propuesto pueda ser extendido para la evaluación de la QoE de vídeo 3D, donde las diferentes representaciones disponibles del vídeo pueden ser obtenidas, además de mediante las variaciones de calidad y resolución espacial, introduciendo criterios como la asimetría entre vistas, que explotan las propiedades de adaptación del sistema de visión humano y es un factor de gran interés en este contexto.

IV. CONCLUSIONES

La evaluación de la QoE en sistemas de DASH incluye muchos aspectos que, en general, no hacen viable la réplica de los experimentos y complica la posibilidad de comparar soluciones o mejoras. Con el objetivo de contribuir en el desarrollo de un testbed que permita automatizar, escalar, replicar y simplificar el proceso de evaluación, se ha desarrollado un sistema que incluye, desde la descripción del proceso de codificación hasta el de la evaluación subjetiva utilizando el estándar ITU-T P.1203. Para el proceso de la emulación de la red y automatización de las pruebas, se ha utilizado la herramienta *Puppeteer* de Google y las opciones que ofrece Chrome DevTools. Además, se ha desarrollado el código que permite extraer la información necesaria en el cliente para la reconstrucción del vídeo visualizado y su adaptación para que pueda ser evaluado con el estándar P.1203. Con el uso de la tecnología Docker como parte del trabajo futuro, se pretende incluir todas las herramientas implementadas en contenedores; esto facilitará la réplica del sistema por parte de terceros. Sin duda, el sistema evolucionará y se ampliará conforme se incorporen nuevos codificadores (como AV1, etc.) tanto en el proceso de codificación como en la reproducción y en la evaluación según el estándar; se desarrollen nuevos reproductores; o se propongan diferentes algoritmos de adaptación. Sin embargo, todas estas nuevas propuestas no suponen cambios en todos los módulos sino la adaptación de manera independiente conforme sea necesario ante la evolución de la tecnología de *streaming*.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente apoyado por el Ministerio de Ciencia, Innovación y Universidades de España

y por la Unión Europea a través de la subvención RTI2018-098085-BC41 (MCUI / AEI / FEDER) y GVA-FSE (PROMETEO/2019/109, "COMTACTS - COMunicación y compuTACión inTeligentes y Sociales").

REFERENCIAS

[1] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper, 2019.

[2] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet", in *IEEE MultiMedia*, vol. 18, No. 4, pp. 62-67, April 2011. doi: 10.1109/MMUL.2011.71

[3] Krishnappa D, Bhat D, Zink M, "DASHing YouTube: An analysis of using DASH in YouTube video service", *38rh Annual IEEE Conference on local computer networks*, pp: 407-415, 2013.

[4] Seufert M Egger S Slanina M Zinner T Hobfeld T Tran-Gia P, "A Survey on Quality of Experience of HTTP Adaptive Streaming", *IEEE Communications Surveys & Tutorials*, vol: 17, Nol 1, pp: 469-492, 2015.

[5] Sani, Y., Mauthe, A., & Edwards, C. "Adaptive Bitrate Selection: A Survey", *IEEE Communications Surveys and Tutorials*, vol. 19, No. 4, pp: 2985–3014. 2017. https://doi.org/10.1109/COMST.2017.2725241

[6] Bentaleb, A., Taani, B., Begen, A. C., Timmerer, C., & Zimmermann, R., "A Survey on Bitrate Adaptation Schemes for Streaming Media over HTTP", *IEEE Communications Surveys and Tutorials*, vol. 21, No. 1, pp: 562-585, 2019.

[7] Juluri P Tamarapalli V Medhi D, "SARA: Segment aware rate adaptation algorithm for dynamic adaptive streaming over HTTP", *IEEE International Conference on Communication Workshop*, ICCW 2015, pp: 1765-1770, 2015.

[8] Hoßfeld T Seufert M Sieber C Zinner T Tran-Gia P, "Identifying QoE optimal adaptation of HTTP adaptive streaming based on subjective studies", *Computer Networks*, vol: 81 pp: 320-332, 2015.

[9] A. Zabrovskiy, E. Kuzmin, E. Petrov, C. Timmerer, and C. Mueller, "AdViSE: Adaptive Video Streaming Evaluation Framework for the Automated Testing of Media Players," *Proc. 8th ACM Multimed. Syst. Conf. - MMSys'17*, pp. 217–220, 2017.

[10] Puppeteer, [Online]. Available https://pptr.dev/, May 2019.

[11] *Shaka Player* Demo. [Online]. Available: https://shaka-player-demo.appspot.com.

[12] ITU-T. Recommendation P.1203, Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport, 2017.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Desarrollo de nuevas tecnologías para la clase invertida en Ingeniería Telemática

Susel Fernandez, Luis Cruz-Piris, Diego Rivera, Ivan Marsa-Maestre
Departamento de Automática,
Universidad de Alcalá
Escuela Politécnica Superior. Campus Universitario, Ctra. Madrid-Barcelona km. 33,600.
28805. Alcalá de Henares. Madrid.
{susel.fernandez, luis.cruz, diego.rivera, ivan.marsa}@uah.es

***Resumen*—Durante los últimos años se está produciendo un cambio de tendencia desde las metodologías de enseñanza que se basan en la clase magistral hacia otro tipo de métodos donde el rol protagonista lo toman los alumnos. La clase invertida o *flipped classroom* es un método que fomenta el trabajo previo de alumno, permitiendo dedicar el tiempo de la clase a realizar otro tipo de actividades como proyectos en grupo, actividades o evaluaciones formativas, que permitan afianzar los conocimientos adquiridos con el autoestudio. Existen numerosas aplicaciones o herramientas que facilitan tareas concretas en este tipo de clases. Estas soluciones presentan algunos inconvenientes como la falta de flexibilidad o el tratarse de productos de pago. En este trabajo se presenta el desarrollo propio de una herramienta software, accesible vía web y multidispositivo, que permite integrar en el desarrollo convencional de una clase, algunas de las técnicas propias de la clase invertida.**

***Palabras Clave*—Clase invertida, docencia, tecnología, aplicación web, telemática**

I. INTRODUCCIÓN

Las clases magistrales son el método de enseñanza mayoritario en los entornos universitario. En los últimos años han surgido nuevas técnicas docentes donde se produce un intercambio entre los roles clásicos, dando el protagonismo a los estudiantes durante el desarrollo de las clases. En el caso de la clase invertida o *flipped classroom* [1], [2] se pretende fomentar el trabajo previo del alumnado, pudiendo dedicar la mayor parte del tiempo de la sesión a otro tipo de actividades como pueden ser el desarrollo de proyectos colaborativos o evaluaciones formativas. Los alumnos, previamente a las clases, tienen acceso al material que se va a impartir en una sesión concreta, pudiendo utilizar el tiempo de las clases para realizar actividades que permitan practicar los conocimientos adquiridos con el autoestudio.

El uso de estas nuevas técnicas de enseñanza presenta grandes retos como puede ser el conseguir motivar a los alumnos para que estudien la materia de cada sesión de forma previa. Metodologías de aprendizaje activo e inductivo [3] suelen ser aplicadas de forma habitual en un entorno de clase invertida. Además de la motivación, también es necesario incluir técnicas que permitan verificar y evaluar el aprendizaje del alumnado. La realización de cuestionarios, discusión de preguntas y casos prácticos, junto con otro tipo de actividades de evaluación formativa, permite verificar el trabajo personal de cada alumno.

En el mercado hay cuantiosas herramientas tecnológicas [4] publicitadas para dar soporte al aprendizaje activo e inductivo. Aunque en la sección II se mostrarán una selección de las más importantes y sus características, de forma general, este tipo de herramientas presentan problemas tales la falta de flexibilidad ante la incorporación de nuevos métodos o técnicas que el docente considera apropiadas en un momento dado. Además, la mayor parte de ellas tiene un coste para su utilización, lo cual dificulta su implantación en proyectos de innovación donde *a priori* no se conoce con seguridad cual de ellas será la mejor opción.

En este trabajo se muestra el planteamiento e implementación inicial para la creación de una aplicación web que pueda ser utiliza como herramienta durante el desarrollo de una clase basada en la metodología *flipped learning* (Sección III). Con esta aplicación se pretende unificar en un único sistema funcionalidades como la presentación de información teórica a través de diapositivas y las encuestas o cuestionarios que son ampliamente utilizados en una clase invertida. Para su diseño inicial se han estudiado tanto alguna de las herramientas comerciales existentes, como las técnicas de *flipped classroom*, seleccionando en esta primera versión los desarrollos que permitan tener las características básicas necesarias para que sea una aplicación de utilidad. Además, en su diseño se han incluido otras funcionalidades que permitan recolectar información adicional, como pueden ser los tiempos de reacción de los estudiantes, con el objetivo que cuantificar su grado de atención en cada momento. Por último, en la sección IV

expondremos las conclusiones que se han obtenido hasta el momento actual de desarrollo y pruebas de esta aplicación.

II. ESTADO DEL ARTE

Tal y como se ha avanzado en la sección I, el auge de las nuevas técnicas para la enseñanza ha promovido la aparición de una serie de herramientas tecnológicas de apoyo a la docencia. La mayoría de ellas comparten características en cuanto a la posibilidad de, mediante el uso de dispositivos móviles u ordenadores, realizar actividades durante el desarrollo de una clase. Sin embargo, existen importantes diferencias en el funcionamiento de cada una de ellas, y en general requieren un uso poco flexible y poco personalizable.

Al tratarse de aplicaciones comerciales, normalmente es necesario el pago de una licencia para obtener la posibilidad de usar todas las funcionalidades que implementan.

Un ejemplo de estas aplicaciones es Socrative [5], que ha sido utilizada con éxito en diversos ámbitos educativos [6], [7]. Cuenta con una licencia gratuita con limitaciones (como, por ejemplo, el número de alumnos máximos a gestionar), y licencias profesionales para poder utilizarla sin ninguna restricción. Esta herramienta permite la realización de encuestas, cuestionarios y otras actividades similares. Sin embargo, presenta ciertas limitaciones en cuanto a las posibilidades de añadir contenido multimedia (sólo permite imágenes) u otro tipo de información no textual a las cuestiones. Tampoco gestiona otras actividades no basadas en preguntas ni permite su calificación. Este tipo de limitaciones son compartidas por otras alternativas como son Kahoot [8] o Poll Everywhere [9]. Al igual que Socrative, estas aplicaciones disponen de una versión limitada gratuita y planes de suscripción para su uso más intensivo. Una desventaja específica de Kahoot es la ausencia de cuestionarios anónimos, que sí permiten tanto Socrative como Poll Everywhere. Sin embargo, no limita el número de respuestas por cuestión, como si hacen las otras dos, dependiendo del plan de suscripción utilizado. La utilización de Poll Everywhere en el aula ha sido estudiada en [10], con resultados en general positivos en cuanto a la mejora de la implicación de los alumnos en la asignatura. Una conclusión que se deriva de este estudio es que una funcionalidad necesaria de este tipo de aplicaciones es la posibilidad de incluir preguntas con respuesta abierta, ya que los alumnos son capaces de implicarse más en ese tipo de actividad. En [11] se describe la experiencia del uso de Kahoot en varias materias, con resultados similares en cuanto a la participación y productividad de los alumnos, sin embargo, esta herramienta no permite precisamente las respuestas abiertas.

Además de las aplicaciones anteriores, se pueden encontrar aplicaciones que tienen como objetivo una integración más amplia de funcionalidades relacionadas con el entorno educativo. De entre estas herramientas se pueden destacar Top Hat [12] y Acadly [13]. La primera de ellas se ha utilizado con éxito para la mejor del aprendizaje de idiomas [14], y ambas son ejemplos de herramientas que pretenden proporcionar un sistema más amplio para su uso en general en el entorno educativo. Incluyen la posibilidad de integración con Sistemas de Gestión del Aprendizaje (Learning Management Systems, LMS), posibilidad de monitorización de la asistencia a clase y la posibilidad de visualizar resultados a lo largo del tiempo por cada participante. Aun así, estás aplicaciones también requieren de una licencia de pago para su utilización, y son difícilmente extensibles a nuevas funcionalidades por parte de los usuarios.

III. FLIPPED LEARNING PLATFORM - FLUAH

El inicio del desarrollo de esta aplicación surge tras haber utilizado otro tipo de soluciones para la clase invertida como puede ser Socrative. Esta aplicación, desde el punto del docente, requiere que durante el desarrollo de la sesión se deba combinar su uso con el de otras aplicaciones para mostrar presentaciones o explicaciones concretas. En ocasiones este hecho puede ocasionar paradas que hagan que el nivel de atención de los alumnos se reduzca. Por este motivo, uno de los primeros requisitos que se fijaron para el desarrollo de la aplicación *Flipped Learning Platform* (FLUAH) fue que se pudieran iniciar actividades, en instantes concretos de cada sesión, sin necesidad de utilizar varias aplicaciones o herramientas. Por otro lado, para facilitar el acceso por parte de los alumnos a la aplicación, se ha optado por desarrollar la herramienta como una aplicación web, donde docentes y alumnos puedan utilizarla simplemente teniendo un equipo con cuente con acceso a Internet y navegador web. Para obtener el grado de interactividad requerido, ha sido necesario crear una aplicación con los elementos que permitan un flujo de información convencional de tipo petición y respuesta (flujo síncrono), como intercambio de mensajes en cualquier momento (flujo asíncrono). Los principales elementos que componen la arquitectura de la aplicación se muestran en la Fig. 1.

Cada usuario de la aplicación accederá a la misma a través del navegador web de su dispositivo. Tras insertar sus credenciales, la aplicación mostrará la página principal que corresponde con su perfil de acceso (profesor o estudiante). Este tipo de peticiones siguen el clásico esquema de peticiones y respuestas síncronas utilizadas en el protocolo HTTP. Otro tipo de acciones, como el inicio de nuevo cuestionario donde el profesor da paso a este y todos sus alumnos deben recibirlo en sus dispositivos, requieren de un esquema peticiones y respuestas asíncronas. El uso de WebSockets [15] permite este flujo asíncrono, y ha sido nuestra elección para implementar estas funcionalidades.

A. *Diseño y funcionamiento básico de la aplicación web*

Esta aplicación se ha pensado no sólo para tener las funcionalidades inicialmente planteadas, sino para poder ser ampliada según surjan nuevas necesidades. Por este motivo se ha optado por basar su desarrollo en un lenguaje de programación como Python y el uso de un entorno de desarrollo web como Django.

El despliegue de la aplicación es algo más complejo que una aplicación web convencional debido a que es necesario gestionar diferentes tipos de peticiones, y que

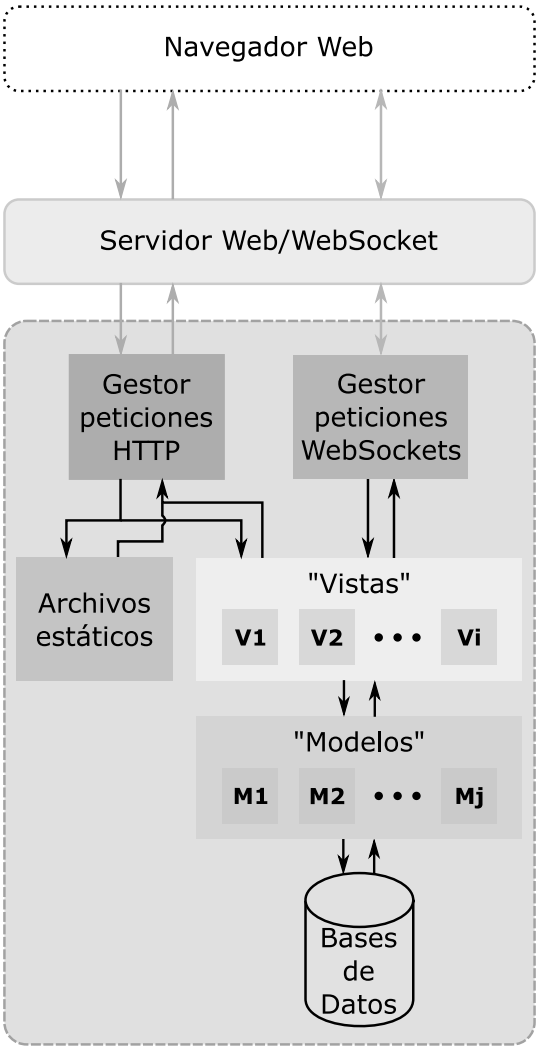


Fig. 1. Diagrama general de la arquitectura de la aplicación *Flipped Learning Platform* (FLUAH).

estas peticiones sean contestadas de forma diferente según su tipo. En la Fig. 1 podemos ver un punto de acceso común al servidor web para todas las peticiones (color verde). Estas peticiones deben clasificarse y redireccionarse al servidor web que gestiona las peticiones síncronas (Apache) o asíncronas (Daphne). En la fase de pruebas, es posible hacer esta clasificación con redirecciones entre los propios servidores web, pero en la fase de despliegue en necesario contar con un proxy como Nginx para hacer esta tarea. Una vez llega cada petición a su punto de entrada correspondiente, pueden darse tres casos:

- El cliente solicita un recurso estático (imagen, video, etc.). La respuesta con este recurso es gestionada de forma directa por el servidor web síncrono.
- El cliente solicita de forma síncrona una página dinámica, la cual varía su contenido dependiendo de la propia petición y el instante de la misma. En este caso, el servidor web envia de forma interna la petición al gestor de peticiones de Django, para que la vista correspondiente realice la acciones que tenga programadas (uso de los modelos para realizar consultas a la base de datos, u otro tipo de procesamiento) y genera el HTML que le será devuelto al cliente, siguiendo una plantilla concreta.
- El cliente o el servidor realizan un intercambio de mensajes asíncronos. Cuando un cliente realiza su primer acceso a la aplicación, se establece un canal basado en WebSockets que permanece abierto hasta que finalice la sesión. Estas peticiones son gestionadas por el servidor web asíncrono, siendo procesadas por vistas de Django desarrolladas para cada fin.

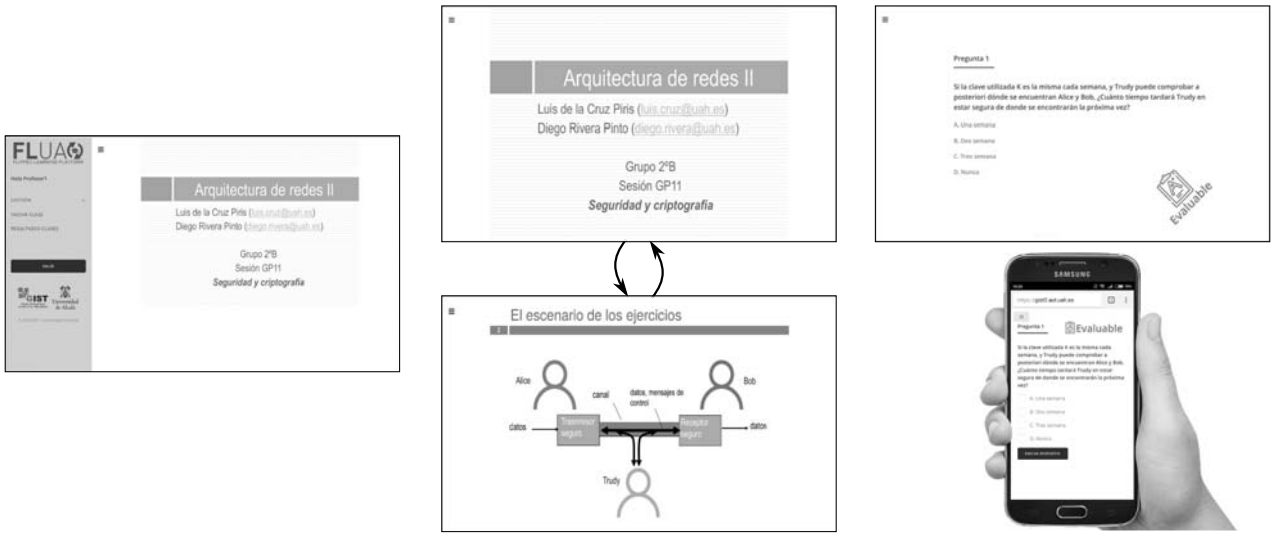
El servidor de despliegue cuenta con un certificado SSL/TLS (Transport Layer Security/Transport Layer Security) para garantizar el cifrado de la información que intercambian los clientes y el servidor.

B. Funcionamiento de la aplicación web

Tras iniciar sesión, a cada cliente se le muestra una página de inicio personalizada dependiendo de su perfil de acceso (profesor o estudiante) y de su propio usuario (estudios, asignaturas, etc.). La aplicación se ha diseñado para poder ser utilizado en inglés o español y, aunque esta selección es automática dependiendo de la configuración del navegador web del cliente, el portal da la opción de cambiar el idioma.

Para describir el funcionamiento de la aplicación, partiremos de un caso de uso concreto. En este caso se ha optado por mostrar un ejemplo desde el punto de vista de un docente en el desarrollo de una clase. En la Fig. 2.a se muestra la página de inicio que visualiza un profesor una vez ha escogido comenzar una sesión concreta (clasificada por estudios, asignatura y grupo). En ella podemos ver como se muestra en el área de trabajo una primera diapositiva con la información de la sesión actual. Para que el profesor pueda utilizar esta aplicación durante todo el desarrollo de la clase, esa área de trabajo se puede poner a pantalla completa, facilitando la proyección directa de la vista del profesor a toda la clase.

Durante el desarrollo de la clase, el profesor puede ir pasando de una diapositiva a otra de la parte teórica (Fig. 2.b). Esta información puede mostrarse a pantalla completa para ser proyectada, pudiendo además ser mostrada o no en los dispositivos de los alumnos. Cuando el profesor programó el contenido de la sesión, fijó unos puntos concretos de ésta donde realizar unos cuestionarios evaluativos. Sin tener que cambiar de pantalla, cuando la sesión llega a ese punto, aparece en pantalla la primera pregunta de la actividad planificada. En ese mismo instante, el servidor inicia el envío de mensaje a todos a los dispositivos de los alumnos de esa asignatura/grupo/sesión que están activos en el sistema. Ese mensaje asíncrono indica al lado cliente que ha comenzado una actividad evaluativa para que este inicie el proceso de peticiones y respuestas que permita mostrar en sus dispositivos los formularios a rellenar (Fig. 2.c). Esta información se presenta de forma adaptada según el dispositivo que utiliza cada cliente. Tras finalizar la actividad, la aplicación retoma la parte teórica de la clase de forma automática.



a) b) c)

Fig. 2. Caso de uso: Sesión con contenidos teóricos y custionarios de evaluación.

IV. CONCLUSIONES

En este trabajo se presenta el desarrollo inicial realizado para crear una aplicación web que facilite el desarrollo de las clases basadas en *flipped classroom*. El desarrollo de esta primera versión ha finalizado al permitir realizar tareas básicas como integrar actividades (evaluativas o no) durante el desarrollo de una sesión, sin necesidad de que el docente utilice varias aplicaciones. El desarrollo de una aplicación propia es un reto desde el punto de vista técnico, pero puede proporcionar grandes ventajas como no depender del pago de las cuotas de suscripción que exigen algunas aplicaciones, o poder personalizar por completo las opciones de esta según las necesidades específicas de cada asignatura. Además, abre una línea para la recopilación de información donde es posible cuantificar de forma activa aspectos como los tiempos de reacción, o la interacción que tienen los alumnos con la aplicación, para poder así medir su grado de atención en cada momento.

Como futuras líneas de trabajo están continuar con el desarrollo de la aplicación y la corrección de fallos para poder contar con una versión estable. Por otro lado, una vez completado este primer ciclo de desarrollo, se han planificado nuevas funcionalidades y el uso de la misma en otras asignaturas del área de Ingeniería Telemática.

AGRADECIMIENTOS

Este trabajo se ha desarrollado dentro del marco de los proyectos de innovación docente UAH/EV953 y UAH/EV1017 de la Universidad de Alcalá.

REFERENCIAS

[1] M. J. Lage, G. J. Platt, and M. Treglia, "Inverting the classroom: A gateway to creating an inclusive learning environment," *The Journal of Economic Education*, vol. 31, no. 1, pp. 30–43, 2000.

[2] I. Marsa-Maestre, J. R. Velasco, E. de la Hoz, and J. M. Gimenez-Guzman, "Una experiencia de"flipped classroom.[en] ingeniería telemática," in *XII Jornadas de Ingeniería Telemática (JITEL 2017)*. Palma de Mallorca (Spain): Universitat De Les Illes Balears, 2015.

[3] A. Prieto, D. Díaz, and R. Santiago, *Metodologías Inductivas: El desafío de enseñar mediante el cuestionamiento y los retos*. Editorial Oceano, 2014.

[4] C. Pimmer, M. Mateescu, and U. Gröhbiel, "Mobile and ubiquitous learning in higher education settings. a systematic review of empirical studies," *Computers in Human Behavior*, vol. 63, pp. 490–501, 2016.

[5] A. B. Pintado and J. M. D. de Cerio, "Socrative: A tool to dinamyze the classroom," *WPOM-Working Papers on Operations Management*, vol. 8, pp. 72–75, 2017.

[6] M. V. Frías, C. Arce, and P. Flores-Morales, "Uso de la plataforma socrative. com para alumnos de química general," *Educación química*, vol. 27, no. 1, pp. 59–66, 2016.

[7] J. Benítez-Porres, "Socrative como herramienta para la integración de contenidos en la asignatura "didáctica de los deportes"," *Universidad Europea de Madrid*, 2015.

[8] L. Rodriguez-Fernandez, "Smartphones and learning: use of kahoot in the university classroom," *Revista Mediterranea Comunicacion-Journal of Communication*, vol. 8, no. 1, pp. 181–189, 2017.

[9] P. Warnich and C. Gordon, "The integration of cell phone technology and poll everywhere as teaching and learning tools into the school history classroom," *Yesterday and Today*, no. 13, pp. 40–66, 2015.

[10] W. M. Kappers and S. L. Cutler, "Polleverywhere! even in the classroom: An investigation into the impact of using polleverywhere in a large-lecture classroom," *The ASEE Computers in Education (CoED) Journal*, vol. 6, no. 2, p. 21, 2015.

[11] M. Fuentes, M. del Mar, M. d. M. Carrasco Andrino, A. J. Pascual, A. R. Martín, C. S. García, and T. Vaello, "El aprendizaje basado en juegos: experiencias docentes en la aplicación de la plataforma virtual"kahoot"," in *XIV Jornadas de redes de investigación en docencia universitaria*. Universidad de Alicante. Instituto de Ciencias de la Educación, 2016.

[12] T. Lucke, U. Keyssner, and P. Dunn, "The use of a classroom response system to more effectively flip the classroom," in *2013 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2013, pp. 491–495.

[13] V. Vashisht, "Acadly: From curiosity to engagement to retention with one ubiquitoues tech tool," in *Reimagine Education Conference 2017*. Philladelphia (USA): QS, 2017.

[14] A. Balula, F. Marques, and C. Martins, "Bet on top hat–challenges to improve language proficiency," in *Proceedings of EDULEARN15 conference*, 2015, pp. 6–8.

