Cost Factor Analysis of QoS in LTE/EPC Mobile Networks

William David Diego Maza, Isabelle Hamchaoui, Xavier Lagrange

Cost Factor Analysis of QoS in LTE/EPC Mobile Networks

William Diego and Isabelle Hamchaoui
Orange Labs Networks
Lannion, France
Email: {william.diego, isabelle.hamchaoui}@orange.com

Xavier Lagrange
Telecom Bretagne - IRISA D2
Rennes, France
Email: xavier.lagrange@telecom-bretagne.eu

***Abstract*—Apart from its tremendous enhanced transfer performances, the LTE/EPC mobile network is intended to distinguish itself from previous mobile technologies as an all-IP system. Nevertheless, far from usual IP QoS management policies, its QoS model inherits many characteristics of circuit oriented model of legacy standards. Additional signalling procedures are required in order to establish an end-to-end dedicated bearer for each desired QoS level. A major issue of this QoS model is the management load related to contexts and bearers established by each user equipment. This paper proposes an analytical model to evaluate the impact of this standard QoS model in terms of Context Load, Processing Load and Memory Access Rate in various LTE/EPC network elements. Simulation results are then presented to evaluate the impact of different realistic scenarios of QoS deployment.**

I. Introduction

Mobile data traffic is continuously increasing, as reported by [1]. Mobile traffic patterns have also greatly evolved. For example, social network and video/music streaming services are widely used today, but did not even exist just a few years ago. This disruptive evolution of mobile usages and services results in an important challenge for operators who struggle to differentiate themselves from competitors. In this extend, the Quality of Service (QoS) seems to be the best way forward, but at what cost?

Many studies related to mobile networks signaling load can be found in the literature, together with various analytical models. For example, [2] evaluates the signaling load related to EPC architectures, [3], [4] to security mechanisms, [5] to mobility and [6], [7] to other LTE/EPC procedures.

Nevertheless, none of these studies addresses the impact of the QoS model defined by the 3GPP standards. In [8] we introduce an analytical model to evaluate the number of incoming signalling messages at LTE/EPC network elements (Processing Load) when this QoS scheme is deployed.

As a complement to that analysis, we present now an extensive cost analysis related to QoS deployment in LTE/EPC mobile networks. We develop in the present paper an analytical model to evaluate the impact of the standard QoS model in terms of *Context Load*, *Processing Load* and *Memory Access Rate*. These three metrics will be defined below. Simulation results based on measurements on Orange networks are also presented.

II. LTE/EPC Networks

A. Quality of Service

The QoS management in LTE/EPC systems is described in [9]. A main transmission path (connection-oriented), called Evolved Packet System (EPS) bearer, must be set up between the User Equipment (UE) and the Packet Data Network GateWay (P-GW) before any traffic can be exchanged between them. Each EPS bearer provides a transport service with specific QoS attributes. When a UE attaches to the network, a default bearer with a "Best Effort" QoS is established. Other bearers can be further set up, one per QoS level. Bearers are operated in connected mode, that is established, or disconnected via signalling protocols.

B. State Machines description

The EPS Mobility Management (EMM) protocol provides procedures related to mobility over the Evolved Universal Terrestrial Radio Access Network (E-UTRAN) like access, authentication and security (e.g. Attach/detach, Tracking Area Update). There are two main EMM states described in the specifications [10], EMM-DEREGISTERED and EMM-REGISTERED.

Once a UE is registered in an LTE/EPC network (EMM-REGISTERED), the EPS Connection Management (ECM) states describes the signaling connectivity between the UE and the EPC [10]–[12]. A UE can be either in CONNECTED state (*ECM-Connected / Radio Resource Control [RRC]-Connected*) or in IDLE state (*ECM-Idle / RRC-Idle*). In the CONNECTED state, the UE has a data connectivity in the E-UTRAN (UE $\leftrightarrow$ evolved NodeB [eNB]), and a signalling connectivity in the EPC (UE$\leftrightarrow$Mobility Management Entity [MME]). After an inactivity period (RRC inactivity timer), the UE switches to IDLE state and its corresponding radio resources are released in the E-UTRAN. Thus, only the resources allocated in the EPC are kept active. Fig. 1 illustrates LTE/EPC states associated with the User-plan and Control-plan status.

In the LTE/EPC network the MME, Serving Gateway (S-GW) and eNB are critical elements, since they manage most of signaling messages; for this reason, we focus our analysis on these three elements and compute the impact of QoS model in terms of *Context Load* (D), *Processing Load* (S) and *Memory Access Rate* (L).

(a) The *Context Load* (D) evaluates the average number of simultaneous active bearers on an LTE/EPC equipment x. This metric represents the measure of memory occupancy because each active bearer is associated with an entry for its context. As the context load increases, memory overflow issues may then appear. Moreover, the latency of lookup increases, where such lookup is performed for each active bearer's context.

(b) *Processing Load* (S) evaluates the average number of incoming signalling messages per unit of time on an LTE/EPC equipment x. This value is referred to as Processing Load as each incoming message generates processing on concerning LTE/EPC equipment x.