[15] I. Fette and A. Melnikov, "The WebSocket Protocol," Internet Engineering Task Force, RFC Editor, RFC 6455, December 2011. [Online]. Available: https://www.rfc-editor.org/info/rfc6455


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Marco de Trabajo para la Coordinación de Testbeds 5G: URLLC como Caso de Uso

Almudena Díaz Zayas, Delia Rico, Bruno García, Pedro Merino
Departamento Lenguajes y Ciencias de la Computación, Universidad de Málaga
Andalucía Tech, Málaga, España, 29071.
adz@uma.es, delia@lcc.uma.es, bgarcía@lcc.uma.es, pedro@lcc.uma.es

***Resumen*—En este artículo se introduce un marco de trabajo para la coordinación de testbeds 5G. El marco permite ofrecer una interfaz de experimentación sobre la arquitectura de red 5G especificada por el 3GPP, la cual se basa en el uso de tecnologías de virtualización de funciones de red y en la separación de los planos de datos y de control. Para demostrar las funcionalidades que ofrece el marco de trabajo, se realiza un análisis previo de su uso en la ejecución de un experimento V2X.**

***Palabras Clave*—Coordinación, testbeds, 5G, experimentación, URLLC.**

I. INTRODUCCIÓN

Las tecnologías NFV (Network Functions Virtualizations) y SDN (Software-Defined Networking) han influido de forma irreversible en la arquitectura de las redes 5G [1] [2]. Ambas proporcionan una mayor flexibilidad: mientras que NFV permite la transferencia de funciones de red tradicionalmente propietarias y de hardware específico a aplicaciones software que se ejecutan en plataformas comerciales [3], SDN separa el plano de control del plano de datos [4]. La arquitectura de los sistemas 5G estandarizada por el 3GPP (3rd Generation Partnership) [5] se basa en la aplicación de estas dos capacidades.

En este contexto, los testbeds 5G [6] tienen que reproducir esta arquitectura y gestionar la infraestructura física y la virtualizada, desacoplando el plano de control y el plano de usuario. Al mismo tiempo, los testbeds tienen que ofrecer funcionalidades de experimentación a los verticales que quieran probar su aplicaciones y servicios sobre una red 5G. El hecho ofrecer una interfaz de experimentación para verticales que no tienen conocimientos sobre redes implica proporcionar una interfaz usable que permita abstraer la complejidad de la configuración de los componentes que integran el testbed.

La interfaz de experimentación también tiene que administrar una gran cantidad de funciones con diferentes objetivos: monitorizar los recursos (para poder recopilar medidas durante la ejecución de los experimentos), evaluar si hay suficiente capacidad para ejecutar los experimentos (con el fin de soportar la ejecución concurrente de experimentos), verificar el estado de los componentes y detectar errores durante la ejecución de los experimentos (y evitar, así, tiempos de espera innecesarios asociados a experimentos fallidos), automatizar la configuración y el control de cada uno de los componentes de la arquitectura (para mejorar la efectividad, confiabilidad y disponibilidad del banco de pruebas), programar los experimentos (y maximizar el uso del testbed), garantizar la seguridad (evitando el acceso no autorizado), etc.

Este artículo introduce un marco de trabajo para la coordinación de testbeds de experimentación que unifica el control y la gestión de los componentes físicos y virtuales de un testbed, a la vez que ofrece una interfaz de experimentación para verticales. Posteriormente, se presenta un análisis previo de un caso de uso URLLC (Ultra Realiable Low Latency Communications) basado en un servicio de control remoto de vehículos autónomos en un entorno V2X (Vehicle-to-Everything). Este marco de trabajo ha sido diseñado en el contexto del proyecto europeo 5GENESIS, cuyo objetivo es la validación de KPIs (Key Performance Indicators) 5G.

El artículo está organizado de la siguiente manera: La Sección 2 introduce el marco de trabajo para la coordinación de testbeds 5G. La Sección 3 presenta el caso de uso de URLLC como demostración. Finalmente, la Sección 4 resume las conclusiones.

II. UN MARCO DE TRABAJO PARA LA COORDINACIÓN DE TESTBEDS 5G

Para describir las funcionalidades ofrecidas por el marco de trabajo de coordinación y su flujo de trabajo, abstraeremos la arquitectura de un testbed 5G en tres capas:

- **Capa de infraestructura física:** Esta capa comprende los componentes que manejan el tráfico de usuario, incluyendo el núcleo de red 5G, la NFVI (Network Functions Virtualization Infrastructure), la red de transporte, la plataforma de edge computing, los elementos de acceso radio y los terminales de

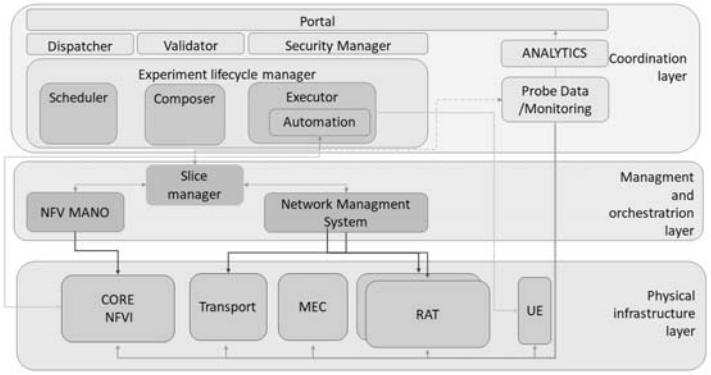


Fig. 1. Arquitectura de la marco de trabajo de coordinación

usuario. Todos los elementos de esta capa son específicos de cada plataforma.

- **Capa de gestión y orquestación:** Esta capa incluye la funcionalidad relacionada con la virtualización (gestión de los slices de red y de los recursos virtualizados), así como funcionalidades tradicionales de gestión de red para controlar las PNFs (Physical Network Functions) y otros elementos de red. Esta capa interactúa tanto con la capa de coordinación como con la capa de infraestructura, permitiendo el mapeo de los experimentos en la infraestructura física y asegurando la correcta gestión de los recursos implicados en este mapeo.
- **Capa de coordinación:** Esta capa es responsable de la coordinación general de la plataforma, proporcionando una supervisión global de la misma y su configuración extremo a extremo para el despliegue del servicio bajo prueba (con su posterior gestión y monitorización).

La Figura 1 muestra la arquitectura del marco de trabajo de coordinación y cómo éste interactúa con el resto de componentes y capas del testbed. El Portal ofrece una serie de menús que guían al experimentador en la definición, configuración y ejecución del experimento. Una vez que el experimento ha sido definido, su ciclo de vida se compone de 5 etapas: comprobación de los recursos disponibles, configuración de los recursos, ejecución y monitorización del experimento, computación de los KPIs y finalización.

El componente "Experiment Lifecycle Manager" (ELCM) supervisará el experimento a lo largo de estas etapas. Con la información proporcionada por el experimentador en el Portal, se compone un descriptor de experimento que contiene la información sobre el servicio/aplicación bajo prueba, las condiciones de red, las configuraciones de los slices, las VNFs que componen el servicio, los KPIs que se quieren validar y los dispositivos de usuario que se van a utilizar en el experimento. El descriptor de experimento contiene toda la información necesaria para ejecutar un experimento y tiene un formato común a todas las plataformas que integran el marco de trabajo de coordinación propuesto en este artículo. La versión inicial del descriptor puede ser consultada en [7].

El descriptor de experimento es traducido por el ELCM a un plan de experimentación específico para cada plataforma. Dicho plan de experimentación está basado en una serie de plantillas que contienen los parámetros de configuración y los comandos de control de los componentes del testbed. Con la información del descriptor se completan las plantillas. El plan de experimentación es un script ejecutable editado en TAP [8]. Por tanto, una vez que el plan de experimentación está disponible, se puede iniciar la ejecución del experimento.

El Slice Manager recibe la información para la configuración del slice y se comunica con el NFV MANO (Network Functions Virtualization Management and Orchestration) para desplegar el slice en la infraestructura virtualizada. Por otra lado, el Slice Manager se comunica con el Network Management System (NMS) para configurar los componentes físicos de la red de transporte, como los routers SDN. Una vez que el slice ha sido desplegado y la infraestructura física está configurada con los parámetros definidos en el descriptor de experimento, el ELCM ejecutará el cuerpo del test, actuando directamente sobre los servicios/aplicaciones y los terminales de usuario (UEs). Simultáneamente, recopilará los resultados y trazas proporcionados por las herramientas de medidas y los componentes del testbed. Los resultados obtenidos serán post-procesados en el módulo Analytics para el cálculo de los KPIs. Los primeros resultados obtenidos se pueden visualizar en [9].

III. CASO DE USO URLLC

A. Introducción a URLLC

Las comunicaciones ultra fiables de baja latencia (URLLC) [10] representan uno de los principales escenarios del 5G, junto con enhanced Mobile Broadband y massive Machine Type Communications. En muchos casos, los escenarios URLLC requieren de una fiabilidad del 99,999% libre de errores y de una latencia extremo-a-extremo del orden de unos pocos milisegundos ($\sim$1ms) [11]. De hecho, sus aplicaciones normalmente son tan críticas para la vida de las personas como los propios requisitos (p.e. Internet Táctil, cirugía remota, control remoto de vehículos autónomos, etc.) y se esperan, de

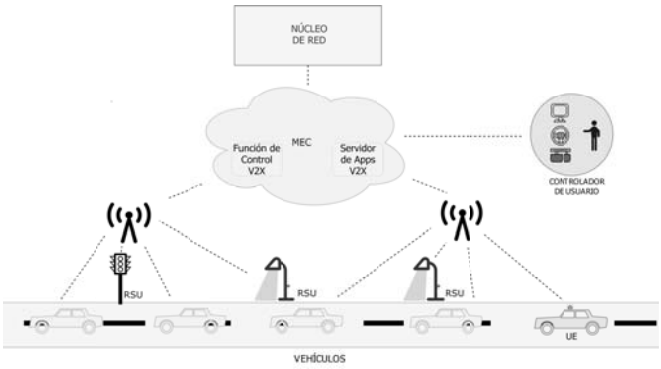


Fig. 2. Escenario del Caso de Uso eV2X.

manera adicional, altas tasas de datos. Estas necesidades exigentes requieren de un alto nivel de testing, situando a URLLC como un caso de uso apropiado para mostrar la coordinación completa del ciclo de vida de un sistema bajo prueba en el marco de trabajo.

El caso de uso seleccionado es Intelligent Transport Systems [12]. Como el estándar de NR-V2X 3GPP Release 16 está todavía en progreso, el experimento está basado en el último escenario vehicular 3GPP eV2X (enhanced vehicle-to-everything) [13] y en definiciones previas de C-V2X [14] para ser consistentes con la nomenclatura de las tecnologías existentes. Sin embargo, usando descriptores adecuados, el marco de coordinación podría soportar también la evaluación de los próximos estándares.

B. Comunicaciones V2X bajo pruebas

El sistema bajo prueba consiste en un entorno eV2X constituido por vehículos autónomos y vehículos controlados en remoto, así como diferentes fuentes de información RoadSide Unit (RSU), conectados bajo la red cellular para conformar un entorno vehicular [14] (ver Figura 2). Siguiendo el estándar 3GPP Cellular V2X, los componentes claves de un sistema V2X son la Función de Control V2X (V2X-CF) y el Servidor de Aplicaciones V2X (V2X-AS) [15]. El primero es la función lógica que se encarga de realizar acciones de control en la red, como proveer al UE con los parámetros necesarios para usar comunicaciones V2X [15]. El segundo es la entidad que se comunica con las aplicaciones V2X de los UE y los RSU y se encarga de la transmisión de datos. V2X-AS provee la infraestructura necesaria para soportar aplicaciones de verticales, implementar servicios V2X y manejar el intercambio de esos mensajes V2X.

Los vehículos autónomos y los RSU implementarían una variedad de aplicaciones, tanto aplicaciones de seguridad [16] para compartir información continua de la carretera (Forward Collision Warning, Control Loss Warning, Wrong Way Driving Warning, etc.) como aplicaciones de control remoto en vehículos y controladores de usuario. La definición de las capas altas y los mensajes específicos V2X está fuera del alcance de 3GPP y debería ser adaptada y reusada de otros estándares (SAE J3161 [17], SAE J2735 [18], IEEE 1609.2 [19], etc.). Tanto las advertencias como la información de control remoto son datos críticos y la red debería priorizarlos.

1) Definición del experimento: Para reducir la latencia al mínimo, V2X-CF y V2X-AS se desplegarán en VNFs tan cerca de los UE como sea posible. En concreto, el lugar óptimo para conseguir los requisitos de computación cerca de los dispositivos en una red 5G sería en el Edge (MEC). Las imágenes VNF y los descriptores VNF (VNFD) de las entidades se subirán a través del Portal, el cual ofrece la posibilidad de indicar el punto de despligue de los VNFs (en este ejemplo, Edge o Data Network). El Portal suministrará una plantilla de NSD (Network Service Description) con la topología de la infraestructura que el vertical editaría para interconectar los VNFs a lo interfaces correspondientes.

El siguiente paso es la selección del slice, en este caso el Portal proporciona una lista predefinida de slices: eMBB, URLLC y mMTC. El slice de URLLC dispondría de grant-free instant uplink, short transmission time interval, fast processing, etc. [20] además de políticas para asignar mayores cantidades de recursos cuando se necesite alta fiabilidad y throughput. El Network Slice Template (NST) de URLLC se actualizará con la información de los componentes NFV que necesiten ser instanciados para configurar el Network Slice Instance (NSI).

El Portal también ofrece una lista de escenarios predefinidos, permitiendo la reproducción de las condiciones de red bajo las cuales se ejecuta el test. Los escenarios ofrecidos inicialmente están basados en los escenarios definidos en los ensayos de redes pre-comerciales 5G de NGMN [21]. Para este experimento se selecciona el escenario Urban Dense Vehicular, caso no ideal de baja velocidad al aire libre con altas cargas de tráfico, densidad de usuarios y concentración de edificios. A la hora de elegir escenario, el número de UEs disponibles se muestra con una breve descripción para que sean seleccionados por el vertical. En este caso, el UE está a bordo de un vehículo conectado ubicado en una pista de prueba.

Finalmente, el vertical elegiría de una lista el caso de prueba a ejecutar. En el contexto del marco de coordinación introducido en este artículo, el caso de prueba define el KPI objetivo, el procedimiento y las medidas que deben ser recogidas a fin de validar el KPI. En principio, la especificación de los casos de prueba recae en 5Genesis,

pero el marco de coordinación puede ser actualizado con nuevos casos de pruebas pertenecientes a otros cuerpos de estandarización. En el caso de las comunicaciones V2X, los casos de pruebas están basados en la arquitectura V2X especificada en [15]. Para mayor claridad, el caso de prueba seleccionado en este ejemplo mide la duración desde la transmisión de paquetes en el UE y el controlador de usuario, hasta la recepción exitosa en la función de control V2X-CF conectada por la interfaz V3 y el servidor de aplicaciones V2X-AS conectado por la interfaz V1, y viceversa. Se pueden ejecutar de igual manera casos de prueba adicionales para medir, por ejemplo, el tiempo de recepción de mensajes de aplicación o la pérdida de paquetes, y calcular otros KPI como la fiabilidad.

Toda esta información será utilizada para componer el descriptor del experimento que será enviado al marco de coordinación. El marco de coordinación usará la información contenida en este descriptor para configurar la infraestructura, reproducir el escenario y configurar las sondas. Posteriormente, el NST se enviará al Slice Manager y una vez que se despligue el slice, la capa de coordinación interactuará con el UE y el controlador de usuario para iniciar la comunicación con las entidades V2X-CF y V2C-AS.

La duración y número de repeticiones de los tests están especificados en los casos de pruebas. Después de cada repetición, la capa de coordinación recogerá las medidas para calcular el parámetro especificado, RTT medio en nuestro caso seleccionado. Las medidas se almacenan en la base de datos común disponible en la capa de coordinación y, tras ejecutar todas las iteraciones, el módulo de Análisis calculará los valores medios y proporcionará el valor final del KPI (latencia extremo-a-extremo).

IV. CONCLUSIONES

En este artículo, se propone un marco de trabajo para la coordinación de testbeds 5G para validar tecnologías 5G y KPIs. El marco expone las características de experimentación a los verticales y automatiza la ejecución de experimentos en una red 5G de extremo a extremo que incluye tecnologías como NFV y SDN y soporta slicing de red. Se presenta en detalle la arquitectura del marco, incluida la descripción de sus componentes, la interconexión con el resto de las capas del testbed y el flujo de trabajo del ciclo de vida de la experimentación cuando se ejecuta un experimento o varios experimentos al mismo tiempo. Finalmente, presentamos un caso de uso en el que el marco se utilizaría para ejecutar un experimento de URLLC y validar los KPI de URLLC. El ejemplo demuestra cómo el marco permite automatizar el servicio de red, ejecutar las aplicaciones, configurar la red, recopilar las mediciones y calcular los KPI.

AGRADECIMIENTOS

Este trabajo ha sido realizado bajo los auspicios del Proyecto 5GENESIS. Este proyecto ha recibido financiación del programa European Union's Horizon 2020 research and innovation (grant agreement no. 815178). Este trabajo también está parcialmente financiado por el Ministerio de Ciencia, Innovación y Universidades de España (FPU grant AP2017-72875) y el proyecto EuWireless (H2020 grant agreement no. 777517).

REFERENCIAS

[1] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "Nfv and sdn key technology enablers for 5g networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2468–2478, Nov 2017.

[2] M. S. Bonfim, K. L. Dias, and S. F. L. Fernandes, "Integrated nfv/sdn architectures: A systematic literature review," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 114:1–114:39, Feb. 2019. [Online]. Available: http://doi.acm.org/10.1145/3172866

[3] ETSI, *Network Operator Perspectives on NFV priorities for 5G*, February 2017, http://portal.etsi.org/NFV/NFV_White_Paper_5G.pdf.

[4] M. Boucadair and C. Jacquenet, "Software-defined networking: A perspective from within a service provider environment." *RFC*, vol. 7149, pp. 1–20, March 2014. [Online]. Available: http://dblp.uni-trier.de/db/journals/rfc/rfc7100-7199.html

[5] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; system architecture for the 5g system; stage 2 (release 16)" 3GPP TS 23.501 v16.0.2, April 2019.

[6] R. Verdone and A. Manzalini, *5G Experimental Facilities in Europe, NetWorld 2020 European Technology Platform, White Paper, Version 11.0*, 2016, https://www.networld2020.eu/wp-content/uploads/2016/03/5G-experimentation-Whitepaper-v11.pdf.

[7] A. Díaz-Zayas et al., "Deliverable d2.3 initial planning of tests and experimentation," February 2019, https://5genesis.eu/deliverables/.

[8] *OpenTAP: An Open Source Test Automation Project*, 2019, http://opentap.io/.

[9] A. Díaz-Zayas et al., "Deliverable d6.1 trials and experimentation (cycle 1)," July 2019, https://5genesis.eu/deliverables/.

[10] ITU-R, "Imt vision - "framework and overall objectives of the future development of imt for 2020 and beyond"," Recommendation ITU-R M.2083, September 2015, https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf.

[11] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; study on scenarios and requirements for next generation access technologies; (release 14)," 3GPP TS 38.913 V14.3.0, June 2017.

[12] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; service requirements for the 5g system; stage 1 (release 16)," 3GPP TS 22.261. V16.0.0, June 2017.

[13] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; enhancement of 3gpp support for v2x scenarios; stage 1 (release 16)," 3GPP TS 22.186 V16.1.0, December 2018.

[14] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; service requirements for v2x services; stage 1 (release 15)," 3GPP TS 22.185 V15.0.0, June 2018.

[15] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; architecture enhancements for v2x services (release 16)," 3GPP TS 23.285 V16.0.0, March 2019.

[16] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; study on lte support for vehicle to everything (v2x) services (release 14)," 3GPP TS 22.885. V14.0.0, December 2015.

[17] S. International, "On-board system requirements for lte v2x v2v safety communications," SAE Standard J3161, January 2012, https://www.sae.org/standards/content/j3161.

[18] S. International, "Dedicated short range communications (dsrc) message set dictionary," SAE Standard J2735, March 2016, https://www.sae.org/standards/content/j2735_201603/.

[19] I. S. Association, "Ieee standard for wireless access in vehicular environments–security services for applications and management messages," IEEE 1609.2, January 2016, https://standards.ieee.org/standard/1609_2-2016.html.

[20] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.

[21] N. Alliance, "Definition of the testing framework for the ngmn 5g pre-commecial networks trails v1," 0. Technical report, Tech. Rep., 2018.


XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Minimización del consumo de energía en redes SDN bajo restricciones TCAM

Jaime Galán-Jiménez, Javier Berrocal, Marino Linaje, Cristóbal Gómez
Escuela Politécnica de Cáceres
Universidad de Extremadura
Avda. de la Universidad, S/N, Cáceres, España
{jaime, jberolm, mlinaje, cgomezx}@unex.es

***Resumen*—Tanto la comunidad investigadora como la industria han realizado grandes esfuerzos para proponer soluciones al problema de consumo de energía en las redes de comunicaciones. La irrupción del nuevo paradigma de interconexión de redes, denominado SDN (*Software-Defined Networking*) abre las puertas para proponer nuevas técnicas que aprovechen el conocimiento global que el controlador SDN posee acerca de la red que gestiona. Sin embargo, los trabajos existentes en este ámbito no tienen en cuenta la restricción del tamaño limitado que tienen las tablas de flujos de los switches SDN. En este trabajo se propone el algoritmo ETAR (*Energy and TCAM-Aware Routing*), que minimiza el consumo de energía de una red SDN y a su vez aplica técnicas de compresión para reducir el número de reglas instaladas en las tablas de flujos. Los resultados obtenidos sobre topologías de red reales indican que es posible conseguir un ahorro de energía significativo, respetando a su vez el límite establecido para el número máximo de reglas que se pueden instalar.**

***Palabras Clave*—Eficiencia energética, SDN, TCAM, Energy-Aware Routing.**

I. INTRODUCCIÓN

Estudios recientes han demostrado que las TIC son responsables del 2% al 10% del consumo total de energía eléctrica en todo el mundo [1], [2]. De hecho, se estima que el sistema de telecomunicaciones europeo puede llegar a requerir 35,8 TWh en 2020 [3]. Por ello, la atención sobre el *green networking* ha ido creciendo en los últimos años [4], [5]. El consumo de energía se debe principalmente a elementos activos de los routers IP como el chasis o los puertos, mientras que la carga de tráfico tiene una influencia mínima sobre éste [6]. Basándose en esta observación, el enfoque denominado *energy-aware routing* (EAR) tiene como objetivo reducir al mínimo el número de elementos de red utilizados, mientras que todas las demandas de tráfico se encaminan procurando que ningún enlace de la red se sobrecargue de tráfico [1], [7]. Precisamente, el apagado o puesta en *standby* de routers o tarjetas de línea puede traducirse en un importante ahorro energético. Sin embargo, no es fácil hacerlo desde un punto de vista práctico, ya que, en primer lugar, la acción de encender o apagar routers y tarjetas de línea necesita de un determinado tiempo de acción que, a la larga, puede traducirse en una reducción en el ciclo de vida de los dispositivos.

Por otro lado, la irrupción de las redes SDN (*Software-Defined Networking*) y la utilización del protocolo OpenFlow [8] ha permitido a los investigadores proponer nuevas soluciones que permitan reducir el consumo de energía de la red aprovechando la filosofía centralizada de esta aproximación. En las redes tradicionales, los dispositivos de red tales como switches y routers actúan como sistemas cerrados. Los usuarios sólo pueden controlarlos a través de interfaces limitadas y específicas del proveedor en cuestión. Por otra parte, la capa de datos y la capa de control están integrados en cada dispositivo, haciendo bastante difícil la tarea de implementar nuevos protocolos de red. Por su parte, SDN es un nuevo paradigma de redes que separa la capa de control de la capa de datos. Proporciona flexibilidad para desarrollar y probar nuevos protocolos y políticas de red en redes reales. De hecho, muchas aplicaciones se han construido utilizando la API de OpenFlow [8] en los últimos años. Por ejemplo, B4 es uno de los primeros trabajos realizados en este ámbito, donde se aplica SDN en la red del *data center* de Google [9]. B4 ha estado en desarrollo durante tres años y durante ese tiempo ha demostrado que puede satisfacer de manera eficiente las demandas tráfico, que es compatible con el rápido despliegue de nuevos servicios de control de la red y que es robusto frente a posibles fallos.

Existen varios trabajos en los que han utilizado OpenFlow para desplegar EAR en una red SDN, pero en la mayoría de ellos se da por hecho que las tablas de flujos de cada switch tienen un número infinito de reglas. Sin embargo, en la práctica esta hipótesis no es cierta, y el espacio de las tablas para reglas se convierte en un cuello de botella significativo para escalar las redes SDN. Además, estas tablas de flujos se implementan utilizando memorias TCAM, que son relativamente caras y consumen

Dir. IP Origen	Dir. IP Destino	Puerto de Salida
10.0.0.1	10.0.0.2	Puerto 2
10.0.0.2	**10.0.0.1**	**Puerto 1**
10.0.0.3	**10.0.0.1**	**Puerto 1**
10.0.0.1	10.0.0.3	Puerto 3
10.0.0.4	**10.0.0.1**	**Puerto 1**

Fig. 1. Tabla de flujos llena antes de la compresión.

	Dir. IP Origen	Dir. IP Destino	Puerto de Salida
	10.0.0.1	10.0.0.2	Puerto 2
	10.0.0.1	10.0.0.3	Puerto 3
Default	*	*	**Puerto 1**

Fig. 2. Tabla de flujos después de aplicar la compresión.

gran cantidad de energía. El tamaño de estas memorias puede albergar entre cientos y miles de entradas [10]–[12], lo que supone una restricción importante a la hora de implementar soluciones EAR. Una asignación ineficiente de reglas puede llevar a una solución de enrutamiento inesperada, causando congestión de la red y afectando a la QoS (*Quality of Service*).

En este trabajo se propone una solución que no sólo tiene en cuenta el ahorro energético, sino que se considera también la limitación del tamaño de las memorias TCAM, asignando rutas de manera eficiente. Para ello, se ha implementado un algoritmo que cumple con todos estos requisitos en un controlador SDN y se han realizado distintos tipos de pruebas sobre topologías de red reales con el objetivo de extraer una serie de conclusiones.

El resto del artículo se describe como sigue. El algoritmo propuesto, denominado ETAR (*Energy and TCAM-Aware Routing*), junto con su pseudocódigo correspondiente se explica en la Sección 2. Las herramientas utilizadas y el entorno sobre el que se han desarrollado las pruebas se indican en la Sección 3. La Sección 4 describe la metodología de las pruebas realizadas, además de un análisis de los resultados obtenidos. Finalmente, la Sección 5 presenta una serie de conclusiones extraídas tras realizar el trabajo presentado.