(c) *Memory Access Rate* (L) evaluates the average number of memory accesses per unit time due to creation, modification or release of contexts, which is evaluated on an LTE/EPC equipment x.

III. LTE-EPC RELEVANT PROCEDURES

In this section we describe briefly the most common procedures in LTE/EPC networks as well as those related to QoS management.

A. Service Request procedure

When data traffic should be transmitted from or to a UE in IDLE state, either the UE or the P-GW performs the procedures specified in [10]. Thus, an ECM connection is setup in the control plane, allowing the UE to receive and send data traffic. Fig. 1 illustrates the bearer states in each LTE/EPC network segment after and before the Service Request procedure (IDLE / CONNECTED states).

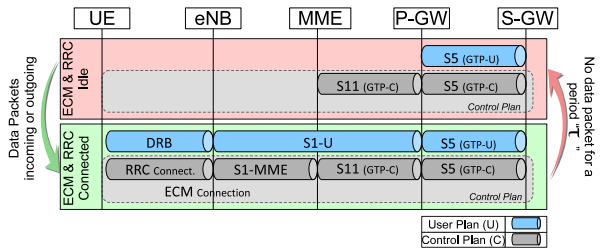

Fig. 1: Bearer states before/after Service Request procedure

B. Dedicated Bearer Activation/Deactivation

The QoS model in LTE [11] has inherited many characteristics of UMTS. The dedicated bearer setup is always initiated by the P-GW, which can be triggered by the UE itself or by data traffic destined to the UE. When downstream data traffic arrives at the P-GW, flows are classified using a Traffic Flow Template (TFT). Each flow is associated with a QoS profile which has been defined beforehand in the TFT. When the QoS profile of an arriving flow is different from "Best Effort", the P-GW initiates a Dedicated Bearer activation procedure. On the other hand, when a UE has data to be transmitted with a QoS level other than "Best Effort", the UE must request a Dedicated Bearer activation.

Once a flow supported by a dedicated bearer is completed and after the expiration of the inactivity timer [13], the dedicated bearer is deactivated. The procedure of bearer deactivation is triggered by the P-GW or the UE, which requires the exchange of messages in a similar way to the dedicated bearer activation procedure.

C. Handover procedure

The handover procedure handles mobility when a UE is in the CONNECTED state (i.e. the UE has a communication in progress). Assuming that X2 interfaces are available on every eNB, we can list two relevant scenarios:

i) Handover without S-GW relocation
ii) Handover with S-GW relocation

The handover preparation, execution and completion phases are performed as specified in [14]. During the handover execution, downstream packets are forwarded from the source eNodeB to the target eNodeB via the X2 interface. In both handover scenarios, the preparation and execution phases are identical.

It is important to highlight that the number of signaling messages used in handover procedures does not depend on the QoS levels used by a UE (i.e. number of EPS bearers per UE). On the contrary, the number of context modifications in the various involved LTE/EPC elements depends linearly on the number of EPS bearers - thus of QoS levels - per UE.

D. Tracking Area Update

While the UE is in CONNECTED state, its location is known by the LTE network at cell level. However, in IDLE state the UE location is only known at Tracking Area List (TAL) level, which is a group of Tracking Areas (TA). A TA is a group of neighbor eNBs, which are defined by the operator. A UE in IDLE state notifies the LTE network of its current TAL by sending a Tracking Area Update (TAU) message every time that it moves to another TAL. When a TAU is triggered, it might involve a MME change, but only with a low probability. Consequently, our analysis takes only into account the case of TAU without MME change.

IV. ANALYTICAL MODEL DESCRIPTION

In this section, we provide a simple analytic model to quantify the *Context Load*, *Processing Load* and *Memory Access Rate* mainly due to QoS procedures described in the previous section.

From this point forward, it is assumed that all UEs are already registered in the LTE/EPC network (EMM-REGISTERED) and a same security association is kept all over the UE life. We assume that each UE is a multitask terminal, capable of supporting n different applications (e.g. voice, video streaming, web, etc.), which can be either originated by itself or by its peer (another UE or a server). Let χ_i denotes the probability that a type-*i* session is originated by a UE.

A. Application Model

Each application is modeled as an ON-OFF state machine as shown in Fig. 2. The average type-*i* session duration (ON state) is denoted by μ_i^{-1} and the average duration of the OFF

state of type-*i* session is denoted by λ_i^{-1}. The average arrival rate α_i of type-*i* sessions is thus:

$$\alpha_i = 1/\left(\lambda_i^{-1} + \mu_i^{-1}\right) \tag{1}$$

Let $\pi_{s,i}$ be the probability that the type-*i* session state is s. The stationary probability of an application state is independent of time and thus satisfies the global balance conditions $\lambda_i \pi_{\text{off},i} = \mu_i \pi_{\text{on},i}$ and $\pi_{\text{off},i} + \pi_{\text{on},i} = 1$, we have thus:

$$\begin{cases} \pi_{\text{off},i} = \frac{\mu_i}{\lambda_i + \mu_i} \\ \pi_{\text{on},i} = \frac{\lambda_i}{\lambda_i + \mu_i} \end{cases} \tag{2}$$

B. User Equipment Model

We now consider that several applications are running on the same UE. Fig. 2 shows the applications states (ON/OFF) together with the UE states (IDLE/ CONNECTED).