II. ENERGY AND TCAM-AWARE ROUTING

El propósito principal de nuestro trabajo es proponer un algoritmo que, teniendo en cuenta el tamaño de las memorias TCAM de los switches SDN, sea energéticamente eficiente, es decir, minimice el consumo de energía global de la red SDN.

A. Pseudocódigo

El código del algoritmo se puede dividir en dos fases:

1) Parte 1: empezamos a calcular todas las posibles rutas de la red teniendo en cuenta las limitaciones de capacidad de los enlaces y el máximo número de reglas de las tablas de flujos de los nodos (algoritmos 1, 2 y 3). Para cada router $u \in V$, tenemos dos colecciones F_u y G_u que van a contener flujos normales y flujos por defecto, respectivamente. Cuando la tabla de rutas esté llena, trataremos de comprimir la tabla de flujos (Fig. 1 y Fig. 2). La idea de comprimir la tabla, consiste en establecer por defecto el puerto que aparezca en el mayor número de reglas. Gracias a esto, tendremos un menor número de reglas instaladas y más espacio disponible para instalar nuevas reglas.

Un ejemplo de la compresión de una tabla de flujos se puede observar en la Fig. 1 (antes de comprimir) y en la Fig. 2 (después de comprimir). En este caso, podemos observar que el tamaño de la tabla de flujos sería de 6 reglas o flujos. Una vez se haya llenado, podemos proceder a comprimirla. Para ello:

1) Buscamos el puerto de salida más usado, el puerto número 1 en este ejemplo.
2) Borramos todos los flujos que utilicen ese puerto de salida.
3) Añadimos un nuevo flujo por defecto, en el que el puerto de salida se corresponde con el más usado, sacado en el paso 1.

Así, ya tendríamos espacio para instalar nuevos flujos y, a partir de ahora, todas los paquetes que no coincidan con ninguno de las flujos de la tabla, se enviarán usando la regla por defecto.

2) Parte 2: en esta fase se intentan borrar los enlaces menos cargados (algoritmo 4). La idea de esto, es apagar los enlaces menos cargados y pasar su tráfico a otros enlaces. De esta manera se consigue ahorrar energía reduciendo el número de enlaces activos.

Antes de pasar a explicar el pseudocódigo, es conveniente explicar el significado de los símbolos utilizados:

- $G = (V, E)$: grafo no dirigido formado por V y E.
- V: conjunto de vértices (en nuestro caso switches).
- E: conjunto de enlaces.
- C_e: capacidad del enlace e.
- R_e: capacidad residual del enlace e.
- C_u: capacidad de la tabla de flujos del switch u.
- u: switch.
- D: conjunto de demandas de tráfico a enviar.
- $D^{st} \in D$: demanda de tráfico del nodo s a t.
- F_u: conjunto de flujos normales instalados en el switch u.
- G_u: conjunto de flujos por defecto instalados en el switch u.
- G': grafo dirigido.
- w_e: peso del enlace e.
- P^{st}: camino más corto desde s hasta t.

A continuación, se va a exponer el pseudocódigo dividido en cuatro partes/algoritmos con su correspondiente explicación:

1) **Algoritmo 1 (Encontrando una ruta factible)**: este algoritmo recibe cómo parámetros de entrada un grafo no dirigido compuesto por el conjunto de enlaces y nodos de la red, la capacidad de cada

Algoritmo 1 Encontrando una ruta factible

Input: un grafo no dirigido $G = (V, E)$, capacidad del enlace $C_e \forall e \in E$, tamaño de la tabla de flujos $C_u \forall u \in V$ y una matriz de demandas de tráfico D.
Output: solución de direccionamiento sobre el grafo G.

1: Capacidad Residual $R_e = C_e \forall e \in E$
2: Inicialmente, $F_u = \emptyset$ y $G_u = \emptyset \ \forall u \in V$
3: Creamos un grafo dirigido $G' = (V, E')$ partiendo de G donde $\forall (u, v) \in E$, añadimos ambas direcciones (u, v) y (v, u) a E'. El peso inicial del enlace $w_e = 1 \ \forall e \in E'$
4: **while** $D^{st} \in D$ no tiene asignada una ruta **do**
5: encontramos el camino más corto P^{st} en G' tal que $R_e \geq D^{st} \ \forall e \in P^{st}$
6: asignamos la ruta P^{st} a la demanda D^{st}
7: actualizamos $R_e := R_e - D^{st} \ \forall e \in P^{st}$
8: actualizamos el peso de los enlaces en proporción al tamaño de $|F_u|$ como en el Algoritmo 2
9: **if** $|F_u| == C_u$ **then**
10: encogemos la tabla de flujos en $u \ \forall u \in P^{st}$
11: **end if**
12: actualizamos F_u y $G_u \ \forall u \in P^{st}$ usando el Algoritmo 3
13: **end while**
14: **return** solución de direccionamiento (si existe) asignada a D

Algoritmo 2 Actualizando el peso de los enlaces

Input: un grafo no dirigido $G = (V, E)$, un conjunto de flujos normales $F_u \ \forall u \in V$ y el máximo valor de capacidad de las tablas de flujos $C_{max} = \max(C_u) \ \forall u \in V$.
Output: pesos de los enlaces de G' ajustados.

1: Creamos un grafo dirigido $G' = (V, E')$
2: **for** (u, v) in E' **do**
3: calculamos el uso del flujo $v : U_v = C_{max} \times |F_v|/C_v$
4: actualizamos $w_{uv} = \max(U_v, 1)$
5: **end for**

Algoritmo 3 Actualizando las listas de flujos normales y por defecto

Input: un grafo no dirigido $G = (V, E)$, el camino más corto P^{st} encontrado en el Algoritmo 1, el conjunto de flujos normales F_u y el conjunto de flujos por defecto $G_u \ \forall u \in P^{st}$ y el puerto por defecto de cada nodo $d(u) \ \forall u \in P^{st}$.
Output: F_u y $G_u \ \forall u \in V$ actualizados.

1: **for** $u \in P^{st}$ **do**
2: **for** $v \in G.neighbour(u)$ **do**
3: **if** $(u, v) \in P^{st} and \ v == d(u)$ **then**
4: $G_u = G_u \cup g^{st}_{uv}$
5: **else if** $(u, v) \in P^{st} and \ v \neq d(u)$ **then**
6: $F_u = F_u \cup f^{st}_{uv}$
7: **end if**
8: **end for**
9: **end for**

enlace, el espacio de las tablas de flujos de cada nodo y el conjunto de demandas de tráfico a enviar. Para empezar, inicializamos la capacidad residual de cada enlace con el valor de la capacidad que tiene cada uno de ellos, las listas de flujos normales y por defecto de cada nodo están vacías y creamos un grafo dirigido a partir del grafo no dirigido (añadiendo los ejes en ambas direcciones), asignando al peso de cada enlace el valor 1. Para cada demanda de tráfico, calculamos su ruta mediante el algoritmo del camino mínimo en el grafo dirigido, teniendo en cuenta que la demanda en cuestión no supere la capacidad residual de ninguno de los enlaces que conforman la ruta. Después, instalamos las reglas necesarias según la ruta que se acaba de calcular. A continuación, actualizamos el valor de la capacidad residual de cada uno de los enlaces de la ruta, para ello, le restamos el valor de la demanda al actual valor de la capacidad residual del enlace. También, actualizamos el peso de los enlaces conforme al tamaño actual de las tablas de flujos de los nodos como se explica en el Algoritmo 2. Para terminar, si el tamaño de la tabla de flujos de los nodos que conforman la ruta calculada es igual al número máximo de reglas que se pueden instalar en ellos, reducimos el número de reglas de la tabla del nodo que cumpla esa condición como se indica en el Algoritmo 3.

2) **Algoritmo 2 (Actualizando el peso de los enlaces)**: este algoritmo lleva a cabo el cálculo de los pesos de los enlaces. Los parámetros de entrada que recibe son un grafo no dirigido formado por el conjunto de nodos y enlaces de la red, la lista de flujos normales (F_u) instalados en los nodos y el valor del número de flujos que se pueden instalar en el nodo de la red que tenga la mayor capacidad de reglas en su tabla de flujos (C_{max}). Primero, creamos un grafo dirigido a partir del grafo no dirigido. Para cada enlace (u, v) de ese grafo dirigido, calculamos el peso utilizando la Ec. 1. Para finalizar, actualizamos el valor del peso de los enlaces con el valor máximo entre el peso que hemos calculado y 1.

$$U_v = C_{max} \times |F_v|/C_v \tag{1}$$

La idea del cálculo de pesos, sería conseguir un balanceo de reglas (compartir las reglas) entre los nodos de la red.

3) **Algoritmo 3 (Actualizando las listas de flujos normales y por defecto)**: en este algoritmo se realiza la actualización de las listas de flujos normales F_u y la lista de flujos por defecto G_u de los nodos que forman parte de la ruta calculada en el Algoritmo 1. Recibe como parámetros de entrada un grado dirigido, la ruta calculada, F_u y G_u de todos los nodos de la ruta y el puerto por defecto de cada router que forma parte de la ruta. Simplemente, añadimos los flujos por defecto instalados a las listas G_u y los flujos normales instalados a las listas F_u de los routers.
4) **Algoritmo 4 (Eliminando los enlaces menos cargados)**: en este algoritmo se borran los enlaces de la red menos utilizados tras calcular las rutas para todas las demandas de tráfico. Recibe como parámetros de entrada un grafo dirigido, además de la capacidad C_e y la capacidad residual R_e de todos los enlaces de la red. Primero calculamos todas las rutas. Después

Algoritmo 4 Eliminando los enlaces menos cargados

Input: un grafo no dirigido $G = (V, E)$, capacidad del enlace C_e y capacidad residual R_e $\forall e \in E$
Output: solución de direccionamiento teniendo en cuenta los enlaces activos.

1: **while** enlaces puedan ser borrados **do**
2: eliminamos el enlace e que no ha sido ya elegido y tiene el menos valor $r_e = C_e/R_e$
3: calculamos una ruta factible con el Algoritmo 1
4: if no existe una ruta factible, ponemos e otra vez en G
5: **end while**
6: **return** ruta factible (si existe)

borramos el enlace que tenga el menor valor dado por la Ec. 2 (el enlace que menos cargado esté de tráfico) y comprobamos que se puedan calcular todas las rutas sin ese enlace, si no se puede calcular alguna de ellas, volvemos a poner el enlace en el grafo; si sí se pueden calcular todas volvemos a borrar el siguiente enlace con el menor valor dado por la Ec. 2 y calcularemos las rutas, lo repetiremos hasta que al borrar un enlace no se pueda calcular la ruta para alguna de las demandas.

$$r_e = C_e/R_e \tag{2}$$

III. HERRAMIENTAS UTILIZADAS

En este apartado se muestran las herramientas utilizadas durante el desarrollo del presente trabajo. Todas ellas se han utilizado en un equipo con las siguientes especificaciones: CPU Intel Core i7-3632QM a 2.20 GHz, memoria RAM de 8GB y Sistema Operativo Ubuntu 16.04.

A. OpenDaylight

OpenDaylight [13] es el controlador SDN que se ha seleccionado y, por tanto, es el controlador en el que se ha desarrollado la aplicación que ejecuta el algoritmo ETAR. La versión utilizada es la Hydrogen Base Edition [14], ya que, esta versión del controlador permite instalar en el controlador nuevos paquetes OSGI. Nuestro interés radica en instalar nuevos módulos en el controlador que ejecuten el algoritmo propuesto.

B. Mininet

Mininet [15] es una herramienta que permite crear una red virtual realista (con kernel, switches y aplicaciones reales) en una sola máquina, en unos segundos y con un solo comando. Se puede interactuar fácilmente con la red, personalizarla o implementarla en hardware real, utilizando para ello la consola de comandos CLI de Mininet o la API. Con Mininet simularemos las topologías de red, llevaremos a cabo la ejecución del algoritmo propuesto y enviaremos las demandas de tráfico obteniendo estadísticas de la red gracias al uso de la herramienta Iperf.

C. Iperf

Iperf [16] es una herramienta que permite medir el máximo ancho de banda disponible en una red IP. Soporta la configuración de diversos parámetros relacionados con los tiempos, buffers y protocolos (TCP, UDP, SCTP con IPv4 y IPv6). Iperf se utiliza junto a Mininet para enviar la matriz de demandas de tráfico y obtener reportes de la red que nos permitan medir el jitter, la pérdida de paquetes o el ancho de banda.

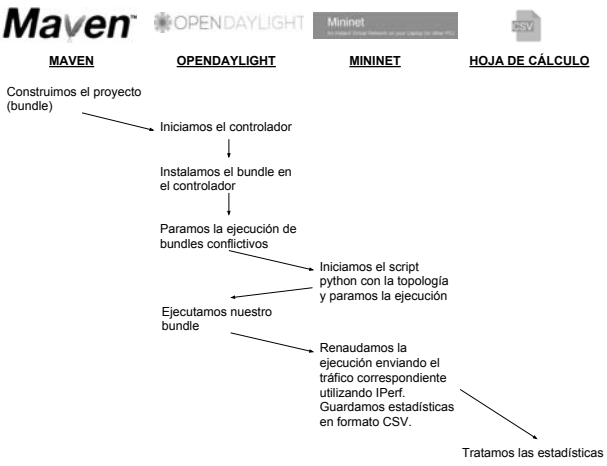


Fig. 3. Metodología de las pruebas.

D. Maven

Maven [17] es una herramienta de gestión y comprensión de proyectos de software. Basado en el concepto POM (Project Object Model, es la representación XML de un proyecto Maven), Maven puede gestionar la construcción de un proyecto, la generación de informes y la documentación de un proyecto a partir de una pieza central de la información.

IV. RESULTADOS EXPERIMENTALES

Las pruebas que se han realizado consisten en ejecutar el algoritmo de optimización del consumo energético y tamaño de las tablas de flujos en dos topologías diferentes, cada una de ellas con una matriz de demandas de tráfico y una matriz de capacidades de los enlaces de la red.

La metodología llevada a cabo para cada prueba se puede observar en la Fig. 3. En resumen, se debe construir el paquete que incluye el algoritmo, lanzar el controlador e instalar dicho paquete en él. Después de esto, se carga la topología y se inicia el envío de tráfico entre cada par origen-destino de la red (Fig. 4). Por último, se obtienen los resultados para ser analizados.

En la Fig.4 se muestra el envío de tráfico con IPerf. Para empezar, contamos con una matriz de tráfico para la topología en cuestión, de la que el controlador tiene conocimiento y que transformamos en directivas de envío de tráfico Iperf. Para terminar, ejecutamos esas directivas y se produce el envío de tráfico en la topología.

A. Topologías

Como se ha mencionado anteriormente, se han utilizado dos topologías para poder probar el algoritmo implementado: Abilene y Nobel-Germany. Ambas topologías se han extraído de la librería SNDLib [18]. En la Tabla I se muestra una comparativa entre ambas.

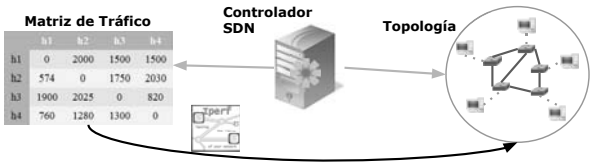


Fig. 4. Envío de tráfico usando IPerf.

Tabla I
COMPARATIVA TOPOLOGÍAS USADAS.

Topología	Nodos	Enlaces
$Abilene$	12	15
$Nobel - Germany$	17	26

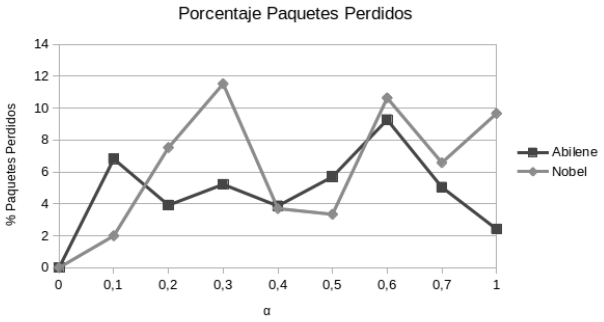


Fig. 5. Porcentaje de Paquetes Perdidos.

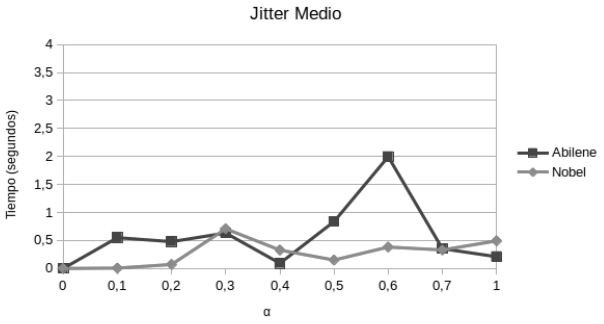


Fig. 6. Jitter Medio.

B. Análisis de Resultados

Antes de pasar a ver y analizar las gráficas de los resultados obtenidos, se define un nuevo parámetro, α, que hace referencia al número máximo de reglas que pueden instalarse en las tablas de flujos de los nodos de la red (Ec. 3), siendo d el número total de demandas de tráfico que se van a enviar en la red. Por ejemplo, si $\alpha = 0.1$, quiere decir que en dicha prueba el número máximo de reglas que se pueden instalar es igual al 10% del número total de demandas que se van a enviar en la red (132 en Abilene y 272 en Nobel-Germany).

$$C_u = (\alpha * d) \tag{3}$$

Entre los resultados obtenidos se encuentran aquéllos que proporciona iPerf directamente: *jitter*, paquetes perdidos y paquetes *out of order*. Por otro lado, se ha calculado el tiempo de cómputo requerido para ejecutar el algoritmo, así como el ahorro de energía obtenido y el número de reglas instaladas en las tablas de flujos de los switches SDN. Para cada parámetro analizado se proporciona una explicación basada en los resultados mostrados en las siguientes gráficas.

1) Porcentaje de Paquetes Perdidos: La Fig. 5 muestra el porcentaje de paquetes perdidos en función de α para las dos topologías consideradas. En ella, se puede observar que ETAR obtiene mejores resultados para Abilene que para Nobel y que, en general, se trata de un valor bajo de paquetes perdidos. De hecho, el porcentaje de paquetes perdidos en Nobel no llega a superar nunca el 12% y los picos más altos los obtiene para $\alpha = 0.3$ y $\alpha = 0.6$. Es en este valor de $\alpha = 0.6$ donde Abilene presenta su valor pico, aunque no llega a sobrepasar el 10%.

2) Jitter Medio: La Fig. 6 muestra el valor de jitter medio en función de α para las dos topologías consideradas. En general, se puede observar que se obtienen valores bajos en términos de jitter, inferiores a 0.5 segundos en la mayoría de los casos, y que son similares para ambas topologías.

3) Porcentaje de Paquetes Out of Order: La Fig. 7 muestra el porcentaje de paquetes *out of order* en función de α para las dos topologías consideradas. El valor obtenido para la topología Abilene ronda entre el 3% y el 8% de paquetes, mientras que para Nobel es algo superior, entre el 3% y el 11%. En ambos casos, los valores máximos se obtienen para $\alpha = 0.4$.

4) Número de Reglas Instaladas: La Fig. 8 muestra el porcentaje de reglas instaladas en el nodo más cargado de la red en función de α para las dos topologías consideradas. En $\alpha = 0$, el número (y por ende el porcentaje) de reglas instaladas es 0, ya que, el número máximo de reglas que se pueden instalar es $C_u = 0$. Para valores de $\alpha < 0.5$, el porcentaje de reglas instaladas es creciente. Sin embargo, para $0.5 \leq \alpha \leq 1$, el número de reglas instaladas se mantiene durante todo el intervalo. Esto es debido a que a partir de $\alpha = 0.5$ el número máximo de reglas instaladas no llega a superar el límite establecido, por tanto, aunque sigamos aumentando el límite, las reglas se van a instalar de la misma manera en todas las pruebas. Además, aunque los valores obtenidos son similares para ambas topologías, se realiza una compresión de las tablas de flujos más eficiente en Nobel que en Abilene para el rango $0.5 \leq \alpha \leq 1$.

5) Tiempo de Cómputo: En la Tabla II se puede observar que el tamaño de la red influye en gran manera en el tiempo de cómputo requerido para ejecutar ETAR. De hecho, se tarda casi el quíntuple de tiempo en ejecutar el algoritmo en la topología Nobel-Germany con respecto a la topología Abilene (5 nodos y 11 enlaces de diferencia, Tabla I).

6) Ahorro de Energía: En la Tabla III se muestra el ahorro de energía obtenido por el algoritmo propuesto para las dos topologías consideradas. Teniendo en cuenta

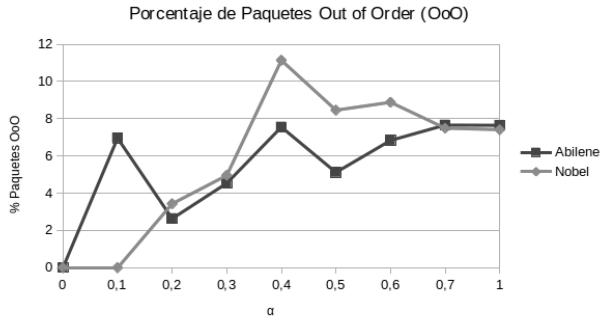


Fig. 7. Porcentaje de Paquetes *Out of Order*.

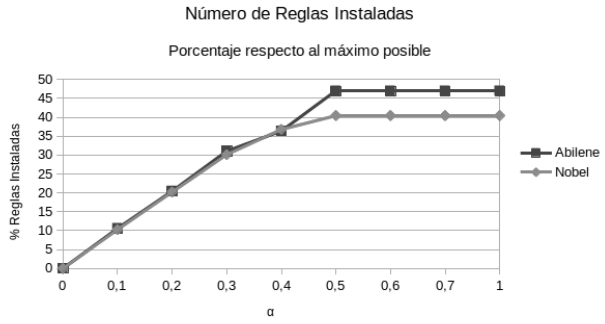


Fig. 8. Porcentaje del número de reglas instaladas.

el tráfico correspondiente a la matriz de tráfico utilizada, es posible conseguir un 11.5% de ahorro para la topología Nobel-Germany, mientras que para Abilene se consigue un mayor ahorro de energía, en concreto un 20%.

V. CONCLUSIONES

Este trabajo estudia el problema de consumo de energía en redes SDN. Aunque existen varios trabajos que tratan de dar solución aprovechando la filosofía centralizada de este nuevo paradigma, en ningún caso se tiene en cuenta la restricción del tamaño limitado de las tablas de flujo de los switches SDN (memorias TCAM). Por ello, en este trabajo se propone el algoritmo ETAR (*Energy and TCAM-Aware Routing*), que minimiza el consumo de energía de una red SDN y aplica técnicas de compresión para reducir el número de reglas instaladas en las tablas de flujos. Los resultados obtenidos sobre topologías de red reales indican que es posible conseguir un ahorro de energía significativo, respetando a su vez el límite establecido en el número máximo de reglas que se pueden instalar.

AGRADECIMIENTOS

Este trabajo ha sido financiado, en parte, por los proyectos 4IE (0045-4IE-4-P) y 4IE+ (0499_4IE_PLUS_4_E) financiados por el programa Interreg V-A España-Portugal (POCTEP) 2014-2020, por el Ministerio de Ciencia, Innovación y Universidades (RTI2018-094591-B-I00), por la Consejería de Economía e Infraestructuras de la Junta de Extremadura (IB18030, GR18112) y por el Fondo Europeo de Desarrollo Regional (FEDER).

Tabla II
TIEMPO DE CÓMPUTO

Topología	Tiempo Cómputo Medio (ms)
$Abilene$	2418
$Nobel - Germany$	10857

Tabla III
AHORRO ENERGÉTICO

Topología	Enlaces	Enlaces Eliminados	% Ahorro
$Abilene$	15	3	20%
$Nobel - Germany$	26	3	11.5%

REFERENCIAS

[1] L. Chiaraviglio, M. Mellia, F. Neri, "Minimizing ISP Network Energy Cost: Formulation and Solutions", IEEE/ACM Transaction in Networking 20 (2011) 463 - 476.
[2] Global Action Plan. http://globalactionplan.org.uk. Último acceso: 15/09/2019.
[3] R. Bolla, F. Davoli, R. Bruschi, K. Christensen, F. Cucchietti, S. Singh, "The Potential Impact of Green Technologies in Next-generation Wireline Networks: Is There Room for Energy Saving Optimization?", IEEE Communications Magazine 49 (2011) 80 - 86.
[4] A. P. Bianzino, C. Chaudet, D. Rossi, J. Rougier, "A Survey of Green Networking Research", IEEE Communication Surveys and Tutorials 14 (2012) 3 - 20.
[5] R. Bolla, R. Bruschi, F. Davoli, F. Cucchietti, "Energy Efficiency in the Future Internet: A Survey of Existing Approaches and Trends in Energy-Aware Fixed Network Infrastructures", IEEE Communication Surveys and Tutorials 13 (2011) 223 - 244.
[6] P. Mahadevan, P. Sharma, S. Banerjee, "A Power Benchmarking Framework for Network Devices", en la International Conferences on Networking (IFIP NETWORKING), 2009, pp. 795 - 808.
[7] M. Gupta, S. Singh, "Greening of the Internet", en la ACM Special Interest Group on Data Communication (SIGCOMM), 2003, pp. 19 - 26.
[8] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, J. Turner, "Openflow: Enabling Innovation in Campus Networks", ACM Computer Communication Review 38 (2008) 69 - 74.
[9] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Holzle, S. Stuart, A. Vahdat, "B4: Experience with a GloballyDeployed Software Defined WAN", en la ACM Special Interest Group on Data Communication (SIGCOMM), 2013.
[10] N. Kang, Z. Liu, J. Rexford, D. Walker, "Optimizing the "One Big Switch" Abstraction in Software-Defined Networks", en la ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT), 2013.
[11] Y. Kanizo, D. Hay, I. Keslassy, "Palette: Distributing Tables in Softwaredefined Networks", en la IEEE INFOCOM Mini-conference, 2013.
[12] B. Stephens, A. Cox, W. Felter, C. Dixon, J. Carter, "PAST: Scalable Ethernet for Data Centers", en la ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT), 2012.
[13] OpenDaylight Controller. https://www.opendaylight.org/. Último acceso: 15/09/2019.
[14] Downloads OpenDaylight. https://www.opendaylight.org/downloads. Último acceso: 15/09/2019.
[15] Mininet An Instan Virtual Network on your Laptop (or other PC) http://mininet.org/. Último acceso: 15/09/2019.
[16] iPerf the TCP, UDP and SCTP network bandwith measurement tool. https://iperf.fr/. Último acceso: 15/09/2019.
[17] Apache Maven Project. https://maven.apache.org/. Último acceso: 15/09/2019.
[18] SNDlib. http://sndlib.zib.de/. Último acceso: 15/09/2019.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

El Cuestionario como Recurso para la Mejora Docente

Guillermo Azuara Guillén, Julián Fernández Navajas, José Mª Saldaña Medina
Grupo CENIT
Departamento de Ingeniería Electrónica y Comunicaciones
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza.
gazuara@unizar.es, navajas@unizar.es, jsaldana@unizar.es

Natalia Ayuso Escuer, Jesús Alastruey Benedé
Departamento de Informática e Ingeniería de Sistemas
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza
nayuso@unizar.es, jalastru@unizar.es

***Resumen*—El cuestionario de preguntas breves es un recurso muy efectivo para mejorar los procesos de enseñanza y aprendizaje. En este trabajo mostramos diferentes experiencias que utilizan el cuestionario para la mejora docente en las asignaturas "Redes de Computadores" y "Diseño y Administración de Redes" que se imparten en el Grado en Ingeniería Informática en la Escuela Universitaria Politécnica de Teruel y la Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza. Se describe la metodología utilizada y se analizan los resultados obtenidos. Como conclusión general, se ha observado que la participación en los cuestionarios es mayor cuanto mayor peso tiene su calificación en la evaluación de la asignatura. La tasa de éxito en las asignaturas objeto de este trabajo son mayores desde que se ha incorporado esta metodología. Por ello, el presente trabajo pone de manifiesto que la utilización de cuestionarios es fundamental para fomentar el trabajo continuo de los estudiantes e implantar una evaluación formativa, lo que facilita la consecución de los objetivos de aprendizaje en estas asignaturas de redes.**

***Palabras Clave*—Cuestionarios, aprendizaje continuo, evaluación formativa, motivación, Moodle, cooperación docente.**

I. INTRODUCCIÓN

Existe una relación entre la motivación de los estudiantes y su rendimiento académico: los estudiantes más motivados son los que obtienen mejores resultados [1]. Si los docentes imparten clases interesantes, los estudiantes se mantendrán motivados y tendrán una buena actitud en las clases. También es importante, en esta misma línea, que los docentes utilicen los recursos a su alcance para obtener información sobre la comprensión de la materia por parte de los estudiantes. Esta interacción permite que tanto docente como estudiantes se centren en reforzar los aspectos que no están siendo comprendidos adecuadamente y así facilitar la consecución de las metas del aprendizaje.

Los cuestionarios se han mostrado como un recurso útil para motivar a los estudiantes y para que profesores y estudiantes obtengan retroalimentación sobre la compresión de contenidos. Un ejemplo clásico es la propuesta de los *One-minute papers*, basados en la experiencia de Charles Schwartz, quien al final de sus clases pedía a los estudiantes que escribiesen en un papel el concepto más importante que habían aprendido en la clase (*What is the most important concept you learned in class today?*) y qué cuestiones habían quedado sin resolver (*What questions remain unanswered?*) [2], [3].

El presente trabajo viene motivado por el curso "El arte de preguntar" impartido por el profesor José Antonio Rojo dentro del programa de formación del profesorado en la Universidad de Zaragoza. Dicho curso mostró el potencial del cuestionario en el proceso de enseñanza-aprendizaje. Además, el profesor José Antonio Rojo coordina un proyecto de innovación docente donde profesores de diversas áreas de conocimiento de la Universidad de Zaragoza ponen en común las acciones que están llevando a cabo en relación al uso del cuestionario como parte de una estrategia de diseño de un aprendizaje continuo y profundo. A raíz de esta cooperación entre profesores, surge la presente contribución donde presentamos nuestra experiencia con distintos tipos de cuestionarios que se han utilizado para intentar mejorar la calidad de la docencia en las asignaturas "Redes de Computadores" y "Diseño y Administración de Redes" del grado en Ingeniería Informática de la Universidad de Zaragoza.

La sección II presenta el cuestionario y algunas consideraciones a tener en cuenta a la hora de formular preguntas. En la sección III se detallan los distintos tipos de cuestionarios trabajados y la metodología utilizada. La sección IV describe el entorno de trabajo. En la sección V se muestran los resultados obtenidos. Finalmente en la sección VI se presentan las conclusiones.

II. EL CUESTIONARIO

Los cuestionarios constan de una serie de preguntas que el profesor plantea a los estudiantes para su resolución en un período de tiempo, normalmente breve. Sus principales objetivos son:

- Realizar un seguimiento de los procesos de enseñanza y aprendizaje.
- Motivar a los estudiantes.
- Valorar si los estudiantes adquieren los conocimientos previstos.
- Detectar las cuestiones que hayan podido quedar poco claras, para volverlas a trabajar si fuese necesario.
- Determinar las causas por las que algunas cuestiones han resultado más complicadas.

Moodle ofrece una actividad denominada *Cuestionario* para el diseño de este recurso. La elaboración del banco de preguntas requiere un trabajo inicial por parte del profesor, pero luego tiene importantes ventajas como la generación aleatoria de cuestionarios, la corrección automática o la generación de datos estadísticos del cuestionario. Además, la plataforma Moodle proporciona información detallada de los resultados de cada pregunta: intentos, índice de facilidad, desviación estándar, calificación aleatoria estimada, ponderación deseada, peso efectivo, índice de discriminación y eficiencia discriminativa [4].

A. *La Pregunta*

A la hora de diseñar un cuestionario, la parte más importante es determinar las preguntas que lo conforman. Serán dichas preguntas las que permitan evaluar cómo se están desarrollando los procesos de enseñanza y aprendizaje. Un rico banco de preguntas es la base de un buen cuestionario.

La plataforma Moodle permite la clasificación de las preguntas por categorías, lo que facilita su utilización posterior en las actividades que se configuren. Asimismo, ofrece múltiples opciones para elaborar preguntas. Es importante que se escoja la opción más conveniente a una pregunta formulada. Entre los distintos tipos de preguntas, las abiertas encierran una mayor dificultad de diseño. En ocasiones, esta complejidad se debe a la existencia de múltiples respuestas correctas debido al uso de términos en inglés o castellano, como por ejemplo *router* y *encaminador*, o de acrónimos en lugar de las palabras que los conforman, como por ejemplo *BGP* y *Border Gateway Protocol*. Incluso las preguntas numéricas pueden presentar dificultades en caso de respuestas que pueden variar con el tiempo, como por ejemplo, el número de sistemas autónomos que tiene una gran compañía. En este caso, es conveniente avisar a los estudiantes para que sean conscientes de esta circunstancia. La observación de una realidad cambiante como Internet forma parte del aprendizaje sobre su topología y tráfico.

III. VARIANTES DE UTILIZACIÓN DE CUESTIONARIOS

Pueden distinguirse distintos tipos de cuestionarios según sus características u objetivos. Por ejemplo, cuestionarios cortos para evaluar el aprovechamiento de una clase teórica o cuestionarios que abordan contenidos abordados en una sesión de laboratorio. Moodle permite configurar los cuestionarios para adaptarlos a los objetivos que se pretenden. En esta sección se presentan los distintos tipos de cuestionarios utilizados en este trabajo.

Todos los cuestionarios se han realizado utilizando Moodle, que permite hacer uso de dos actividades, *Cuestionario* y *Taller*. La principal diferencia es que el *Taller* permite la evaluación por pares.

A. *Cuestionario previo a práctica de laboratorio*

Se realizan evaluaciones individuales previas a la realización de las prácticas. Estos cuestionarios se realizan tras una lectura previa del guión de cada práctica y se contestan en base a la teoría. El objetivo es que el estudiante prepare adecuadamente la práctica. Algunas de las preguntas pueden ser las que se formulan al final de la práctica, en cuyo caso el estudiante deberá anticipar los resultados que va a obtener en la sesión de laboratorio y luego cotejarlos con los obtenidos. El resto de preguntas hacen referencia a aspectos de la práctica que se consideran importantes para un aprovechamiento adecuado de la sesión de trabajo. Se realizan mediante la opción *Taller* o *Cuestionario* de Moodle.

B. *Cuestionario de seguimiento de práctica de laboratorio*

Los estudiantes tiene a su disposición una serie de preguntas cortas relacionadas con los resultados de cada sesión práctica. El objetivo es que el estudiante reciba realimentación sobre su comprensión del trabajo de laboratorio. Se realizan mediante la opción *Taller* o *Cuestionario* de Moodle.

C. *Cuestionario breve al finalizar el tema*

El profesor solicita a los estudiantes que aporten preguntas tipo test con cuatro opciones de respuesta sobre el tema que se acaba de impartir. Esta actividad es voluntaria. Las preguntas recibidas son revisadas (enunciado claro y conciso, respuesta correcta, nivel adecuado de dificultad, etc.) y se modifican si es necesario antes de introducirse en un banco de preguntas que va creciendo cada curso. A continuación, Moodle genera un cuestionario seleccionando automáticamente 10 preguntas al azar de su banco de preguntas. La realización del cuestionario es voluntaria y su resultado no se considera para el cálculo de la nota final de la asignatura. Los estudiantes disponen de 5 minutos para completarlo. Se permiten dos intentos. Las preguntas pueden cambiar en cada intento.

D. *Cuestionario sobre una lectura relacionada con los contenidos teóricos*

Se plantean unos cuestionarios a realizar en casa tras una lectura previa. Esta actividad permite mostrar la transferencia del aprendizaje a otro contexto de aplicación. Las lecturas presentan casos reales donde además se pueden trabajar aspectos como la perspectiva de género o la sostenibilidad del medio ambiente. Cada cuestionario consta de en torno a 5 preguntas.

IV. ENTORNO DE TRABAJO

Se han utilizado los cuestionarios descritos en la docencia de las siguientes asignaturas:

- Redes de Computadores (RC)
- Diseño y Administración de Redes (DAR)

Ambas asignaturas, correspondientes al grado en Ingeniería Informática, se imparten en dos centros de la Universidad de Zaragoza: la Escuela Universitaria Politécnica de Teruel (EUPT) y la Escuela de Ingeniería y Arquitectura (EINA). RC y DAR se imparten en el cuatrimestre de otoño del segundo y cuarto curso del grado, respectivamente.

En las siguientes secciones no se distingue entre grupos de DAR, pero sí entre grupos de RC (RC-EUPT y RC-EINA) puesto que la utilización de los cuestionarios ha sido diferente. Los datos mostrados pertenecen al curso 2018-19, excepto los del cuestionario breve al finalizar el tema en RC-EUPT, que corresponden al curso 2017-18.

V. RESULTADOS Y VALORACIÓN

En esta sección se van a detallar los tipos de cuestionarios utilizados en los grupos RC-EUPT, RC-EINA y DAR, así como los resultados obtenidos. La Tabla I recoge las características principales de los cuestionarios utilizados, y la Tabla II muestra la participación y las calificaciones de dichos cuestionarios.

A. Redes de computadores (EUPT)

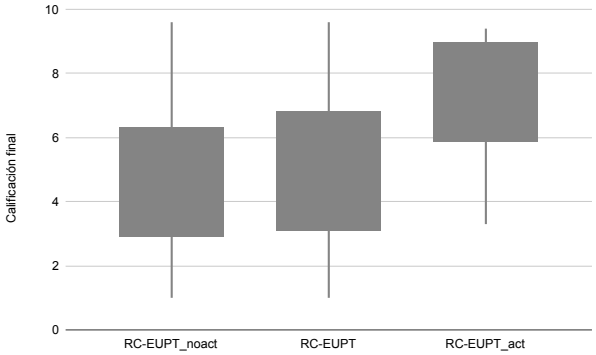
A1. Cuestionarios previos a las prácticas de laboratorio: los estudiantes tienen a su disposición una serie de preguntas sobre la práctica. Su contestación implica trabajar la teoría de la práctica y buscar información adicional. Es obligatorio realizar estos cuestionarios, cuya calificación tiene un peso del 30 % en la nota de la práctica. Respecto al curso anterior, la nota media de las prácticas ha mejorado, pasando de 7.5 a 7.8. El hecho de asistir a la práctica con el guión leído ha podido influir en este resultado. La cantidad de trabajo para el profesor ha sido muy elevada y los resultados a nivel cuantitativo no han sido muy distintos respecto a cursos anteriores. Los estudiantes más motivados completaban los cuestionarios con buena nota. Se debe mejorar el planteamiento de esta actividad.

A2. Cuestionarios de seguimiento de prácticas de laboratorio: se utilizan desde hace mucho tiempo. Es una forma clara y ordenada de recoger la información de la práctica y verificar si los estudiantes han entendido los conceptos.

A3. Cuestionarios breves al finalizar el tema: esta experiencia se realizó en el curso académico 2017-18 sobre un grupo de 29 estudiantes. Se plantearon cuestionarios tras finalizar dos temas. Los estudiantes tenían un plazo para poder realizar el cuestionario. Dentro de este plazo, cuando accedían al cuestionario tenían 5 minutos para completarlo (10 preguntas). La media de las calificaciones de los cuestionarios contestados fue de 4.6/10 (primer intento) y 7.4/10 (último intento) en el Tema 2 y 6.3/10 (primer intento) y 9.3/10 (último intento) en el Tema 3. La participación no ha sido muy alta y sólo se ha ofertado en dos de cinco temas. Se trabajará para mejorar la participación, que fue del 48 % en el Tema 2 y del 10 % en el Tema 3. El descenso de participación entre ambos temas puede obedecer a la coincidencia con trabajos y pruebas de otras asignaturas.

A4. Efecto de la participación en los resultados de la asignatura: se observa cierta correlación entre la participación activa en los cuestionarios y las probabilidades de presentarse al examen y aprobar la asignatura. La Figura 1 muestra un diagrama de caja con la distribución de las notas finales de RC-EUPT, donde la columnas izquierda, central y derecha se corresponden con la distribución de las notas finales de los estudiantes no activos, del grupo completo y de los estudiantes activos, respectivamente. Consideramos como estudiantes activos a los que han completado al menos un 75 % de los cuestionarios. El 100 % de los estudiantes activos se presentan a examen, frente al 88 % de los no activos. Por otra parte, el 67 % de los estudiantes activos que se presentan al examen aprueban, frente al 42 % de los no-activos.

Figura 1. Distribución de notas finales de RC-EUPT.



Consideramos que este tipo de cuestionario es útil para estudiantes y docentes. Para los estudiantes supone una sobrecarga de trabajo baja, ya que se complementa muy bien con la tarea de estudio del tema. El tiempo necesario para que el docente prepare estos cuestionarios es moderado, y tiene la ventaja de que se puede aprovechar material de cursos anteriores.

B. Redes de computadores (EINA)

En el curso académico 2018-19 se matricularon 123 estudiantes, que se repartieron en dos grupos de teoría (mañana y tarde) y seis de prácticas.

B1. Cuestionarios previos a las sesiones de laboratorio: se planteó un cuestionario Moodle para su realización antes de las sesiones prácticas. Para resolver las 5 preguntas de cada cuestionario, los estudiantes tienen que haber leído el guión de prácticas y la teoría relacionada. El objetivo es que el estudiante acuda al laboratorio sabiendo el trabajo a realizar y de esta forma aproveche mejor la sesión de prácticas. El plazo para realizar el cuestionario finaliza antes de la práctica, por tanto, se fijaron distintos plazos

Tabla I

CARACTERÍSTICAS DE LOS DISTINTOS TIPOS DE CUESTIONARIOS. LEYENDA: : EN CASA, : LIMITADOS POR FECHA, : TEMPORIZADO 5 MINUTOS, OP: OPTATIVO, FONDO GRIS: SU CALIFICACIÓN SE CONSIDERA PARA EL CÁLCULO DE LA NOTA FINAL DE LA ASIGNATURA.

	preLab	segLab	tema	lectura
RC-EUPT				
RC-EINA				OP
DAR				

Tabla II

PARTICIPACIÓN Y RESULTADOS DE LOS CUESTIONARIOS. *: EL NÚMERO DE ESTUDIANTES MATRICULADOS EN *tema* ES DISTINTO PUES SE CORRESPONDE CON UN CURSO ACADÉMICO DIFERENTE.

	RC-EUPT			RC-EINA			DAR
	preLab	segLab	tema*	preLab	segLab	lectura	segLab
Estudiantes matriculados	36		29	123			14
Participación media (%)	94.7	94.7	48	79.3	51.5	52.4	85.7
Nota media (sobre 10)	8.2	8.3	5.5	7.2	7.8	7.7	7.8

para los grupos que realizan las sesiones de laboratorio en distintas semanas (grupos diferenciados en Moodle). El peso de estos cuestionarios en la nota final es del 10 %. La participación media fue del 79.3 % (97.6/123), y la calificación media ha sido 7.3

Los valores de participación y calificación media fueron bastante elevados. Se consiguió que un significativo porcentaje de estudiantes acuda a las clases de prácticas con el guión leído con detenimiento, lo repercutió en provechosas sesiones de trabajo.

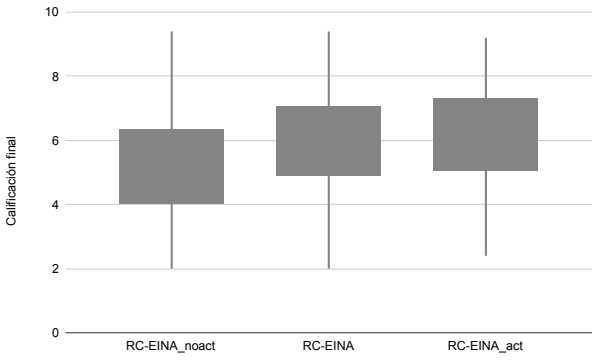
B2. Cuestionarios de seguimiento de prácticas de laboratorio: después de cada sesión de prácticas, los estudiantes tenían disponible un cuestionario Moodle con unas 20 cuestiones sobre los objetivos de la práctica. Dichas cuestiones son similares a las que se plantean en la parte de prácticas del examen escrito, tanto en la primera como en la segunda convocatoria. Aunque se obtiene una calificación, es orientativa puesto que no se considera para el cálculo de la nota final de la asignatura. La participación media ha sido del 51.5 % (63.4/123), y la calificación media ha sido 7.8.

En las sesiones de laboratorio se ha observado que la mayoría de estudiantes realizan esta actividad durante la sesión de prácticas. Posiblemente la inmediatez de la evaluación estimule esta forma de realización.

B3. Cuestionario breve sobre una lectura relacionada con los contenidos teóricos: los estudiantes realizaron dos cuestionarios Moodle tras una lectura previa en su casa. Cada cuestionario constaba de 5 preguntas. La participación media fue del 52.4 % (64.5/123), y la calificación media fue 7.7. La participación en este tipo de cuestionarios fue menor de la esperada. En general, los estudiantes que los realizaron fueron los que obtuvieron las mejores notas finales de asignatura.

B4. Efecto de la participación en los resultados de la asignatura: al igual que en el caso de RC-EUPT, se observan diferencias entre el rendimiento de los estudiantes dependiendo de si su participación en los cuestionarios es activa o no. La Figura 2 muestra un diagrama de caja con la distribución de las notas finales de RC-EINA, donde la columnas izquierda, central y derecha se corresponden con la distribución de las notas finales de los estudiantes no activos, del grupo completo y de los estudiantes activos, respectivamente. Consideramos como estudiantes activos a aquellos que han completado al menos un 75 % de los cuestionarios, en este caso, al menos 8 cuestionarios de los 12 planteados. El 93 % (64/69) de los estudiantes activos se presentan a examen, frente al 67 % (36/54) de los no activos. Las diferencias son similares si comparamos las notas finales: el 84 % (54/64) de los estudiantes activos que se presentan al examen superan la asignatura, frente al 58 % (21/36) de los no-activos.

Figura 2. Distribución de notas finales de RC-EINA.



C. Diseño y Administración de Redes

El número de estudiantes matriculados en los grupos de la EINA y la EUPT es de 12 y 2, respectivamente.

C1. Cuestionarios previos a las prácticas de laboratorio: se han realizado tres entregas individuales previas para la realización de tres prácticas. Estas entregas se realizan tras una lectura previa del guión de cada práctica y se contestan en base a la teoría anticipando lo que va a suceder en la sesión de laboratorio. La respuesta a las

preguntas es obligatoria para poder realizar la práctica, pero no se califica. También se entregan las respuestas con posterioridad a la realización de la práctica y se comparan. La realización de estas entregas tiene varias ventajas. En primer lugar, obliga a los estudiantes a leerse el enunciado de la práctica. En segundo lugar les obliga a consultar la teoría para poder responder. Por último, en el caso de que no contesten correctamente, lo que es habitual, les hace conscientes de que deben profundizar en el estudio y de que las prácticas sirven para ello. El resultado principal de obligar a entregar este informe previo es que la práctica se realiza de manera más fluida (comparada con años anteriores en que no se pedía dicha entrega previa) y se aprovecha mejor el tiempo, lo cual puede comprobarse en vista de la mejor calidad y calificación de los guiones presentados posteriormente.

VI. CONCLUSIONES

En lo que al grupo de Redes de Computadores de la EUPT y a los grupos de Diseño y Administración de Redes se refiere, se puede concluir que el cuestionario previo a la práctica ayuda a su preparación a aquellos estudiantes que realizan un trabajo continuo de la asignatura, y penaliza a los que no lo hacen. La carga de trabajo para el profesor es importante, ya que se realizan dos evaluaciones por práctica. Para los estudiantes también es más trabajo, si bien, es básicamente el que deberían hacer para preparar la práctica. Respecto al cuestionario breve al final de los temas que se hace en Redes de Computadores de la EUPT, se cree que es útil, pero se trabajará la forma de aplicarlo para que haya una mayor participación (aunque fue del 50 % , un 30 % de los que hicieron algún cuestionario no llegaron a 4, por lo que se puede suponer que no lo habían preparado adecuadamente).

En cuanto al grupo de Redes de Computadores de la EINA, se cree que los cuestionarios han sido un estímulo importante para que los estudiantes realicen trabajo continuo de la asignatura. Se estima que han influido positivamente en los resultados de la primera convocatoria del curso 2018-19 (febrero de 2019). Se han observado valores satisfactorios tanto en el número de estudiantes presentados al examen (81 % de los matriculados, 100/123), como en el número de aprobados (75 % de los presentados, 75/100). También se ha constatado que la participación activa en los cuestionarios aumenta las probabilidades de presentarse al examen (93 % de los matriculados activos, 64/69) y superar la asignatura (84 % de los presentados activos, 54/64).

De manera global, en todas las asignaturas es fundamental la realización de cuestionarios previos a las sesiones de prácticas. La participación está por encima de prácticamente el 80 % de los estudiantes matriculados y las notas medias por encima del 7.

Respecto a los cuestionarios de seguimiento de las prácticas, la participación más alta se da en DAR, con un 85.7 % frente al 51.5 % en RC-EINA. El hecho de que la calificación del cuestionario se considere como parte de la nota final de asignatura parece que incentiva la participación de los estudiantes. La baja participación en actividades cuyo resultado no tiene efecto en la nota final también se ha podido observar en el cuestionario al finalizar el tema implantado en RC-EUPT y en el cuestionario de lectura de RC-EINA (34 % y 52.4 % de los estudiantes, respectivamente).

La nota media sólo ha sido notablemente inferior en el cuestionario al finalizar el tema de RC-EUPT. Posiblemente la causa se debe a que dicho cuestionario es el único temporizado. Como conclusión, habría que plantear un período de entrenamiento para que no penalizara demasiado en la nota final y desmotivara a los estudiantes, ya que se ha constatado que en los segundos intentos la nota ha sido mucho mayor.

Dadas las experiencias realizadas, se recomienda la realización de las actividades de tipo cuestionario integradas en la plataforma de aprendizaje (en nuestro caso Moodle). Asimismo, se recomienda que las calificaciones de dichos cuestionarios se utilicen en el cálculo de la nota final para la evaluación, ya que se ha observado que incrementa la participación y motivación de los alumnos.

Este trabajo ha permitido poner en común las experiencias desarrolladas por profesores de distintos centros universitarios que están aplicando una metodología similar. Esto sin duda, va a redundar de manera positiva en el proceso de enseñanza y en el aprendizaje de los estudiantes. Por ello, para próximos cursos se seguirá cooperando para seguir mejorando la utilización de los recursos del presente trabajo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por los proyectos TIN2016-76770-R, TIN2016-76635-C2-1-R (Agencia Estatal de Investigación/Fondo Europeo de Desarrollo Regional, UE), el Gobierno de Aragón (grupos de investigación T31_17R, T45_17R y T58_17R) y FEDER 2014-2020 ”Construyendo Europa desde Aragón”.

Este artículo recoge una mejora docente aprobada en el curso 2018-19 dentro del Programa de Incentivación de la Innovación Docente en la Universidad de Zaragoza (PIIDUZ_18_094)

REFERENCIAS

[1] H. Afzal, I. Ali, M. Aslam Khan, and K. Hamid, “A study of university students’ motivation and its relationship with their academic performance,” *International Journal of Business and Manageme*, 2010.

[2] D. R. Stead, “A review of the one-minute paper,” *Active learning in higher education*, vol. 6, no. 2, pp. 118–131, 2005.

[3] G. Light, S. Calkins, and R. Cox, *Learning and teaching in higher education: The reflective professional*. Sage, 2009.

[4] “Documentación de moodle,” [Online]. Consultado mayo 2019. [Online]. Available: https://docs.moodle.org/all/es/Reporte_de_estad%C3%ADsticas_de_examen#Estad.C3.ADsticas_generales_del_examen

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Interfaz Gráfica para la Gestión SDWN de un Entorno WLAN

Pedro Fortón, José Mª Saldaña, Julián Fernández-Navajas, José Ruiz-Mas
Departamento de Ingeniería Electrónica y Comunicaciones – Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza
C/ María de Luna 1, Edif. Ada Byron, 50018 Zaragoza, España.
jsaldana@unizar.es, navajas@unizar.es, jruiz@unizar.es

***Resumen*- En este trabajo se presenta una interfaz gráfica para la gestión de una red de área local inalámbrica que integra un conjunto de puntos de acceso Wi-Fi coordinados. La interfaz interactúa con una aplicación de red que se encarga de realizar un balanceo de carga y una gestión de la movilidad. La aplicación gráfica es capaz de obtener y mostrar la información almacenada en un sistema que incluye un número de puntos de acceso, y la muestra de manera amigable para el gestor de la red. Finalmente, la aplicación permite la gestión remota del sistema, ajustando sus parámetros o realizando traspasos entre puntos de acceso.**

***Palabras Clave*- SDWN, Wireless LAN, 802.11, interfaz gráfica**

I. INTRODUCCIÓN

En los últimos años se ha experimentado un gran crecimiento en el uso de dispositivos móviles (teléfonos, tabletas, portátiles, etc.), lo que ha llevado a un aumento de la necesidad de permanecer conectados a Internet en todo momento y lugar (casa, trabajo, ocio, etc.). Aunque actualmente el despliegue de redes de datos 3G o 4G está generalizado en muchos países, las redes inalámbricas basadas en IEEE 802.11 [1], conocidas como redes Wi-Fi, siguen siendo usadas con mucha frecuencia por los usuarios, debido a su compatibilidad con diversos dispositivos, su ancho de banda, su presencia en la mayoría de los hogares y, por supuesto, su gratuidad en muchos lugares públicos (bares, restaurantes, hoteles, aeropuertos, etc.).

Actualmente el número de redes Wi-Fi que podemos encontrar en una misma localización es muy alto, compartiendo todas ellas unos recursos espectrales limitados. Esa situación se conoce como Wi-Fi *Jungle* [2], ya que los puntos de acceso *(Access Point*, APs) comparten canales y la interferencia es elevada.

En los lugares en los que es posible coordinar los puntos de acceso, como centros comerciales, empresas o edificios de servicios públicos, se suelen desplegar soluciones denominadas WLAN *Enterprise*, que son redes inalámbricas de clase empresarial que mejoran en ciertos aspectos la versión para uso personal, que se suele denominar SOHO *(Small Office / Home Office)*.

Debido a que las soluciones WLAN Enterprise actuales son tecnología propietaria y de alto coste, algunos trabajos han estudiado y desarrollado herramientas de código abierto que realizan ese servicio utilizando hardware de bajo coste, centrándose en la movilidad y el uso eficiente de los recursos inalámbricos. Algunas soluciones se basan en la adaptación de las ideas de SDN *(Software Defined Networks)* a entornos inalámbricos y se suelen denominar SDWN *(Software Defined Wireless Networks)*. Simultáneamente, se introduce el término "punto de acceso virtual ligero" *(Light Virtual Access Point*, LVAP) [3], creado para cada terminal cuando éste accede a la red y al que permanece conectado mientras le sea posible.

Gracias a esta abstracción, un mismo punto de acceso físico puede albergar varios LVAP, dando acceso a cada terminal mediante diferente BSSID *(Basic Service Set Identifier)*. Todo ello facilita la gestión del controlador a la hora de decidir qué AP da servicio a un determinado terminal o estación (STA), para lo cual debe ser capaz de "trasladar" el LVAP de un AP a otro, siguiendo el desplazamiento del terminal asociado, de una manera transparente para él [4].

La Fig. 1 ilustra el funcionamiento del LVAP: un controlador puede moverlo entre diferentes AP físicos, de forma que la estación siempre recibe las tramas de la misma dirección TA *(Transmitting Address)*, y "piensa" que siempre está conectada al mismo AP. Esto requiere un "parche" en el driver de la tarjeta inalámbrica del AP, para que se modifique TA en función de la STA destino. Es decir, un mismo AP utilizará un BSSID distinto en función de la STA destino.

Dentro del proyecto H2020 Wi-5 se desarrollaron algunas soluciones, llamadas *aplicaciones*, en las que se permitía el balanceo de carga y la gestión de la movilidad de los usuarios en escenarios con un número elevado de AP [5]. Sin embargo, el proyecto Wi-5 no planteó el desarrollo de una interfaz amigable que permitiera la monitorización del estado de la red,

o el análisis de la información generada por el sistema. La información se mostraba mediante trazas en la consola del equipo controlador de la aplicación, lo que limitaba las posibilidades de explotación de los datos generados por el sistema.

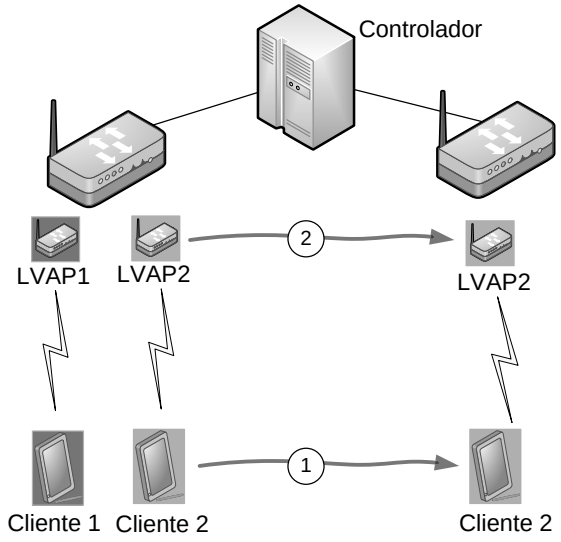


Fig. 1. Traslado de LVAP por orden del controlador.

Partiendo de la necesidad de un mejor uso de los datos generados por el conjunto de aplicaciones disponibles, el presente artículo expone el desarrollo de una interfaz gráfica usable y amigable, que presente los datos en tiempo real y permita la interacción con el sistema. Entre las múltiples aplicaciones de red desarrolladas por el proyecto Wi-5, se ha elegido desarrollar dicha interfaz gráfica para la más completa de todas, *Smart AP Selection*. Esta aplicación combina los resultados de varias aplicaciones desarrolladas durante el proceso del proyecto, y permite una gestión de la movilidad de las STA y el balanceo de carga [6].

II. DESCRIPCIÓN DEL ENTORNO ODIN WI-5

En esta sección se describirán en primer lugar los elementos que componen el sistema, para posteriormente explicar la aplicación de gestión de la movilidad y balanceo sobre la que se ha añadido la interfaz gráfica.

A. *Elementos del sistema*

Se parte de la solución Wi-5 [5], que se desarrolló como una extensión de la plataforma Odin [4], presentada en [7], que se compone de los siguientes elementos (ver Fig. 2):

- El controlador *(Controller)*, implementado como un módulo para el Floodlight OpenFlow Controller (http://www.projectfloodlight.org/), sobre el cual, el grupo de investigación de la Universidad de Zaragoza desarrolló un conjunto de aplicaciones (*Smart Functionalities* o Wi-5 *apps*) que permiten balancear la carga, gestionar la movilidad, obtener estadísticas de red en tiempo real, etc.
- Los AP, que son gestionados por el controlador, y deben realizar traspasos suaves de terminales (*seamless handoff*) mediante el uso de LVAP. Además realizan funciones de monitorización de terminales, tráfico y potencia de señal, y envían la información al controlador, que es quien decide qué hacer en cada situación.

B. *Aplicación de gestión de la movilidad y balanceo de carga*

La aplicación *Smart AP Selection* forma parte del Wi-5 *controller*, que está integrado dentro de Floodlight Controller. Esta aplicación utiliza un método proactivo para realizar una gestión de la movilidad de los usuarios (asignar cada dispositivo al AP que resulte óptimo), y simultáneamente un balanceo de carga (conseguir una distribución equilibrada entre los AP) [6].

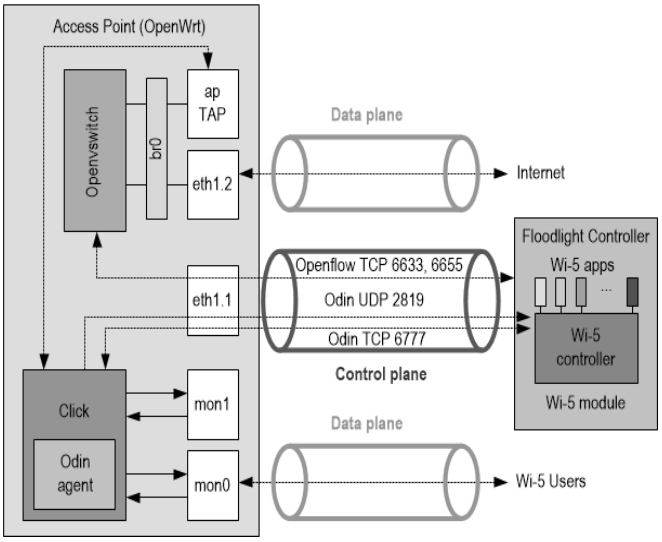


Fig. 2. Elementos del sistema Wi-5.

Mediante el protocolo Odin [4], solicita de manera periódica a los diferentes puntos de acceso de la red un escaneo en sus respectivos canales. La información obtenida en tiempo real permite al controlador, mediante diferentes algoritmos, la asignación a los AP de las estaciones conectadas a la red, según la posición en la que se encuentren *(mobility management)*, así como distribuir óptimamente el tráfico entre los AP disponibles *(load balancing).* El tiempo empleado en la iteración principal debe ser lo suficientemente breve como para permitir al controlador tomar decisiones de traspaso entre puntos de acceso para usuarios que caminan. En las pruebas realizadas durante el proyecto Wi-5, se mostró que, para seguir adecuadamente a un usuario caminando, estas decisiones se deben tomar cada 2 segundos como máximo [6].

La función principal de la aplicación realiza, en el siguiente orden, estas acciones (Fig. 3):

1. Inicializa las variables que se utilizarán a lo largo de toda la ejecución.
2. Solicita un escaneo activo a todos los AP conectados al controlador, solicitando su nivel de RSSI *(Received Signal Strength Indicator*, el indicador de fuerza de la señal recibida) recibido de cada una de las estaciones conectadas.
3. Con la información obtenida a través de cada una de las estaciones, la función crea y almacena una matriz de atenuación. Haciendo uso de esta matriz, un algoritmo organizará la red en busca de una configuración óptima.
4. Finalmente, las decisiones del algoritmo se implementarán en la red, realizando los traspasos oportunos. Estos traspasos se consiguen mediante la asignación de un LVAP a un nuevo AP. El LVAP está asociado a cada cliente, y "viaja" con él, de forma que cada cliente se comunica siempre con la MAC del LVAP.

Para calcular la matriz de atenuación (paso 3) después de cada escaneo periódico, se realiza un cálculo en el cual se obtiene la atenuación suavizada (RSSI) para cada par AP-estación [6]:

$$RSSI\ suavizada = \alpha * nuevo\ RSSI + (1-\alpha) * RSSI\ histórico \quad (1)$$

Una vez finalizado el proceso que calcula la matriz de atenuación, da comienzo el proceso que asigna las estaciones a sus puntos de acceso óptimos (paso 4), teniendo en cuenta el RSSI suavizado y el número de estaciones que hay en cada AP.

III. DESARROLLO DE UNA INTERFAZ GRÁFICA

En esta sección se explica en detalle el desarrollo y los elementos de la interfaz gráfica que se ha desarrollado, y que permite al usuario interactuar con la aplicación *Smart AP Selection*, para obtener información en tiempo real sobre el estado de la red y permitir una gestión manual de los resursos (canales, traspasos, etc.) que facilite el desarrollo de algoritmos para su automatización.

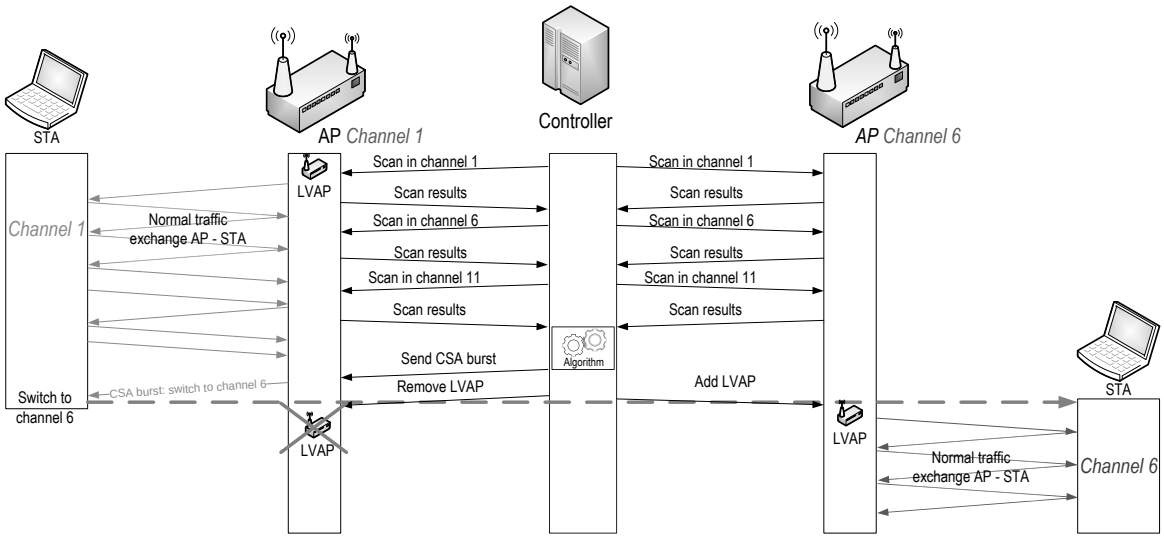


Fig. 3. Esquema de funcionamiento de *SmartAPSelection*.

El código de la interfaz desarrollada se encuentra disponible en el repositorio github del proyecto Wi-5:

https://github.com/Wi5/odin-wi5-gui-client

A. Arquitectura del sistema

La arquitectura del sistema (Fig. 4) está compuesta por diferentes capas:

- *Odin*. Contiene la aplicación *Smart AP Selection* con las funciones correspondientes de asignación de canales y terminales y traspasos mediante los diferentes algoritmos implementados y los datos recopilados mediante el intercambio de flujos de información entre el controlador y los diferentes AP.
- *Servidor*. Contiene los servicios necesarios para la extracción de datos de la capa Odin y su almacenamiento NoSql y el uso de sus funciones. Además dispone de un servidor web para interactuar con ella.
- *Cliente* (interfaz gráfica propiamente dicha), independiente del servidor, que ataca la capa servidor mediante peticiones HTTP a su servidor web.

Por otra parte, la visión global del sistema, presentada en la Fig. 5, refleja el flujo que sigue la información, desde que es generada por los puntos de acceso, hasta que es consumida por el usuario final del interfaz gráfico.

B. Pasarela de datos

El correcto funcionamiento de toda la aplicación desarrollada depende de la gestión de los datos compartidos entre las capas, para lo que se define la llamada pasarela de datos, diseñada a partir de un servicio que implementa FloodLight y que es la base de datos intermedia que alberga la información necesaria para la aplicación web.

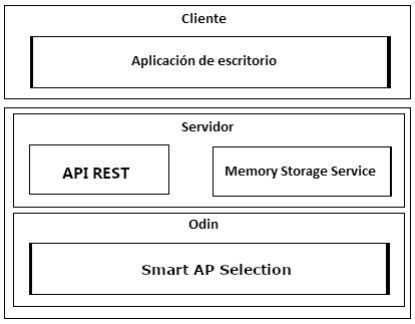


Fig. 4. Arquitectura del sistema desarrollado.

La pasarela incluye una función que inicia las tablas NoSql al inicio del proceso de la aplicación, antes de iniciar los escaneos periódicos. Posteriormente, las diferentes funciones atacarán estas tablas usando operaciones CRUD *(Create, Read, Update and Delete)*. El servicio puede almacenar objetos Java en las tablas.

La información que genera cada uno de los procesos de *Smart AP Selection*, así como la información obtenida directamente de los puntos de acceso, se vuelca sobre la pasarela, donde los datos actualizados están disponibles para los clientes que la requieran.

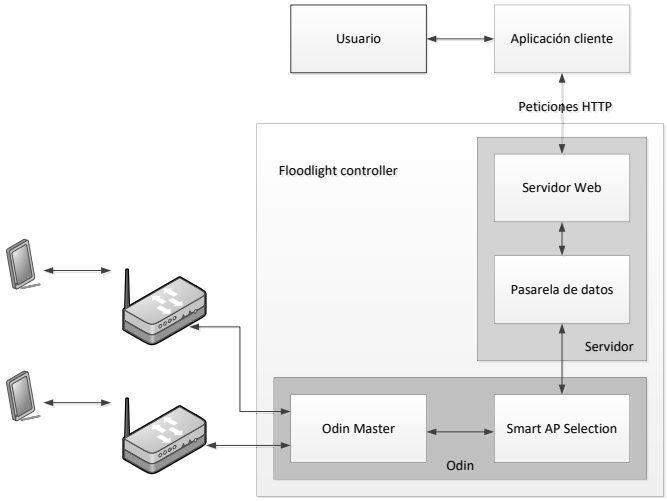


Fig. 5. Visión global de los elementos del sistema.

Para el traspaso seguro de información entre la aplicación web y *Smart AP Selection* se cuenta con el servicio de memoria presente en el sistema Floodlight. Este servicio actúa como una base de datos NoSql, dando persistencia a los datos mientras el sistema esté en ejecución. Su estilo NoSql requiere de la creación de entidades donde se guarde la información que interesa al usuario. Dichas entidades, al generarse en memoria cacheada, deben ser creadas en cada ejecución del programa. Las entidades creadas se almacenarán en forma de tablas NoSql de objetos Java.

Se busca una correcta interacción de la aplicación *Smart AP Selection* con el servicio que actúa de pasarela. Por tanto, durante el periodo entre la finalización del proceso y el reinicio del escaneo periódico, se comprueba si el cliente ha realizado alguna petición para modificar los parámetros de la aplicación y, si es así, se realiza la modificación antes de que el bucle principal se reinicie (Fig. 6).

Los pies de las figuras y de las tablas deben seguir el formato mostrado bajo la Fig. 1 y sobre la tabla I. Si es posible, utilice un formato vectorial (como EPS o PDF) para representar diagramas. Los formatos de tipo *raster* (como PNG o JPG) suelen generar ficheros muy grandes y pueden perder calidad al ampliarlos.

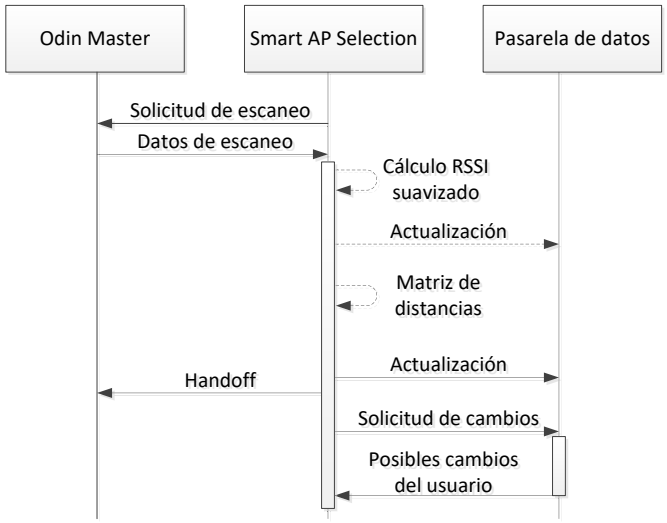


Fig. 6. Diagrama de secuencia de la información.

C. Servicio web

El servicio web, juntamente con la pasarela, actúa como intermediario entre el cliente y la aplicación *Smart AP Selection* independizando un servicio del otro. El servicio web dispone de diferentes *endpoints*:

- Obtención de todos los clientes que están actualmente conectados o lo hayan estado en algún momento.
- Estaciones que están actualmente conectadas.
- Agentes que actualmente están conectados.
- Se indica al sistema que un AP debe cambiar el canal donde está emitiendo. Se debe indicar la dirección IPv4 del AP y el canal al cual se desea cambiar.
- Solicitud del *handoff* de una estación de manera manual. El *handoff* requerirá de dos parámetros: la dirección IP del AP al que se desea mover la estación, y la dirección física MAC que identifica el cliente.
- Petición que permita que el usuario pueda solicitar un escaneo de un canal por parte del AP. Los AP pueden escanear en cualquier momento, si no se da el caso que se encuentren ya escaneando. En ese caso, este *endpoint* rescata los datos del último escaneo e informa de ellos al usuario.
- Obtención de los parámetros de la aplicación.
- Solicitud de modificación de algún parámetro.
- Envío de la matriz de potencias, que se mostrará al usuario.

Todos los *endpoint* propuestos responden mediante objetos JSON. De esta manera se estandariza la comunicación entre los servicios.

D. Cliente

El cliente se ha desarrollado como una aplicación de escritorio que mediante peticiones HTTP que obtiene la información que requiere de la aplicación *Smart AP Selection*. Con el fin de realizar una capa mantenible y robusta, se optó por un modelo modelo-vista-controlador [8] (MVC) (Fig. 7). Dicho modelo subdivide una aplicación en tres partes principales, con el objetivo de separar la lógica de negocio y sus datos, de la capa donde interactúa el usuario y todos los controladores encargados de responder a los eventos generados. El acceso a los datos se realiza mediante un conjunto de peticiones al servicio HTTP dispuesto por el controlador.

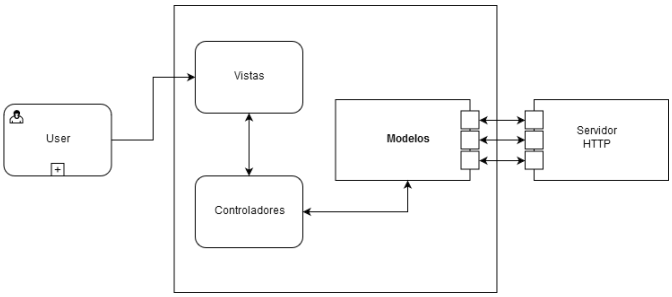


Fig. 7. Modelo MVC de la aplicación de escritorio.

Se han diseñado las diferentes vistas de la aplicación, presentando los datos de la red de una manera amigable y en tiempo real. Se diseñan tres vistas diferentes acordes a los diferentes casos de uso planteados: *network*, *clientes* y *estadísticas*. Cada una tiene el objetivo de mostrar de una manera agradable y útil la compleja información que produce el sistema.

Para la pantalla *network* (Fig. 8) se propone un gráfico piramidal, con el controlador arriba, los diferentes puntos de acceso un nivel por debajo, las estaciones (STA) otro nivel por debajo, representando las conexiones mediante líneas continuas entre los nodos. El gráfico indicará el nivel de señal (RSSI) de cada estación. De la misma manera, se utilizarán diferentes colores en la línea (variando de verde a rojo), según sea la calidad de la señal.

Tras un estudio de los requerimientos y de las posibles tecnologías compatibles con Java, se optó por la librería GraphStream (http://graphstream-project.org/), que permite mostrar grandes cantidades de datos mediante grafos en tiempo real.

En la vista *network*, el usuario puede ver el estado de la red en tiempo real, incluyendo los puntos de acceso, estaciones conectadas y su estado de conexión, así como una matriz de distancias (o atenuación) entre las estaciones conectadas y los puntos de acceso. Además, necesita de datos actualizados de manera constante, por tanto, se desarrolló una tarea que, de manera periódica, solicita los datos del último escaneo, obteniendo los últimos cambios que se han producido en la red. Esta tarea se realiza con *Timeline*, componente de la librería de JavaFX, y se ejecuta periódicamente, según se indique en el parámetro *intervalo de escaneo*. Además, desde esta vista el usuario podrá modificar algunos de los parámetros de la aplicación *Smart AP Selection*.

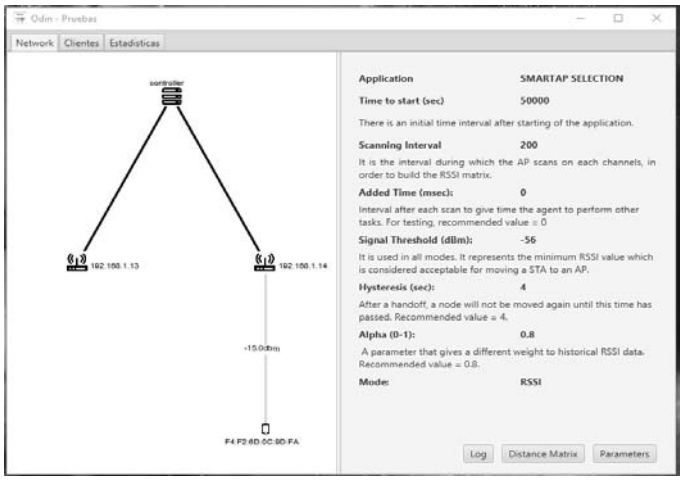


Fig. 8. Captura de pantalla de la vista general.

La matriz de distancias entre puntos de acceso y clientes (Fig. 9) se muestra con los mismos colores que se muestran en el gráfico inicial, indicando de esta manera al usuario la calidad de las diferentes conexiones entre puntos de acceso y estaciones.

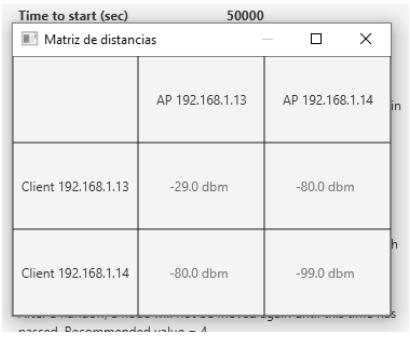


Fig. 9. Captura de pantalla del pop up con la matriz de atenuación

La vista de *estadísticas del sistema* (Fig. 10) presenta de manera gráfica los datos del último escaneo. Estas son las gráficas que se han considerado útiles:

- El tiempo en el aire de todas las estaciones conectadas a un punto de acceso escogido.
- El RSSI suavizado de cada estación para un punto de acceso escogido.
- El número de paquetes de cada estación para un punto de acceso escogido.

Estos tres parámetros permiten al administrador de la red comprobar que el algoritmo de balanceo está funcionando correctamente. Si una estación ocupa mucho tiempo en el aire, enviando pocos paquetes y con un RSSI bajo, estará probablemente posicionada en un punto de acceso no óptimo, recibiendo peor señal de la que podría tener en otro punto de acceso. En el caso opuesto, una estación con buena señal debería ser capaz de enviar grandes cantidades de paquetes ocupando un menor tiempo de aire.

En la vista se muestran los datos del último escaneo realizado por los puntos de acceso. Durante un periodo de tiempo los agentes miden el número de paquetes que envía una estación, el tiempo que ha tardado en enviar ese número de paquetes, y el nivel de señal con la cual se envían dichos paquetes.

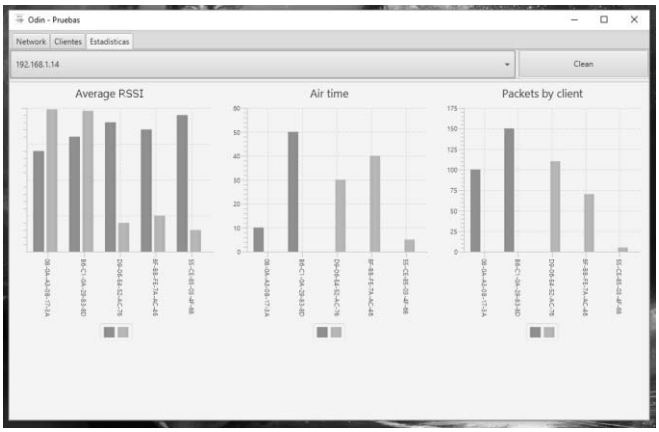


Fig. 10. Gráficas de estadísticas de la red Wi-Fi.

Finalmente, la Fig. 11 muestra la pantalla *clientes* y en particular una captura de pantalla que corresponde a la "solicitud de handoff" que el gestor del sistema puede realizar, para solicitar que una estación cambie a otro AP.

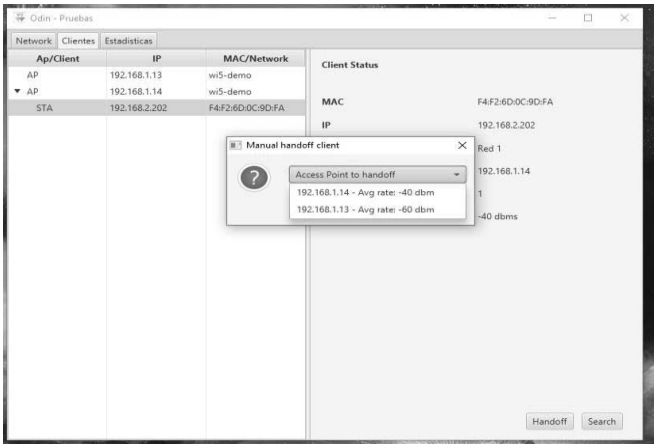


Fig. 11. Ventana emergente de solicitud de handoff.

IV. CONCLUSIONES

En este trabajo se ha presentado el desarrollo de una interfaz gráfica para la gestión de una WLAN que integra un conjunto de AP Wi-Fi coordinados. Se ha presentado el sistema, así como la aplicación Smart AP Selection, que se encarga de realizar un balanceo de carga y una gestión de la movilidad. La aplicación es capaz de obtener y mostrar la información almacenada en un sistema complejo como es Wi-5, y se muestra ahora de manera amigable para el gestor de la red. Además, la aplicación permite la gestión remota del sistema, ajustando sus parámetros o realizando traspasos entre puntos de acceso.

El trabajo realizado con las diferentes tecnologías usadas (*Java*, *JavaFx*, *FloodLight*, entre otras) y la interrelación entre ellas ha concluido en un sistema robusto, capaz de cumplir con los objetivos y requisitos requeridos, obteniendo una solución usable, atractiva, y mantenible, y que puede ser usado como base para siguientes trabajos, ya que ha sido diseñado con este objetivo.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto T31_17R del Gobierno de Aragón.

REFERENCIAS

[1] IEEE 802.11 group. “IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: IEEE 802.11 Wireless Network Management”, IEEE Std 802.11v-2011, pp. 1-433, Feb. 2011.

[2] F. den Hartog, J. de Nijs, “The role of regulation in preventing Wi-Fi over-congestion in densely populated areas,” Australian Journal of Telecommunications and the Digital Economy, vol. 5, no. 2, Jun. 2017.

[3] Y. Grunenberger, F. Rousseau, “Virtual Access Points for Transparent Mobility in Wireless LANs”, WCNC, IEEE, p. 16.

[4] J. Schulz-Zander, P. L. Suresh, N. Sarrar, A. Feldmann, T. Hhn, R. Merz, “Programmatic Orchestration of WiFi Networks”. G. Gibson, N. Zeldovich (Eds.), USENIX Annual Technical Conference, USENIX Association, pp. 347-358, 2014.

[5] Faycal Bouhafs, Michael Mackay, Alessandro Raschella, Qi Shi, Frank den Hartog, Jose Saldana, Jose Ruiz-Mas, Julian Fernandez-Navajas, Ruben Munilla, Jose Almodovar, Niels van Adrichem, “Wi-5: A Programming Architecture for Unlicensed Frequency Bands,” IEEE Communications Magazine, vol. 56, no. 12, pp. 178-185, December 2018. doi: 10.1109/MCOM.2018.1800246.

[6] Jose Saldana, Ruben Munilla, Salim Eryigit, Omer Topal, Jose Ruiz Mas, Julian Fernandez-Navajas, Luis Sequeira, “Unsticking the Wi-Fi Client: Smarter Decisions using a Software Defined Wireless Solution,” in IEEE Access, vol. 6, pp. 30917-30931, 2018.
doi: 10.1109/ACCESS.2018.2844088

[7] Luis Sequeira, Juan Luis de la Cruz, Jose Ruiz-Mas, Jose Saldana, Julian Fernandez-Navajas, Jose Luis Almodovar, “Building a SDN Enterprise WLAN Based On Virtual APs”, IEEE Communications Letters, vol. 21, no. 2, pp. 374-377, Feb. 2017.

[8] González, Y. D., & Romero, Y. F. (2012). Patrón Modelo-Vista-Controlador. Revista Telem@ tica, 11(1), 47-57.

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

Análisis de costes del despliegue de una arquitectura de red basada en SDN/NFV

Jesús Calle-Cancho*, David Cortés-Polo*, Javier Carmona-Murillo†, José-Luis González-Sánchez*, Francisco-Javier Rodríguez-Pérez†

* Centro Extremeño de Investigación, Innovación Tecnológica y Supercomputación (CénitS), Carretera Nacional 521, km 41.8 10.071 Cáceres. España.

† Departamento de Ingeniería de Sistemas Informáticos y Telemáticos, Universidad de Extremadura, Av. Universidad s/n, 10.003, Cáceres, España.

jesus.calle@cenits.es, david.cortes@cenits.es, jcarmur@unex.es, joseluis.gonzalez@cenits.es, fjrodri@unex.es

***Resumen*—En los últimos años, el incremento exponencial del tráfico de datos móviles unido al despliegue de nuevos servicios sobre las redes actuales, han propiciado que los operadores de red busquen nuevas tecnologías para adaptarse. Entre éstas, destacan especialmente las redes definidas por software (SDN) y la virtualización de funciones de red (NFV), como dos de las tecnologías que se ajustan a la naturaleza dinámica de las redes de próxima generación. Dado que la migración a estas nuevas tecnologías conlleva un coste elevado para los operadores, éstos buscan soluciones que ofrezcan un aumento de las capacidades de la red, reduciendo lo máximo posible los costes de infraestructura (CAPEX) y los costes de operación (OPEX). En este trabajo se desarrolla un análisis y evaluación de costes sobre la Red Científico Tecnológica de Extremadura (RCT), de forma que se puedan cubrir las necesidades de las redes de próxima generación, reduciendo costes y proporcionando agilidad en el despliegue de sus servicios.**

***Palabras Clave*—Redes Definidas por Software, Virtualización de Funciones de Red, Costes de operación, Costes de infraestructura.**

I. INTRODUCCIÓN

Debido al continuo desarrollo de las redes de comunicaciones móviles durante los últimos años, la realización de una gestión eficiente de la red se ha convertido en uno de los mayores desafíos en entornos de red de próxima generación. Además, los avances en las tecnologías móviles y la aparición de multitud de servicios ofertados por los operadores, han impulsado la creación de nuevos mecanismos y arquitecturas de gestión de redes de manera eficiente, capaces de soportar estas nuevas tecnologías emergentes, servicios y aplicaciones móviles. Por lo tanto, se espera que las redes de próxima generación sean capaces de proporcionar nuevos servicios de acuerdo a las demandas específicas de los usuarios. Estos avances han generado un crecimiento exponencial del tráfico global de datos móviles, experimentando un incremento del 46% entre 2017 y 2022, estimándose un consumo mensual de 77.5 exabytes en 2022 [1].

Por otro lado, todos estos cambios están teniendo un gran impacto en las estrategias económicas llevadas a cabo por los operadores de red. Cuando los operadores introducen nuevos servicios, expandiendo su infraestructura de red y/o optimizando sus recursos, tienen que tener en cuenta el coste de estas acciones. El aumento de los requisitos de eficiencia y disponibilidad por parte de los usuarios va ligado a un incremento de los costes de todo ello. Por lo tanto, las decisiones de planificación y diseño de las redes deben tener en cuenta las estimaciones de costes con la mayor precisión posible. Para ello se utilizan modelos basados en CAPEX y OPEX [2], [3].

En definitiva, los operadores de red requieren de nuevos mecanismos y soluciones para cubrir de manera eficiente (técnicamente y económicamente) las necesidades que han ido apareciendo en estos entornos de red tan cambiantes. Estos mecanismos deben ser capaces de controlar y reservar dinámicamente los recursos de la red, para proporcionar la flexibilidad requerida por los operadores de red [4]. Por ello, SDN [5] y NFV [6] son vistos como una gran oportunidad para hacer frente a la naturaleza dinámica de las redes de próxima generación y, al aprovisionamiento de servicios de manera flexible, proporcionando beneficios desde el punto de vista técnico, pero también económico, ya que estos enfoques tienen el potencial de conducir a importantes reducciones de costes de capital y costes de operación.

Este artículo presenta un análisis del despliegue de mecanismos de próxima generación basados en SDN y NFV, con el principal objetivo de reducir los costes de capital y operación. Para ello, se ha elaborado un modelo de costes basado en el CAPEX y OPEX, para analizar de manera cuantitativa el despliegue de una arquitectura

de red virtualizada basada en SDN/NFV, sobre la Red Científico Tecnológica de Extremadura .

El resto del artículo está organizado de la siguiente manera. La sección II presenta un estado del arte de SDN y NFV. En la sección III se define y presenta el modelo de costes que se utilizará para realizar el análisis del despliegue de una arquitectura de red virtualizada sobre la RCT. La sección IV muestra los resultados comparativos entre los costes necesarios para el despliegue de una arquitectura de red basada en SDN/NFV y una arquitectura de red tradicional. Por último, en la sección V se presentan las conclusiones.

II. ESTADO DEL ARTE: SDN Y NFV

La tecnología de las Redes Definidas por Software surge como un nuevo paradigma de red, cuya principal característica es la separación del plano de datos del plano de control, con el objetivo de simplificar la gestión y configuración de la red [7]. La arquitectura propuesta por SDN es considerada como una importante oportunidad para gestionar las redes tradicionales, las cuales son complejas y difícilmente gestionables. SDN proporciona una vista global de la red a través de un controlador de red centralizado. Por lo tanto, el plano de control queda centralizado en el controlador de la red que, a su vez, gestiona el plano de datos a través de protocolos abiertos como OpenFlow [8]. SDN proporciona agilidad, permitiendo a los administradores de red una gestión dinámica de flujos, optimizando recursos ante las necesidades cambiantes de las aplicaciones, que pueden tener diferentes requerimientos en cuanto a características y calidad de servicio [9].

Por otro lado, la aparición del paradigma de NFV ha sido muy importante desde el punto de vista del aprovisionamiento de servicios de telecomunicaciones. Este paradigma tiene como principal objetivo el desacoplar las funciones de red de los dispositivos físicos en los cuales se ejecutan. Además, NFV tiene el potencial de proporcionar reducciones significativas de CAPEX y OPEX, y de facilitar y flexibilizar el despliegue de nuevos servicios con más agilidad [10], permitiendo alcanzar los requisitos de baja latencia y alta fiabilidad requerida por los servicios que se ofrecerán en las futuras redes 5G [11].

Tal y como se muestra en la Fig. 1, los paradigmas SDN y NFV están íntimamente relacionados [12] y, con una integración eficiente de ambos paradigmas, se podría conseguir un importante ahorro de costes y una mayor flexibilidad en la provisión de servicios.

La virtualización de red es un método en el que los recursos físicos de una red se dividen en segmentos o *slices*. Cada *slice* está aislada de las demás y pueden compartir infraestructura física en paralelo, con las ventajas que ello conlleva. Por ello, se acuñó el termino de Cloud-RAN [13] que permite desacoplar el procesamiento de banda base de las propias estaciones base. Esta tecnología permite que el procesamiento se lleve a cabo en un centro de datos, reduciendo costes de forma significativa. Cloud-RAN ofrece un despliegue

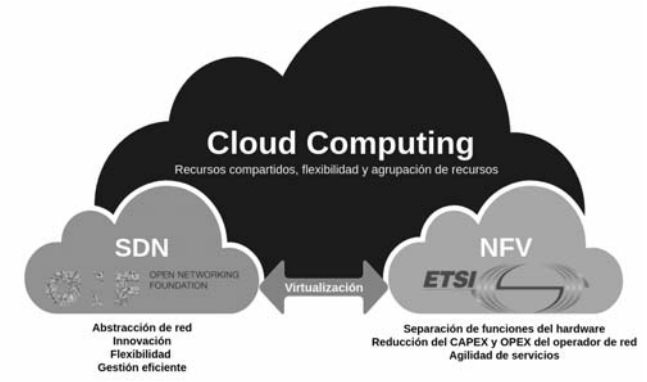


Fig. 1. Relación entre SDN y NFV.

más fácil y flexible de nuevas tecnologías, a diferencia de las estaciones base (BS) tradicionales. De esta manera, aparece el concepto de SBS (BS definida por software), en la que un número determinado de BS virtuales pueden ser desplegadas simultáneamente.

Numerosos estudios establecen que SDN y NFV conducen a una reducción significativa del CAPEX para los operadores de red [3], [12]. En relación al OPEX [14], los procesos operativos automatizados podrían reducir la intervención humana, reduciendo los costes de personal y las operaciones de red que generen fallos. En lo que respecta al CAPEX, una provisión de funciones y servicios flexible, ágil y óptima puede reducir los costes de equipamiento y posponer las inversiones [15].

III. MODELO DE COSTES

En esta sección se presenta el modelo de costes que se utilizará en secciones posteriores, para analizar y evaluar de manera cuantitativa el despliegue de una arquitectura de red virtualizada.

El modelo utilizado establece que los costes totales de un determinado operador pueden ser divididos en dos, principalmente: CAPEX y OPEX [2], [3].

- Costes de capital (CAPEX). También llamados inversiones en bienes de capitales, ya que son inversiones de capital que crean beneficios. El CAPEX está íntimamente ligado a los costes de la infraestructura fija de la empresa y se amortizan con el tiempo. En general, el CAPEX empieza a tener sentido cuando una empresa invierte en una determinada compra de un activo fijo o para añadir valor a un activo existente con una vida útil que se extiende más allá del año actual. Para un operador de red, el CAPEX incluye los costes relacionados con la compra de terrenos y/o edificios (por ejemplo, para alojar el equipamiento tecnológico), la infraestructura de red y el despliegue inicial de dicha infraestructura. Hay que tener en cuenta que la compra de equipamiento siempre tiene que ser incluida en el CAPEX, independientemente de si el pago se realiza de una vez o se extiende a lo largo del tiempo.

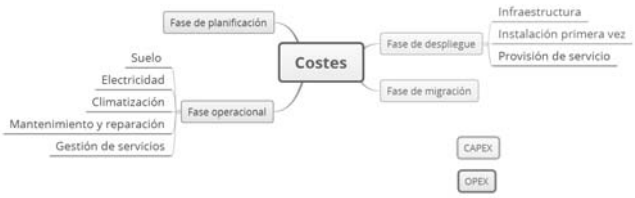


Fig. 2. Costes de infraestructura y operación considerados.

- Costes de operación (OPEX). Se trata de un coste permanente para el funcionamiento de un producto, negocio o sistema. También llamados costes de funcionamiento o costes operativos. El OPEX no contribuye a los costes del despliegue de la infraestructura; representan el coste necesario para mantener operativa la infraestructura, incluyendo los costes de operaciones técnicas y comerciales, administración, etc. Para un operador de red, el OPEX está formado por los costes de alquiler (edificio, terreno, equipos de red,etc), los costes de energía y los costes de mantenimiento, administración y operación, entre otros.

Las estimaciones del CAPEX y OPEX realizadas se han llevado a cabo utilizando el modelo que se presenta en la Fig. 2. En ella se pueden observar los costes que han sido considerados durante el estudio y análisis.

Además, para poder evaluar y comparar convenientemente los costes del despliegue de una arquitectura de red virtualizada, que haga uso de los nuevos paradigmas emergentes, es necesario analizar también los costes a los que los operadores tendrían que hacer frente si se usara una arquitectura de red tradicional. Desde el punto de vista de la tecnología, la Fig. 3 muestra una comparativa entre ellas. En realidad, una RAN (Red de Acceso Radio) virtualizada (C-RAN) se trata de una evolución de una RAN tradicional, en la que el elemento principal pasa a ser una BS virtualizada (vBS), es decir, un número determinado de BS virtuales pueden ser desplegadas simultáneamente sobre una BS física [16]. La estación base que albergue varias vBS la denominaremos SBS. Concretamente, una C-RAN consta de tres componentes principales: la antena que junto al RRH (Remote Radio Head) se ubican en una localización remota y estarán controlados por las vBS, el propio *pool* de vBS que está compuesto por procesadores de alto rendimiento que hacen uso de tecnologías de virtualización en tiempo real y una red de baja latencia para conectar las antenas al *pool* de vBS.

A. CAPEX y OPEX en una red de acceso tradicional

En esta sección se modelan los costes relacionados con una red de acceso radio tradicional (RAN), que nos permitirán evaluar el impacto relativo al despliegue de la infraestructura de red necesaria y su mantenimiento como si fuera un operador de red tradicional.

Se asume que en un área A existen un determinado número de operadores (N_{op}) que se encuentran dando cobertura a un conjunto de clientes. El número de

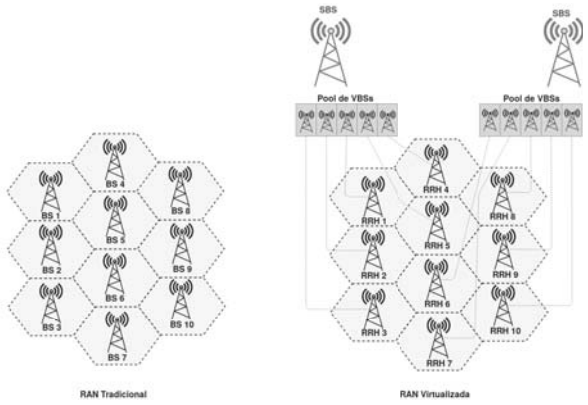


Fig. 3. Comparativa entre una RAN tradicional y una RAN virtualizada.

estaciones base requeridas para cubrir el área A depende, por un lado, del radio de cobertura de la estación base (R_{max}) y, por otro lado, del número total de usuarios en el área A (N_{UT}). Por lo tanto, el número total de estaciones base requeridas por cada operador viene dado por la Ec. 1.

$$N_{BSO} = \frac{A}{\pi \cdot R_{max}^2} \tag{1}$$

Si λ refleja la densidad de usuarios por unidad de área, se puede definir el número total de usuarios por operador (N_{UOp}) como se muestra en la Ec. 2.

$$N_{UOp} = \lambda \cdot A = \pi \cdot \lambda \cdot N_{BSO} \cdot R_{max}^2 \tag{2}$$

Siendo C_{dBS} los costes asociados al despliegue inicial de una estación base y C_{BS} los costes de cada estación base (equipamiento), se puede definir el CAPEX total de una arquitectura RAN tradicional como se muestra en la Ec. 3.

$$CAPEX_{RAN} = \sum_{i=1}^{N_{op}} \sum_{j=1}^{N_{BSO_i}} \left(C_{dBS} + C_{BS} \right)_{ij} \tag{3}$$

Obteniendo N_{BSO} de la Ec. 2 y, aplicándolo a la Ec. 3, se obtiene el coste de la infraestructura para todos los usuarios (C_{infra}), tal y como se muestra en la Ec. 4.

$$C_{infra} = \overbrace{N_{op} \cdot N_{UOp}}^{N_{UT}} \cdot \frac{C_{dBS} + C_{BS}}{\pi \cdot \lambda \cdot R_{max}^2} \tag{4}$$

Además, si se quisiera conocer el coste de la infraestructura por usuario ($C_{infra-u}$), dividiremos la Ec. 4 entre el número de usuarios totales en la red (N_{UT}), tal y como muestra la Ec. 5.

$$C_{infra-u} = \frac{C_{infra}}{N_{UT}} = \frac{C_{dBS} + C_{BS}}{\pi \cdot \lambda \cdot R_{max}^2} \tag{5}$$

Por otro lado, la Ec. 6 define el OPEX total de una arquitectura RAN tradicional. Para ello se tendrán en cuenta la energía consumida por BS (P_{BS}), el coste del kWh (C_{kWh}), los costes asociados a la operación,

mantenimiento y administración ($C_{OA\&M}$) y los costes de alquiler (C_{ALQ}).

$$OPEX_{RAN} = \sum_{i=1}^{N_{op}} \sum_{j=1}^{N_{BSO_i}} \left(P_{BS} \cdot C_{kWh} + \right. \\ \left. + C_{OA\&M} + C_{ALQ} \right)_{ij} \tag{6}$$

P_{BS} engloba la potencia de todas las antenas que conforman la estación base. Considerando que una estación base está formada por un número de antenas determinado (N_a), con una potencia determinada cada una (P_a), se puede estimar la potencia eléctrica total de cada BS, tal y como se muestra en la Ec. 7.

$$P_{BS} = N_a \cdot P_a \tag{7}$$

Por último, se define TCO como el coste total del operador, que resulta de sumar los dos costes definidos anteriormente, tal y como se muestra en la Ec. 8.

$$TCO_{RAN} = CAPEX_{RAN} + OPEX_{RAN} = \\ = \sum_{i=1}^{N_{op}} \sum_{j=1}^{N_{BSO_i}} \left(C_{dBS} + C_{BS} \right)_{ij} + \\ + \sum_{i=1}^{N_{op}} \sum_{j=1}^{N_{BSO_i}} \left(P_{BS} \cdot C_{kWh} + C_{OA\&M} + C_{ALQ} \right)_{ij} \tag{8}$$

B. CAPEX y OPEX en una red de acceso virtualizada

A continuación se modelan los costes relacionados con una red de acceso radio virtualizada, que permitirán evaluar los costes relativos a la infraestructura de red necesaria para desplegar un operador de red virtualizado.

Se asume que en un área A existe un determinado número de estaciones base virtuales vBS, desplegadas sobre SBS, que se encuentran dando cobertura a un conjunto de usuarios. El número de SBS requerido para cubrir el área A estará determinado por el radio de cobertura de la SBS (R_{max}) y por el número total de usuarios en el área A (N_{UT}), tal y como se muestra en la Ec. 9.

$$N_{SBS} = \frac{A}{\pi \cdot R_{max}^2} \tag{9}$$

Si λ refleja la densidad de usuarios por unidad de área y N_{sl} define el número de *slices* (vBS) por cada SBS, se tiene que el número total de usuarios en un área A es N_{UT} (Ec. 10).

$$N_{UT} = \lambda \cdot A = \sum_{i=1}^{N_{sl}} \left(\pi \cdot \lambda \cdot N_{SBS} \cdot R_{max}^2 \right) \tag{10}$$

Además, C_{SBS} representa el coste de una SBS. Se asume que el coste total de una SBS aumenta linealmente con respecto al número de vBS desplegadas en ella, tal y como se muestra en la Ec. 11 [17].

$$C_{SBS} = C_{BS} \cdot (1 + 0,2 \cdot (N_{sl} - 1)) \tag{11}$$

Siendo C_{dSBS} el coste asociado al despliegue inicial de una SBS, se puede definir el CAPEX total de una arquitectura C-RAN como se muestra en la Ec. 12.

$$CAPEX_{C-RAN} = C_{dSBS-total} + C_{SBS-total} = \\ = N_{SBS} \cdot (C_{dSBS} + C_{SBS}) \tag{12}$$

Obteniendo N_{SBS} de la Ec. 10 y, aplicándolo a la Ec. 12, se obtiene el coste de la infraestructura para todos los usuarios ($C_{infra-sbs}$), tal y como se muestra en la Ec. 13.

$$C_{infra-sbs} = \frac{N_{UT}}{N_{sl} \cdot \pi \cdot \lambda \cdot R_{max}^2} \cdot (C_{dSBS} + C_{SBS}) \tag{13}$$

Además, el coste de la infraestructura por usuario ($C_{infra-sbs-u}$) queda definido por la Ec. 14.

$$C_{infra-sbs-u} = \frac{C_{dSBS} + C_{SBS}}{N_{sl} \cdot \pi \cdot \lambda \cdot R_{max}^2} \tag{14}$$

Por otro lado, la Ec. 15 define el OPEX para una arquitectura C-RAN. Estará formado por los costes derivados del consumo energético de las SBS, que serán calculados a partir de la potencia de la SBS (P_{SBS}) y el coste del kWh (C_{kWh}). Además, también se tendrán en cuenta costes asociados a la operación, mantenimiento y administración ($C_{OA\&M}$) y los costes de alquiler (C_{ALQ}). Hay que tener en cuenta que cada SBS tiene más capacidades que una BS tradicional, por lo que P_{SBS} dependerá del número de vBS que se desplieguen sobre ella.

$$OPEX_{C-RAN} = \sum_{i=1}^{N_{SBS}} \left(P_{SBS} \cdot C_{kWh} + \right. \\ \left. C_{OA\&M} + C_{ALQ} \right)_{i} \tag{15}$$

Considerando que cada vBS está formada por un número de antenas determinado (N_a), se puede estimar que la potencia eléctrica de cada SBS (P_{SBS}) como se indica en la Ec. 16.

$$P_{SBS} = N_{sl} \cdot N_a \cdot P_a \tag{16}$$

Por último, se define el TCO como el coste total del operador relativo al despliegue de una RAN virtualizada. En la Ec. 17 se define este parámetro.

$$TCO_{C-RAN} = CAPEX_{C-RAN} + OPEX_{C-RAN} = \\ = \sum_{i=1}^{N_{SBS}} \left(C_{dSBS} + C_{SBS} \right)_{i} + \\ + \sum_{i=1}^{N_{SBS}} \left(P_{SBS} \cdot C_{kWh} + C_{OA\&M} + C_{ALQ} \right)_{i} \tag{17}$$

IV. CASO RCT: RESULTADOS NUMÉRICOS

Extremadura cuenta con una infraestructura de fibra óptica a la que la Junta de Extremadura denominó como Red Científico Tecnológica de Extremadura (RCT). Tiene como principal objetivo proporcionar servicios de comunicaciones avanzadas y de altas prestaciones a los centros tecnológicos y de investigación ubicados en la región.

Todo ello surgió ante la necesidad de la Comunidad Autónoma de Extremadura de contar con una infraestructura de telecomunicaciones tecnológicamente avanzada que soportara accesos telemáticos de última generación. Por ello se acometieron distintos proyectos cuya finalidad fue el tendido de una red de fibra óptica que permitiera conformar una red troncal de telecomunicaciones de alta velocidad que cumpliera con los siguientes objetivos:

- Unir las infraestructuras de los campus y/o edificios de la Universidad de Extremadura.
- Interconectar los centros sanitarios de referencia, tecnológicos y principales centros administrativos de la región.
- Dar cobertura a proyectos e iniciativas impulsadas por la Administración Regional en el ámbito de la investigación científica e innovación tecnológica.
- Fomentar el intercambio de información y conocimiento entre las universidades, centros de investigación y centros tecnológicos, llegando a ser la columna vertebral del desarrollo tecnológico en la región.
- Permitir la conexión con otras redes de investigación de ámbito nacional y europeo.

Dada la gran capacidad de transmisión de datos que ofrece la Red Científico Tecnológica de Extremadura, se exceden las necesidades de los centros tecnológicos conectados inicialmente, bajo cuya filosofía surgió la necesidad del despliegue de dicha red. Por ello, se puede reutilizar el sobredimensionamiento de red para la prestación de servicios de acceso a las infraestructuras de telecomunicaciones tanto a clientes públicos como privados. Con esta infraestructura, se dispondría de una red troncal de telecomunicaciones que conectaría los municipios de Extremadura, y que podría ser utilizada como red troncal por un operador de telecomunicaciones para ofrecer servicios de banda ancha en dichos municipios, tal y como se muestra en la Fig.4. Por lo tanto, en esta sección se evaluará de forma cuantitativa la implantación de mecanismos de nueva generación en la RCT, con el objetivo de cuantificar los costes operacionales y de capital.

Tabla I
PARÁMETROS TENIDOS EN CUENTA PARA LOS CÁLCULOS DEL CAPEX, OPEX Y TCO.

Parámetro	Valor
$C_{dBS} = C_{dSBS}$	5 000 €
C_{BS}	15 596 €
P_a	615 W
N_a	4
C_{kWh}	0,14 €
$C_{OA\&M}$	4 000 €/año
C_{ALQ}	1 000 €/año
R_{max}	2 km

La Tabla I muestra los parámetros por defecto utilizados [15], [18], [19], [20].

Además, los valores de referencia para el número de estaciones base han sido: 10, 20, 30, 50 y 100 [21]. En el caso del escenario C-RAN se ha considerado que sobre cada SBS se pueden desplegar hasta 5 vBS.

En primer lugar se han calculado los costes de capital en relación al número de estaciones base desplegadas, teniendo en cuenta los valores establecidos en la Tabla I. La Fig. 5 muestra la comparativa de los costes de capital entre los dos tipos de arquitectura desplegadas (RAN y C-RAN). En ambos casos, el CAPEX aumenta de manera lineal con respecto al número de estaciones base desplegadas. Tal y como se muestra en la Fig. 5, se pueden conseguir ahorros de los costes de capital del orden del 70% cuando se despliegan 100 BS bajo una arquitectura C-RAN, con respecto a una arquitectura RAN tradicional.

Por otro lado, la Fig. 6 refleja la evolución de los costes

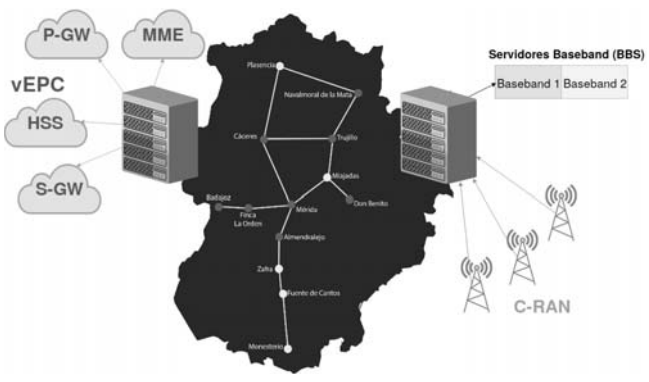


Fig. 4. Arquitectura de la red virtualizada.

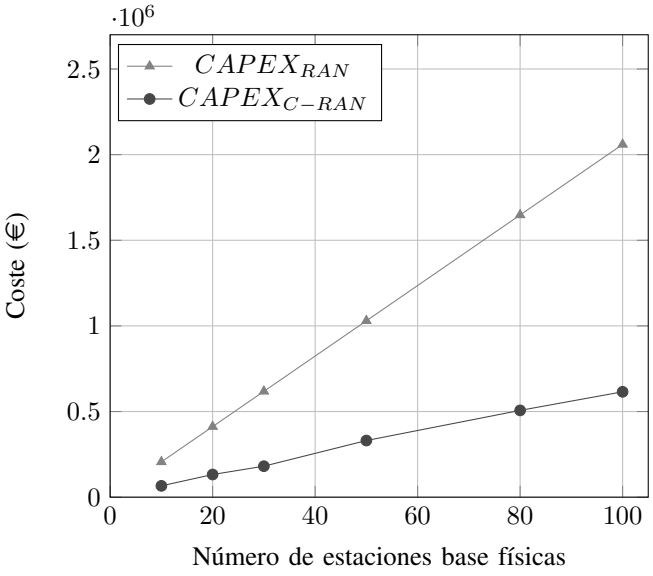


Fig. 5. Análisis de CAPEX: RAN vs C-RAN.

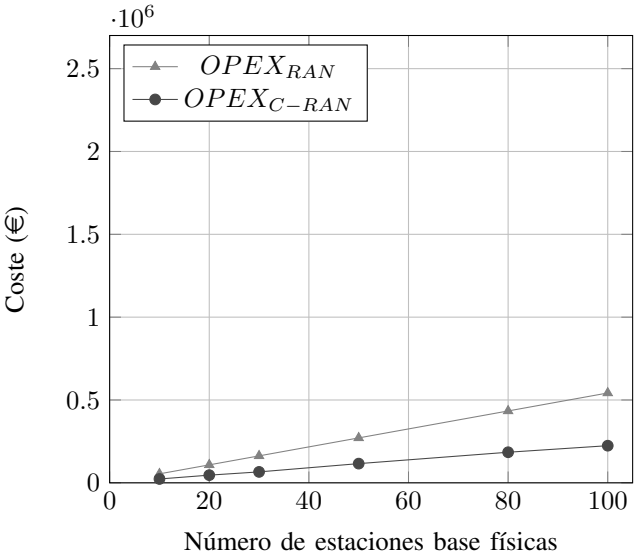


Fig. 6. Análisis de OPEX: RAN vs C-RAN.

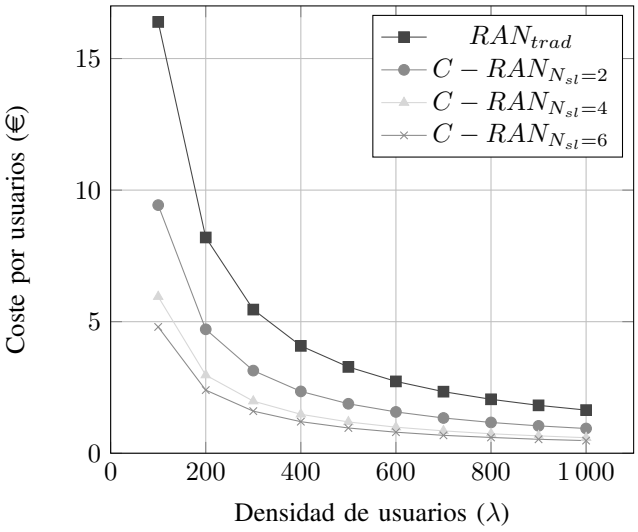


Fig. 8. Comparativa de costes relativos de RAN vs C-RAN en función de la densidad de usuarios.

de operación con respecto al número de estaciones base físicas desplegadas en un área determinada. En el caso del OPEX, al ser costes relacionados con el mantenimiento de la infraestructura, en este artículo se han considerado los costes asociados al primer año de operación, es decir, OPEX anual. En el caso de C-RAN se consiguen ahorros de OPEX de aproximadamente un 60% con respecto a la arquitectura RAN.

La Fig. 7 muestra la influencia del número de estaciones base desplegadas en el coste total del operador de red. Además, las reducciones de CAPEX y OPEX en el caso de la arquitectura C-RAN se ven reflejadas en el TCO. Este coste total está muy influenciado por los costes de capital para ambos despliegues analizados.

Por último, se ha calculado el coste de la infraestructura en función de la densidad de usuarios en un determinado área, tal y como se muestra en la Fig. 8. Para el caso de la arquitectura C-RAN, a medida que el número de *slices* por SBS aumenta, el coste de la inversión en infraestructura por usuario disminuye. Además, el coste por usuarios en la RAN tradicional es mucho mayor que el coste por usuarios en la arquitectura C-RAN. Cuando hay una mayor densidad de usuarios en un área determinada, los costes se disminuyen considerablemente, pero siempre los del despliegue de red tradicional están por encima de los del despliegue de red virtualizado.

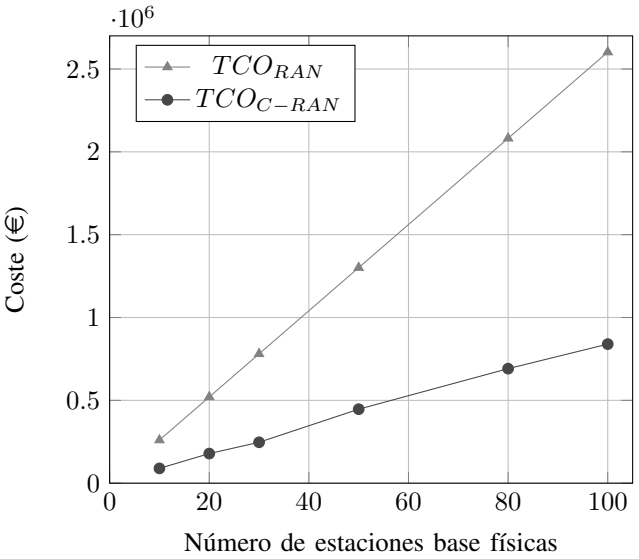


Fig. 7. Análisis de TCO: RAN vs C-RAN.

V. CONCLUSIONES Y TRABAJOS FUTUROS

Debido al rápido crecimiento del tráfico de datos móviles y los nuevos servicios demandados por los usuarios, los proveedores de servicios de telecomunicaciones han propuesto nuevas formas de prestación de servicios en términos de flexibilidad, agilidad y ahorro de costes a través de la gestión y ahorro de los costes de operación y los costes de capital.

Tecnologías como NFV y SDN han surgido para permitir que la infraestructura de red sea más flexible y abierta. Además, los proveedores de servicios pueden implementar servicios innovadores que reduzcan los costes de operación, administración y mantenimiento.

SDN y NFV están estrechamente relacionados y todas estas tecnologías van a ser muy importantes en los entornos de prestación de servicios dado que estas arquitecturas están definidas e integradas con redes de próxima generación o redes 5G.

En este trabajo se propone un enfoque para reducir el impacto de NFV y SDN en el coste total de despliegue y mantenimiento para un operador de red. Se presenta un modelo de análisis de costes que proporciona una visión técnica y económica para decidir qué arquitectura es la más acertada a la hora de desplegar las tecnologías de red y los servicios que han de ser desarrollados. Además, se establece un análisis cuantitativo para evaluar los beneficios y desventajas de estas tecnologías emergentes.

El modelo cuantitativo se utiliza para realizar experimentos que muestran una reducción significativa de los costes cuando se aplican mecanismos de virtualización en las redes de acceso. El análisis realizado revela que la arquitectura propuesta proporciona importantes ahorros de CAPEX, OPEX y TCO. Los resultados experimentales obtenidos de la comparación entre el caso tradicional y el caso virtualizado muestran que:

- El CAPEX para la red de acceso virtualizada se puede reducir hasta un 70 por ciento con respecto al caso tradicional.
- En comparación con la red de acceso tradicional, el OPEX para la red de acceso virtualizada puede reducirse hasta un 59 por ciento.
- El TCO para la arquitectura virtualizada se puede reducir hasta un 68 por ciento.
- La reducción de costes aumenta cuando mayor es el número de particiones virtuales.

Finalmente, se establecen algunas líneas de trabajo futuro con el objetivo de mejorar el diseño del modelo de costes para futuros análisis y evaluaciones:

- Analizar otros parámetros como el Retorno en la Inversión (ROI), con la finalidad de calcular el beneficio real de cualquier tipo de inversión que se haga en la infraestructura.
- Aplicar optimización matemáticas sobre el modelo de costes propuestos para ajustar parámetros críticos de los operadores de red.
- Modelar nuevos escenarios con volúmenes de tráfico estimados y demandas de capacidad para los años siguientes.
- Adecuar el modelo de costes para evaluar despliegues de celdas de redes 5G.

AGRADECIMIENTOS

Este trabajo está financiado en parte por el Fondo Europeo de Desarrollo Regional (FEDER) Programa Operativo 2014-2020 de Extremadura a través del proyecto CultivData (2018.14.02.332A.444.00), y, en parte, a través de los Fondos Europeos de Desarrollo Regional bajo el proyecto IB18003.

REFERENCIAS

[1] Cisco Systems Inc. Cisco Visual Networking Index: Forecast and Trends, 2017-2022. White Paper. February 2019.

[2] S. Verbrugge, D. Colle, M. Pickavet, P. Demeester, S. Pasqualini, A. Iselt, A. Kirstädter, R. Hülsermann, F.-J. Westphal, and M. Jäger. Methodology and input availability parameters for calculating OpEx and CapEx costs for realistic network scenarios. In *Journal of Optical Networking*, vol. 5, no. 6, pp. 509-520, 2006.

[3] B. Naudts, M. Kind, S. Verbrugge, D. Colle, and M. Pickavet. How can a mobile service provider reduce costs with software-defined networking? In *International Journal of Network Management*, vol. 26, no. 1, pp. 56-72, 2016.

[4] 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 and Interfaces. In *Computer Standards and Interfaces*, vol. 54, pp. 216-228, 2017.

[5] M. Jammal, T. Singh, A. Shami, R. Asal, and Y. Li. Software-Defined Networking: State of the Art and Research Challenges. In CoRR, vol. abs/1406.0124, 2014.

[6] Y. Li and M. Chen. Software-Defined Network Function Virtualization: A Survey In IEEE Access, vol. 3, pp. 2542-2553, 2015.

[7] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti. A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks. In *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1617-1634, 2014.

[8] OpenFlow specifications. [Online] July 2019. Available: https://www.opennetworking.org/sdn-resources/openflow

[9] W. Chin, Z. Fan, and R. Haines. Emerging technologies and research challenges for 5G wireless networks. In *IEEE Wireless Communications*, vol. 21, no. 2, pp. 106-112, 2014.

[10] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee. Network function virtualization: Challenges and opportunities for innovations In *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90-97, 2015.

[11] J. A. Cabrera, R. Schmoll, G. T. Nguyen, S. Pandi and F. H. P. Fitzek. Softwarization and Network Coding in the Mobile Edge Cloud for the Tactile Internet. In *Proceedings of the IEEE*, vol. 107, no. 2, pp. 350-363, 2019.

[12] 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. In *IEEE Communications Surveys Tutorials*, vol. 18, pp. 236-262, 2016.

[13] I.-F. Akyildiz, P. Wang, and S.-H. Lin. SoftAir: A software defined networking architecture for 5G wireless systems. In *Computer Networks*, vol. 85, pp. 1-18, 2015.

[14] E. Hernández-Valencia, S. Izzo, and B. Polonsky. How will nfv/sdn transform service provider opex?. In *IEEE Network*, vol. 29, no. 3, pp. 60-67, 2015.

[15] N. Zhang, and H. Hämmäinen. Cost Efficiency of SDN in LTE-based Mobile Networks: Case Finland. In *International Conference and Workshops on Networked Systems*, 2015.

[16] D. Pompili, A. Hajisami, and H. Viswanathan. Dynamic provisioning and allocation in Cloud Radio Access Networks (C-RANs). In *Ad Hoc Networks*, vol. 30, pp. 128-143, 2015.

[17] M. Rahman, C. Despins, and S. Affes. Analysis of CAPEX and OPEX benefits of wireless access virtualization. In *IEEE International Conference Communications Workshops (ICC)*, pp. 436-440, 2013.

[18] S.F. Yunas, J. Niemelä, M. Valkama, and T. Isotalo. Techno-economical analysis and comparison of legacy and ultra-dense small cell networks. In *Local Computer Networks*, 2014.

[19] S.F. Yunas, J. Niemelä, M. Valkama, and T. Isotalo. Relation between base station characteristics and cost structure in cellular systems. In *IEEE Personal, Indoor and Mobile Radio Communications*, vol. 4, pp. 2627-2631, 2004.

[20] LTE Multi-Platform Base Station. [Online] July 2019. Available: https://www.winncom.com/en/products/998-03-536

[21] CNMC. Información Geográfica de las estaciones base LTE de la Comunidad Autónoma de Extremadura.[Online] July 2019. Available: http://data.cnmc.es/datagraph/jsp/graph/mapa.jsp

XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: 978-84-09-21112-8

ImAc: Soluciones de Accesibilidad para Medios Inmersivos

Mario Montagud, Isaac Fraile, Einar Meyerson, Sergi Fernández

Media & Internet Area

Fundación i2CAT

C\ Gran Capità 2-4 Edifici Nexus I, Barcelona (Spain)

{mario.montagud, isaac.fraile, einar.meyerson, sergi.fernandez}@i2cat.net

***Resumen*- La accesibilidad es un requisito fundamental para cualquier servicio (multimedia). Este artículo presenta las contribuciones del proyecto europeo ImAc, que explora cómo integrar eficientemente servicios de accesibilidad y tecnologías asistivas en el ámbito de los medios inmersivos, centrándose principalmente en vídeo 360º y audio espacial. En primer lugar, se describe la metodología centrada en usuario adoptada en el proyecto, formado por un consorcio inter-disciplinar. En segundo lugar, se presenta la plataforma extremo-a-extremo siendo desarrollada en el proyecto, manteniendo la compatibiliad con las plataformas e infraestructuras broadcast actuales, así como con las tecnologías y estándares existentes. En particular, se está desarrollando la tecnología necesaria para aumentar los servicios broadcast tradicionales con contenidos inmersivos accesibles proporcionados via broadband. El artículo concluye con una breve descripción de los pilotos planificados, así como de los escenarios y evaluaciones consideradas.**

***Palabras Clave*- Accesiblidad, Audio Espacial, Realidad Virtual, Subtítulos, Vídeo 360º**

I. INTRODUCCIÓN

En los últimos años se está investigando mucho en el ámbito del consumo de vídeo online y/o contenidos TV. Como prueba de evidencia, algunos trabajos recientes han tratado de integrar servicios de accesibilidad (ej. [1]), escenarios multi-pantalla (ej. [2]) y contenidos inmersivos (ej. [3]) en servicios TV interactivos. Todas estas contribuciones responden a necesidades y demandas, y aportan un indudable valor añadido a la experiencia de consumo de contenidos. Sin embargo, todavía no se han considerado de manera conjunta. Como respuesta a ello, el proyecto europeo H2020 ImAc (Immersive Accessibility, www.imac-project.eu) trata de explorar cómo los servicios de accesibilidad se pueden integrar de manera eficiente en el ámbito de los medios inmersivos, garantizando compatibilidad con las tecnologías, formatos y recursos existentes. En cuanto a servicios de accesibilidad, ImAc considera subtítulos (incluyendo audio subtítulos y subtítulos de lectura fácil), audio descripción y lengua de signos. En cuanto a medios inmersivos, ImAc considera vídeo 360º y audio espacial. Asimismo, ImAc también considera tecnologías y funcionalidades asistivas, tales como interacción por voz, zoom, mecanismos de guiado y ayuda, etc.

Con tal de conseguir los objetivos planteados, se deben aportar soluciones reales, resultados experimentales y recomendaciones en este ámbito, dando respuesta a una serie de necesidades y desafíos de investigación, de entre los que se pueden destacar los siguientes:

- ¿Cuáles son los requisitos para posibilitar servicios inmersivos que sean realmente inclusivos y accesibles? ¿Cómo determinarlos de manera precisa?
- ¿Cómo se pueden extender las tecnologías y sistemas (inmersivas/os) actuales para soportar e integrar de manera eficiente servicios de accesibilidad?
- ¿Cómo se pueden extender las plataformas e infraestructuras broadcast para integrar de manera eficiente contenidos broadband inmersivos y accesibles?
- ¿Qué tipo de tecnologías asistivas pueden aportar en mayor medida y son susceptibles de ser adoptadas?
- ¿Qué modos y formatos de presentación de contenidos son más viables y adecuados?
- ¿Qué mecanismos y opciones de personalización deberían proporcionarse con tal de ajustarse a las necesidades y/o preferencias específicas de los usuarios?
- ¿Qué escenarios y casos de uso se verían beneficiados en mayor medida por las contribuciones del proyecto?

En especial, también se pueden destacar multitud de retos y objetivos de investigación en cuanto al consumo interactivo de contenidos se refiere:

- Los formatos para los contenidos audiovisuales tradicionales están bien definidos y especificados en estándares. Sin embargo, los formatos inmersivos se encuentran todavía en fase de especificación y evolución.
- Los formatos y contenidos inmersivos son más restrictivos en cuanto a consumo de recursos (ej. procesado, ancho de banda…) y presentación adaptativa de los mismos.
- En los contenidos tradicionales, el campo de visión viene determinado en producción. Sin embargo, en los medios inmersivos, los usuarios disponen de libertad de exploración alrededor del área 360º, por lo que los usuarios podrían perderse escenas importantes fuera de su campo de visión actual.
- Los contenidos gráficos tradicionales suelen tener un formato plano con un área de visionado delimitada. Sin

embargo, los contenidos inmersivos suelen tener formato esférico o incluso 3D. Ello, unido a la libertad de exploración citada anteriormente, supone un reto para el diseño de interfaces de usuario apropiadas, así como para delimitar el área de visionado confortable para las mismas.

- El consumo de contenidos multimedia tradicionales es muy frecuente en la sociedad actual. Sin embargo, los usuarios disponen de experiencia muy limitada en cuanto al consumo de contenidos inmersivos y, especialmente, en cuanto al uso de dispositivos de consumo para los mismos, como son las gafas de Realidad Virtual (RV) o Head Mounted Displays (HMD).
- Las modalidades de interacción en contenidos y dispositivos de consumo tradicionales (ej. mandos a distancia, teclado, ratón…) son bien conocidas y altamente adoptadas. Sin embargo, nuevas modalidades de interacción son necesarias para dispositivos de RV, como son el control por gestos, uso de controladores, etc. Esto se magnifica ante la imposibilidad de ver los controles externos, debido al uso del HMD, y de tener en cuenta los requisitos de accesibilidad.
- Existe una variedad de soluciones de accesibilidad para contenidos tradicionales, por lo que se pueden considerar las mismas como punto de referencia (baseline) cuando se proponen soluciones nuevas o mejoras. Sin embargo, no existen soluciones de accesibilidad para medios inmersivos, por lo que se parte de cero, sin referencia ni experiencia previa.

Este artículo ofrece una visión general del proyecto ImAc, destacando algunas de sus contribuciones en el ámbito y contexto de investigación introducidos en esta sección, centrándose principalmente en el consumo de contenidos. En la Sección II se repasa el estado del arte, enumerando algunos proyectos precedentes y destacando los aspectos innovadores y desafíos extra en ImAc. En la Sección III se describe la metodología centrada en usuario que se ha adoptado en el proyecto. A continuación, se describen las extensiones propuestas a las plataformas extremo-a-extremo broadcast existentes para ofrecer soluciones a los requisitos identificados. Asimismo, se destacarán las actividades de estandarización del consorcio en este ecosistema extremo-a-extremo. Finalmente, en la Sección V se concluye el artículo con una descripción de los pilotos planificados, así como de los escenarios y evaluaciones consideradas. Se detallarán algunos de los resultados obtenidos y lecciones aprendidadas.

II. Estado del Arte

ImAc cuenta con la experiencia, conocimientos y contribuciones fruto de proyectos europeos anteriores relacionados, en los que miembros del consorcio han participado (siendo incluso coordinadores de los mismos).

Por un lado, el proyecto HBB4ALL ya abarcó el tema de los servicios de accesibilidad en el emergente ecosistema multimedia híbrido broadcast broadband, bajo el paraguas del estándar Hybrid Broadcast Broadband TV (HbbTV) standard [4]. Este ecosistema híbrido permite complementar los servicios broadcast tradicionales con contenidos broadband adicionales, que pueden ser consumidos en la misma TV conectada y/o en segundas pantallas (ej. tablets, smartphones…), de manera interactiva y personalizada. HBB4ALL aprovechó dichas posibilidades para integrar servicios de accesibilidad en contenidos broadcast [1]. ImAc busca la misma historia de éxito, pero centrándose en entornos inmersivos, e incorporando dos novedades importantes: 1) soporte para conseguir Sincronización Inter-Dispositivo y streaming adaptativo basado en Dynamic Adaptive Streaming over HTTP (DASH); y 2) soporte para escenarios multi-pantalla en los que sólo intervienen tecnologías web (broadband).

Por otro lado, el proyecto ImmersiaTV desarrolló una plataforma extremo-a-extremo para posibilitar experiencias TV multi-pantalla immersivas y personalizables [3]. Mediante la introducción de nuevos mecanismos de interacción y formatos en cuanto a narrativa (storytelling), las contribuciones de ImmersiaTV permiten enriquecer los servicios TV convencionales mediante la inclusión de escenas 360º e información contextual (ej. notificaciones, vídeos superpuestos, imágenes…). Estos tipos de servicios inmersivos e interactivos se están extendiendo en ImAc, prestando más atención a los formatos inmersivos (ej. codificación avanzada para vídeos 360º, audio 3D…) y, especialmente, integrando servicios de accesibilidad y tecnologías asistivas.

III. Metodología

La incorporación de soluciones de accesibilidad en nuevas tecnologías, desde su inicio, contribuye a una implantación y adopción más efectiva. En base a esta premisa, ImAc se construye sobre tres pilares principales: 1) identificación de requisitos; 2) desarrollo e integración; y 3) validación y diseminación. El proyecto ha adoptado una metodología *user-centric* o centrada en usuario (Fig. 1), en la que los usuarios finales, profesionales y agentes interesados juegan un rol muy relevante en cada una de las etapas del proyecto, mediante la organización de talleres, sesiones de intercambio de ideas y análisis de propuestas, tests, y la asistencia a eventos. Todo ello permite identificar con gran precisión los requisitios de accesibilidad, opciones de personalización y escenarios de interés. "*El diseño para usuarios con usuarios*" representa una herramienta muy potente para identificar limitaciones y proporcionar de manera precisa las expectativas y necesidades de los perfiles de los consumidores considerados.

Los resultados de las actividades *user-centric* a su vez determinan las soluciones tecnológicas necesarias para proporcionar los requisitos identificados, así como para posibilitar los servicios y escenarios planteados.

Una premisa esencial de ImAc consiste en que las soluciones a proporcionar deben ser compatibles con las tecnologías, formatos, infrastructuras y prácticas actuales en el ámbito broadcast. Ello maximizará la re-usabilidad, inter-operabilidad y las probabilidades de implantación y adopción masiva. En este contexto, el consorcio está contribuyendo activamente a la estandarización de tecnologías, formatos y recomendaciones bajo el marco de organismos internacionales, como son: World Wide Web Consortium (W3C); Moving Picture Experts Group (MPEG); e International Organization for Standardization (ISO).

Finalmente, los pilotos y las acciones de diseminación considerados, junto a las estrategias de explotación implantadas, no sólo permiten validar las contribuciones del proyecto, sino también refinarlas en base a los resultados obtenidos y las impresiones recogidas (ver Fig. 1).

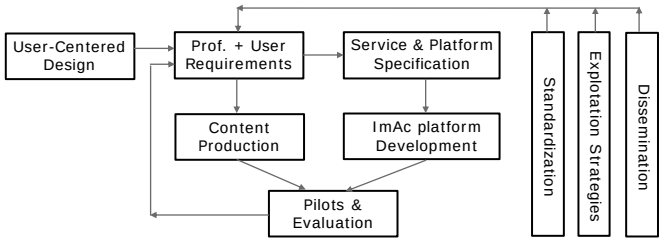


Fig. 1. Metodología centrada en Usuario utilizada en ImAc.

IV. Plataforma Extremo-a-Extremo

Con tal de conseguir los objetivos planteados, ImAc está desarrollando una plataforma extremo-a-extremo, compuesta por diferentes bloques y componentes encargados de la producción, edición, gestión, preparación, distribución y consumo de contenidos inmersivos y de accesibilidad. La Fig. 2 esquematiza los niveles lógicos / bloques principales que componen la plataforma. Los módulos o componentes funcionales en dichos bloques se indican también en la citada figura, donde color verde indica que los componentes asociados se están desarrollando en ImAc, color naranja que se tratan de componentes relevantes para ImAc, pero que han sido desarrollado en otros proyectos (ej. ImmersiaTV [3]), y color blanco se utiliza para componentes que existen normalmente en las plataformas broadcast, pero que no son esenciales para ImAc. Además, la Fig. 3 proporciona una visión general de la arquitectura de la plataforma desarrollada y de la interacción entre los componentes de la misma.

A continuación, se proporciona una breve descripción de cada bloque de la plataforma, enfatizando las extensiones propuestas en el proyecto. Las actividades de estandarización desarrolladas en este contexto también se detallan.

A. Producción de Contenidos

Los escenarios y servicios considerados en ImAc incluyen contenidos audiodivisuales a ser distribuidos vía broadcast, o bien vía plataformas de vídeo bajo demanda de los broadcasters, así como contenidos inmersivos y de accesibilidad a ser distribuidos vía broadband (ver Fig. 3). Sin embargo, en el proyecto sólo se considera el desarrollo de herramientas de producción y edición de contenidos de accesibilidad, asumiendo que los contenidos tradicionales e inmersivos han sido producidos mediante otras herramientas.

Por tanto, el bloque de *Producción de Contenidos* de la plataforma incluye una serie de herramientas web para la producción y edición de los siguientes contenidos de accesibilidad: subtítulos; audio descripción; y vídeos con interpretación de lengua de signos. Dichas herramientas proporcionan los metadatos necesarios para la señalización y presentación de dichos contenidos, asociados e integrados con contenidos inmersivos específicos, y que a su vez pueden estar relacionados con contenidos broadcast.

Más información sobre las funcionalidades del editor de subtítulos desarrollado se puede encontrar en [5].

Estos editores de contenidos de accesibilidad podrán ser utilizados por los agentes interesados bajo un modelo de distribución *Software as a Service* (SaaS).

B. Proveedor de Servicios

Este bloque de la plataforma incluye componentes para la ingesta y gestión de contenidos, como un Media Asset Management (MAM), así como para la planificación de la emisión / distribución de contenidos. Un componente clave de este bloque desarrollado en ImAc es el gestor de contenidos de accesibilidad, demonimando *Accessibility Content Manager* (ACM). En concreto, el ACM es el componente a través del cual se suministran y catalogan los contenidos inmersivos, se gestiona y valida el proceso de creación de los contenidos de accesibilidad y, finalmente, se ordena la preparación de los contenidos para su distribución.

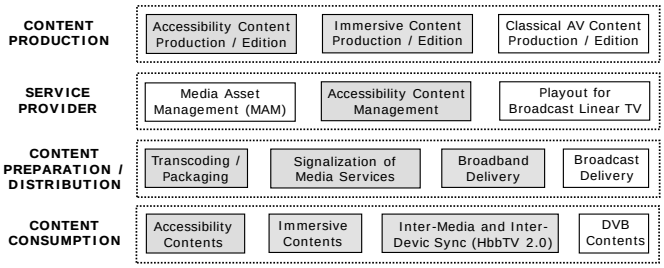


Fig. 2. Bloques y componentes principales de la plataforma ImAc.

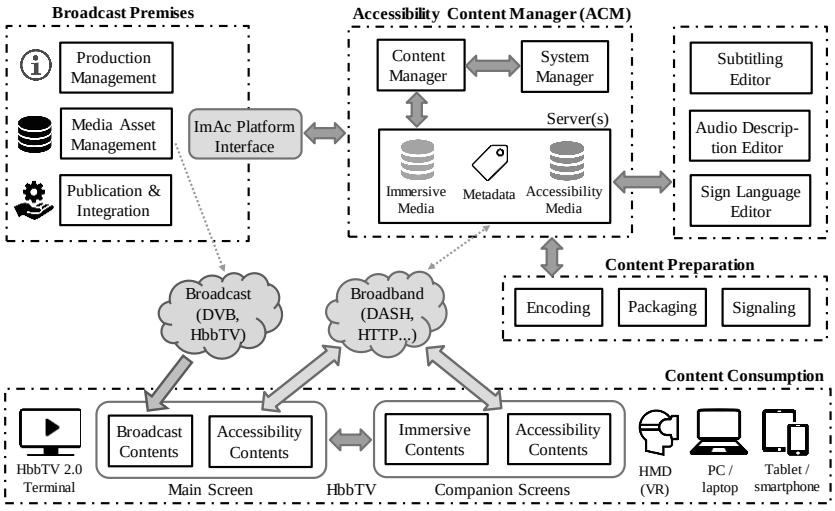


Fig. 3. Bloques y componentes principales de la plataforma ImAc.

C. Preparación y Distribución de Contenidos

Este bloque de la plataforma incluye componentes para preparar los contenidos suministrados y creados para su distribución. Dichos componentes se encargan de codificar los contenidos en múltiples calidades, de segmentarlos, señalizar su disponibilidad, y proporcionar una descripción sobre los mismos. El proyecto se centra en la distribución de contenidos vía broadband, mediante el uso de DASH y Content Delivery Networks (CDNs). Sin embargo, también se prevé el uso de DASH en coordinación con la transmisión broadcast vía Digital Video Broadcasting (DVB), bajo el paraguas del estándar HbbTV, utilizando los recursos y servicios de los broadcasters del consorcio.

En este contexto, ImAc está explorando la especificación de extensiones a los formatos, tecnologías y soluciones de señalización existentes. Dichas extensiones se definen bajo la premisa de mantener la compatibilidad (*backward compatibility*) con los estándares existentes relacionados. En particular, se están presentando propuestas para la extensión de formatos de subtítulos en W3C (ej. formato Internet Media Subtitles and Captions, IMSC) y en MPEG (ej. en Omnidirectional Media Format, OMAF), así como para la señalización de servicios de accesibilidad y parámetros sobre los mismos en MPEG (ej. en DASH).

D. Consumo de Contenidos

Este bloque está compuesto por un portal para el listado y selección de contenidos, e información sobre los mismos, así como para la configuración de ajustes iniciales (ver Fig. 4), y un reproductor web (ver Fig. 5) para la presentación de contenidos inmersivos (vídeos 360º y audio espacial) y de accesibilidad (subtítulos, audio subtítulos, audio descripción, y vídeos con interpretación de lengua de signos) de manera personalizada e interactiva. Se ha diseñado una Interfaz de Usuario amigable e intuitiva (Fig. 5), adaptada a las características de los dispositivos y contenidos RV, y que permite ser magnificada para una mejor accesibilidad.

El reproductor soporta diferentes modos de presentación para cada uno de los servicios de accesibilidad, determinados por los requisitos identificados en las actividades centradas en usuario llevadas a cabo en el proyecto.

En cuanto a la presentación de subtítulos, se proporcionan diferentes opciones de personalización, como tamaño, color, posición e idioma. Subtítulos de lectura fácil también se soportan, siguiendo un proceso de validación previa estándar. Se trata de subtítulos con una estructura semántica simplificada para mejorar la accesibilidad de personas con dificultades de lectura o cognitivas.

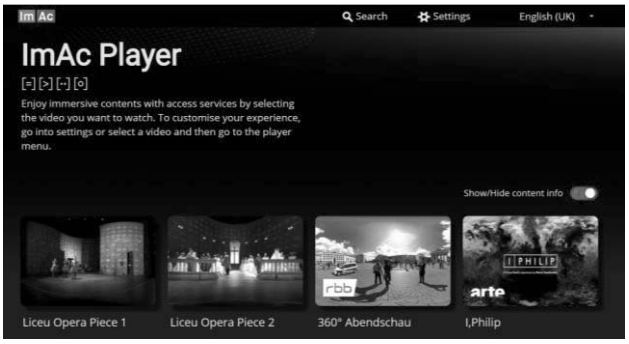


Fig. 4. Portal ImAc.

El reproductor soporta dos modos de presentación principales para los subtítulos. Por un lado, se pueden presentar siempre centrados en el campo de visión del usuario, independientemente de hacia dónde esté mirando. De este modo siempre serán visibles, siguiendo al usuario (*user-referenced*). Por otro lado, se pueden/os renderizar junto a regiones o elementos visuales específicas en el área 360º, tomando como referencia el vídeo o mundo virtual (*world-referenced*). En ambos casos, se pueden proporcionar indicadores visuales (flechas o radar) para guiar a los usuarios hacia la persona que habla. Como ejemplo, la Fig. 6 muestra la presentación de subtítulos junto a un radar que indica hacia dónde esta la persona que habla en comparación al campo de visión actual. El color del indicador se corresponde con el de los subtítulos para una mejor identificación. Asimismo, también se ha desarrollado un algoritmo que permite la presentación adaptativa de subtítulos en cuanto a su tamaño y cantidad de caracteres que se muestran, en función del área de renderizado efectiva [6].

En cuanto a la presentación de vídeos con lengua de signos, se soportan opciones de personalización en cuanto al idioma, tamaño y posición. Además, también se pueden añadir identificadores visuales, como son flechas y un radar, así como texto (nombre, descripción…) y emoticonos/pictogramas para ayudar al usuario a identificar a la persona que habla. Como ejemplo, la Fig. 7 muestra la presentación de un vídeo de lengua de signos junto con flechas y un emoticono con un rostro identificativo.

Los vídeos con lengua de signos siempre se presentan en modo superpuesto al vídeo 360º, en una posición determinada dentro del campo de visión del usuario, independiente de hacia dónde esté mirando.

Fig. 5. Interfaz de Usuario del reproductor ImAc.

Fig. 6. Presentación de subtítulos centrados en el campo de visión junto a un radar para guiar al usuario.

Fig. 7. Presentación de vídeo de lengua de signos con un emoticono para identificar a la persona que habla y con flechas para guiar al usuario.

Otro mecanismo innovador en ImAc con respecto a este servicio de accesibilidad es que se han especificado soluciones de señalización compatibles con los estándares existentes para indicar los periodos de in/actividad del o de la intérprete. De esta manera, se pueda mostrar / ocultar el vídeo de manera automática, con tal de contribuir a una posible mayor inmersión. Los potenciales beneficios están siendo evaluados, bajo la premisa que minimizar la cantidad de elementos visuales durante el visionado contribuye a una mayor inmersividad y *engagement*.

En cuanto a la presentación de audio descripción, además de explorar diferentes narrativas, ImAc saca partido de las posibilidades del audio espacial (Ambisonics), evaluando el impacto de diferentes combinaciones de espacialidad de audio para mejorar la accesibilidad e inmersión. En particular, se están explorando tres modos:

- *Modo Clásico*: voz superpuesta tradicional, como viniendo del cielo.
- *Modo Estático*: como si alguien te susurrase a la oreja.
- *Modo Dinámico*: el audio está referenciado a las acciones dinámicas que ocurren en el entorno 360º, por lo que se puede percibir dónde ocurren las mismas.

El reproductor permite la selección de los diferentes modos (si están disponibles), así como la personalización del nivel de volumen con respecto al audio principal, y la presentación de pistas de audio descripción secundarias referenciadas a escenas o elementos visuales específicas/os.

En cuanto a la presentación de audio subtítulos, el reproductor también soporta diferentes combinaciones de audio espacial (bien mediante voces humanas o sintéticas), y para ajustar su nivel de volumen, independientemente del del audio principal. Se trata de un servicio de accesibilidad menos implantado que los anteriormente citados, pero que ImAc está explorando y proponiendo soluciones para su presentación dinámica y personalizada.

El reproductor se puede ejecutar en dispositivos de consumo tradicionales (ej. TV conectadas, PC, portátiles, tablets, smartphones…) y en dispositivos de RV, como HMDs. Asimismo, el player soporta diferentes modalidades de interacción, como vía el ratón, teclado, pantalla táctil, giroscopio, o controles RV. Asimismo, el portal y reproductor ImAc permiten la interacción por voz. Por un lado, es posible controlarlos vía comandos de voz. Por otro lado, se proporciona confirmaciones de voz a la ejecución de comandos. Para ello se han desarrollado APIs para la conversión bidireccional entre sentencias de voz y controles del portal/reproductor. Dichas funcionalidades se han integrado en un servidor o pasarela intermedia entre dispositivos de control de voz y el portal/reproductor, que permite la asociación entre los mismos. Ello permite conectar el portal/reproductor con controladores de voz externos, como Amazon Echo (Alexa) o Google Home, para aquellos dispositivos de consumo que no dispongan de un motor de reconocimiento de voz o incluso de micrófono, como son muchos HMDs. Además, este tipo de asistentes de voz son cada vez más utilizados en la sociedad actual. Las APIs desarrolladas son modulares y extensibles, y permiten, por ejemplo, la conexión de controladores remotos, mediante controles táctiles, u otras modalidades de interacción futuras (ej. gestos).

Este bloque de la plataforma también incluye la tecnología necesaria para posibilitar escenarios multi-pantalla de manera sincronizada, tanto en escenarios web como en escenarios HbbTV [3].

En este contexto, ImAc está explorando la especificación de soluciones estándar para los modos de presentación de subtítulos (ej. en W3C y MPEG OMAF), para la señalización de vídeos con interpretación de lengua de signos discontinuos (ej. en MPEG DASH), así como para la especificación de directrices y recomendaciones en cuanto a Interfaces de Usuario y modalidades de interacción en RV (ej. en ISO, donde se ha iniciado un grupo de trabajo centrado en estos aspectos).

Más información sobre el portal, reproductor y sus funcionalidades se puede encontrar en [5] y en [7]. Su versión actual puede ser accedida a través del siguiente enlace: https://imac.gpac-licensing.com/player/

Asimismo, demo videos se pueden visualizar en: https://bit.ly/2Wqd336

V. Pilotos

Más allá del desarrollo de los componentes tecnológicos necesarios, ImAc engloba tanto la creación de contenidos como la planificación y ejecución de pilotos, compuestos por una serie de tests y demostraciones. Esto es esencial para evaluar el rendimiento de la plataforma y, especialmente, para determinar los beneficios aportados a los usuarios. En concreto, se han planificado dos iteraciones de pilotos, que a su vez están precedidas de sus correspondientes pre-pilotos. Los pre-pilotos persiguen validar el rendimiento adecuado de la tecnología desarrollada, validar la metodología de evaluación propuesta, así como para determinar las condiciones de tests más adecuadas a considerar en los pilotos. Estos pilotos se ejecutan en diferentes acciones en España, Alemania y Reino Unido, aunque también se consideran pilotos abiertos a través de las webs del proyecto y de proveedores de contenidos.

Las herramientas de edición de contenidos de accesibilidad y de gestión de contenidos son evaluadas por usuarios profesionales. Los contenidos creados y el reproductor multimedia son evaluados por usuarios finales, con perfiles específicos según cada servicio de accesibilidad y condiciones de test. Las evaluaciones subjetivas incluyen cuestionarios estandarizados sobre usabilidad e inmersión, así como preguntas ad-hoc sobre preferencias y aspectos de percepción.

A continuación, se detallan algunas de las acciones piloto planteadas con respecto al bloque de *Consumo de Contenidos*, destacando algunos de los resultados y/o conocimientos adquiridos.

A. Piloto 1

La primera iteración de pre-pilotos y pilotos con usuarios finales se centró en tres aspectos principales:

1. Determinar la proporción y tamaño del área de seguridad o campo de visión confortable (esto es, sub-región del campo de visión) para la presentación de elementos visuales en pantalla, cuando se utilizan HMDs. A diferencia de pantallas convencionales con relación de aspecto 16:9 en las que existe un área de seguridad determinada por la recomendación R95 de la European Broadcasting Union (EBU) [8], en los HMDs la relación de aspecto es aproximadamente 1:1, y existe una deformación en los cantos (efecto esférico). En este sentido, la hipótesis inicial se basó en que una relación de aspecto 16:9 para el área de seguridad proporcionaria una experiencia de visionado más confortable que una relación de aspecto 1:1. Por tanto, se evaluó la presentación de subtítulos y de vídeos con interpretación de lengua de signos utilizando ambas relaciones de aspecto, y con diferentes tamaños especificados como porcentaje del campo de visión efectivo (ver Fig. 8). La hipótesis se confirmó, y se concluyó además que porcentajes entre un 60% y 70% del campo de visión efectivo para el área de seguridad del proporcionan un visionado más confortable.
2. Evaluar la usabilidad de las dos versiones de Interfaz de Usuario o menús del reproductor diseñadas, incluyendo los controles necesarios para proporcionar las funcionalidades perseguidas. Uno de ellos era un menú tradicional (Fig. 9), siguiendo una filosofía similar al reproductor de Youtube. El otro ocupaba la mayor parte de la pantalla (Fig. 10), y estaba enfocado a mejorar la accesibilidad, sobretodo para usuarios con baja visión. Se determinó que los menús tenían margen de mejora, en cuanto a: su ajuste al área de seguridad; mecanismos de activación / desactivación de servicios de accesibilidad propuestos; integración en entornos de RV; implicaciones por disponder de dos menús diferentes; y usabilidad del menú tradicional en pantallas pequeñas (ej. smartphones). Con las lecciones aprendidas, se diseñó una nueva versión de menú más amigable e intuitivo (ver Fig. 5), incluyendo el uso de iconos universales de accesibilidad y funcionalidades de magnificación para mejorar la accesibilidad, sin necesidad de utilizar 2 menús diferentes. Este nuevo menú se evaluará en el piloto 2.

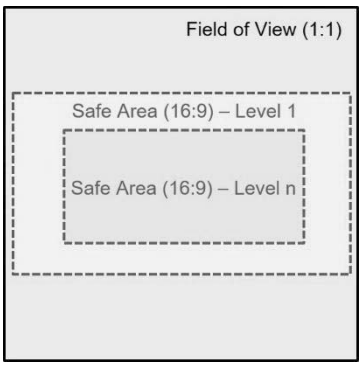


Fig. 8. Áreas de seguridad o de campo de visión confortable en HMDs.

Fig. 9. Menú tradicional (primera versión).

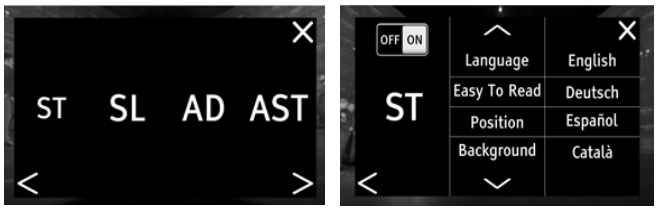


Fig. 10. Menú para mejor accesibilidad (usuarios con baja visión).

3. Modos de presentación y uso de indicadores para subtítulos. Se evaluó el primer modo de presentación de subtítulos planteado en el proyecto, en el que los subtítulos se muestran centrados en el campo de visión, independientemente de en qué zona del área 360º esté mirando el usuario. Dicha solución se comparó con la solución existente más implantada hasta el momento para vídeos 360º, consistente en presentar los subtítulos en 3 posiciones fijas, equiespaciados cada 120º [9]. Asimismo, se comparó el uso de flechas, de un radar, o de modos de auto-posicionamiento para asistir a los usuarios en la identificación de la persona que habla en cada momento. Algunos de los resultados y conclusiones que se obtuvieron se pueden consultar en [5]. Aunque se sigue investigando en estos aspectos, los resultados obtenidos revelan que la presentación de subtítulos centrados en el campo de visión es preferida por los usuarios. Con respecto a los mecanismos de guiado, las flechas son la solución más intuitiva, aunque el radar es preferido para usuarios con más contacto con los video-juegos y cuando hay diferentes personas hablando en el área 360º. Los mecanismos de auto-posicionamiento también son muy útiles. Sin embargo, en modo RV pueden resultar en mareos, debido al movimiento del mundo virtual sin movimiendo del usuario, así que se están explorando estrategias de transición adecuadas, así como su uso como mecanismo de rescate, y no de aplicación contínua.

Los resultados y lecciones aprendidas en el piloto 1 han servido para refinar las contribuciones relacionadas, y a su vez evaluarlas en la segunda iteración de pilotos, con tal de determinar las mejoras y beneficios obtenidos.

B. Piloto 2

La segunda iteración de pilotos considera todas las mejoras y extensiones a las contribuciones evaluadas en el piloto 1, así como modos de presentación e interacción para los demás servicios de accesibilidad considerados en ImAc.

En primer lugar, se va a evaluar la percepción sobre el diseño y usabilidad del portal ImAc y del menú del reproductor.

En cuanto a subtítulos, se van a evaluar modos de presentación mixtos, comparando elementos centrados en el campo de visión con elementos referenciados a objetos y escenas específicas. También se han añadido mejoras en el diseño gráfico de los indicadores visuales, así como elementos visuales adicionales (texto y emoticonos) para representar efectos de sonidos. Los beneficios aportados por la presentación de subtítulos de lectura fácil con respecto a la presentación de subtítulos tradicionales también se van a evaluar para distintos perfiles de usarios.

En cuanto a lengua de signos, audio descripción y audio subtítulos, se van a evaluar las opciones de presentación descritas en la Sección IV.D. Para el caso de audio descripción, se va a evaluar además la interacción vía el menú accesible (magnificado) y por voz, para usuarios con deficiencias de visión y ciegos.

C. *Pilotos Abiertos*

Con las lecciones aprendidas a través de las acciones piloto previas, incluyendo tests subjetivos con usuarios en entornos controlados, se pretenden ejecutar pilotos abiertos a través de los servicios de los broadcasters del consorcio.

En particular, se consideran dos escenarios. El primero de ellos consiste en proporcionar los recursos y servicios del proyecto, en cuanto al consumo interactivo de contenidos, en la web de los broadcasters y otros proveedores de contenidos para su consumo bajo demanda. El segundo escenario consiste en enriquecer programas emitidos por los canales terrestres de los broadcasters con contenidos inmersivos y de accesibilidad complementarios, a ser consumidos en escenarios multi-pantalla, de manera personalizada. Para ello, se utilizarán las funcionalidades proporcionadas por el estándar HbbTV, extendiéndolas para dar soporte a aplicaciones web capaces de reproducir contenidos inmersivos y de accesibilidad, y que pueden ser controladas por voz. Este mismo tipo de escenarios se replicarán en entornos 100% web, para que los usuarios que no dispongan de TV compatibles con las versiones recientes de HbbTV también puedan disfrutar de este tipo de servicios innovadores.

En dichas acciones piloto se evaluará el alcance en cuanto a número de usuarios y contenidos consumidos, así como el comportamiento de los usuarios en cuanto a duración de sesiones y funciones habilitadas.

VI. Conclusiones y Trabajo Futuro

Todos los ciudadanos tienen el derecho de poder utilizar las tecnologías y servicios existentes, así como de interpretar los contenidos e información disponibles. Ello contribuye a una inclusión global y a la igualdad de oportunidades. En este contexto, ImAc se centra en proporcionar soluciones de accesibilidad para los medios inmersivos, considerando tanto entorno de consumo bajo demanda online, como servicios broadcast enriquecidos con contenidos broadband.

A través de un consorcio inter-disciplinar y la adopción de una metodología centrada en usuario, ImAc está proponiendo extensiones a las plataformas broadcast existentes para posibilitar los escenarios considerados, de manera compatible con los recursos, tecnologías y formatos existentes. ImAc tiene como objetivo garantizar que las experiencias inmersivas sean inclusivas en diferentes idiomas, abordando no sólo las necesidades de personas con dificultades de audición y visión, sino también de personas con dificultades cognitivas o de aprendizaje, con bajo nivel de alfabetización y ancianos.

Las contribuciones de ImAc en este contexto se han presentado en este artículo. Se han descrito las acciones pilotos planteadas y ejecutadas, junto a los escenarios considerados, así como se han destacado algunos de los resultados obtenidos y lecciones aprendidas.

El trabajo futuro viene marcado por el plan de trabajo del proyecto hacia el piloto 2 y pilotos abiertos, así como algunas acciones complementarias que se pretender añadir (ej. considerar técnicas de procesado de señal – ej. como en [10] - y audio basado en objetos, para mejorar la accesibilidad). Asimismo, la accesibilidad en entornos de RV 3D, con libertad de exploración, navegación e interacción con los mismos, es un ámbito de investigación que se plantea como continuación del proyecto ImAc.

Agradecimientos

Este trabajo se ha financiado por el programa H2020 de la Unión Europea, en el contexto del proyecto ImAc, con referencia 761974. El trabajo de Mario Montagud ha sido adicionalmente financiado por el Ministerio de Ciencia, Innovación y Universidades, en el contexto de una Ayuda Juan de la Cierva – Incorporación, con referencia IJCI-2017-34611.

Referencias

[1] P. Orero P., C. A. Martín, M. Zorrilla, “HBB4ALL: Deployment of HbbTV services for all”, IEEE BMSB’15, Ghent (Belgium), June 2015.

[2] F. Boronat, D. Marfil, M. Montagud, J. Pastor, “HbbTV-Compliant Platform for Hybrid Media Delivery and Synchronization on Single and Multi-Device Scenarios”, IEEE Transactions on Broadcasting, 64(3), pp. 721-746, 2018.

[3] J. A. Núñez, M. Montagud, I. Fraile, D. Gómez, S. Fernández, “ImmersiaTV: an end-to-end toolset to enable customizable and immersive multi-screen TV experiences”, Workshop on Virtual Reality, co-located with ACM TVX 2018, Seoul (South Korea), June 2018.

[4] Hybrid Broadcast Broadband TV (HbbTV) 2.0.2 Specification. 2018. HbbTV Association Resource Library, https://www.hbbtv.org/resource-library, February 2018.

[5] B. Agulló, M. Montagud, I. Fraile, “Making interaction with virtual reality accessible: rendering and guiding methods for subtitles”, AI EDAM, To Appear in 2019.

[6] C. Hughes, M. Montagud, Peter tho Pesch, “Disruptive Approaches for Subtitling in Immersive Environments”, ACM TVX 2019, Manchester (UK), June 2019.

[7] M. Montagud, I. Fraile, E. Meyerson, M. Genís, S. Fernández, “ImAc Player: Enabling a Personalized Consumption of Accessible Immersive Content”, ACM TVX 2019, Manchester (UK), June 2019.

[8] European Broadcasting Union (EBU) Recommendation R95, “Safe areas for 16:9 television production”, https://tech.ebu.ch/publications/r095, Last Access in June 2019.

[9] A. Brown, J. Turner, J. Patterson, A. Schmitz, M. Armstrong, M. Glancy, “Subtitles in 360-degree Video”. ACM TVX 2017, Hilversum (The Netherlands, June 2017.

[10] Y. Zhao, E. Cutrell, C. Holz, M. R. Morris, E. Ofek, A. D. Wilson, “SeeingVR: A Set of Tools to Make Virtual Reality More Accessible to People with Low Vision”, ACM CHI’19, Glasgow (UK), May 2019.