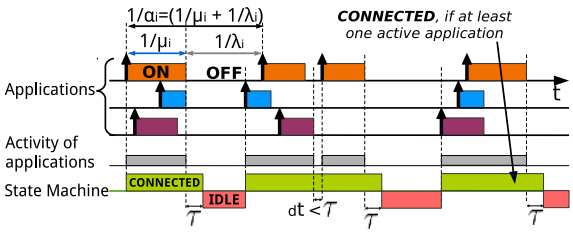

Fig. 2: Modeled RRC States

Let T_0 be the duration of a period of total inactivity (no active application) and T_1 be the duration of a period of activity (at least one application active) on this terminal. We only assume that the duration of the OFF state for application i denoted by X_i is an exponential random variable. Hence, $P(X_i > t) = e^{-\lambda_i t}$ and we have:

$$\Pr(T_0 > t) = \prod_{i=1}^{n} \Pr(X_i > t) = e^{-(\sum_{i=1}^{n} \lambda_i)t} \tag{3}$$

Let $\overline{T}_0$ and $\overline{T}_1$ be respectively the average of T_0 and T_1. We have:

$$\overline{T}_0 = 1/\sum_{i=1}^{n} \lambda_i \tag{4}$$

The stationary probability π_0 that all applications are inactive can be expressed similarly as (2):

$$\pi_0 = \frac{\overline{T}_0}{\overline{T}_0 + \overline{T}_1} \tag{5}$$

Note that π_0 can also be expressed as:

$$\pi_0 = \prod_{i=1}^{n} \pi_{\text{off},i} \tag{6}$$

Let τ be the RRC Inactivity timer, which is used in order to switch a UE from CONNECTED state to IDLE state after a period of data inactivity. The probability π_{idle} that a UE is in IDLE state is $\pi_{\text{idle}} = \Pr(T_0 > \tau)\pi_0$. Combining (2), (3) and (6) we get:

$$\pi_{\text{idle}} = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i} e^{-(\sum_{i=1}^{n} \lambda_i)\tau} \tag{7}$$

C. Context Time Duration

Let T_{idle} be the time during which a UE is in IDLE state and $T_{\text{connected}}$ be the time during which a UE is in CONNECTED state. We have:

$$\Pr(T_{\text{idle}} > t) = \Pr(T_0 > t + \tau \mid T_0 > \tau) \tag{8}$$

Due to the memory-less propriety of exponential random variable T_0, we thus have $\Pr(T_{\text{idle}} > t) = \Pr(T_0 > t)$. From (3) we thus have:

$$\overline{T}_{\text{idle}} = 1/\sum_{i=1}^{n} \lambda_i \tag{9}$$

The stationary probability π_{idle} that a UE is in IDLE state can be also expressed as $\pi_{\text{idle}} = \overline{T}_{\text{idle}}/(\overline{T}_{\text{idle}} + \overline{T}_{\text{connected}})$. Therefore, combining equations (7) and (9) we get:

$$\overline{T}_{\text{connected}} = \overline{T}_{\text{idle}} \frac{1 - \pi_{\text{idle}}}{\pi_{\text{idle}}} \tag{10}$$

D. System Model

The average number of transitions from at least one active application to non-active application is $\frac{1}{\overline{T}_0 + \overline{T}_1}$ per unit of time. Let β be the average number of transitions from CONNECTED to IDLE states per unit of time, we thus have:

$$\beta = \frac{\Pr(T_0 > \tau)}{\overline{T}_0 + \overline{T}_1} \tag{11}$$

Combining equations (2), (3), (4), (5) and (6) we get:

$$\beta = \sum_{i=1}^{n} \lambda_i \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i} e^{-(\sum_{i=1}^{n} \lambda_i)\tau} \tag{12}$$

Let P_{ue} be the probability that a session is originated by the UE. It may be estimated as:

$$P_{\text{ue}} = \frac{\sum_{i=1}^{n} \alpha_i \chi_i}{\sum_{i=1}^{n} \alpha_i} \tag{13}$$

E. Dedicated Bearer Model

A bearer-inactivity timer ϕ is set for each dedicated bearer and is managed at P-GW level [13], [15]. Once timer ϕ expires, the P-GW triggers the Dedicated Bearer deactivation procedure.

Let j be an application supported by a dedicated bearer and following the previously defined applications model. Let Ω_j be the average number of transitions from ON to OFF states per unit time of the dedicated bearer carrying type-*j* application. Using the same procedure as in the equation (11), we thus have:

$$\Omega_j = \frac{\lambda_j \mu_j}{\lambda_j + \mu_j} e^{-\lambda_j \phi} \tag{14}$$

Fig. 3 shows the modeled behaviour of the dedicated bearer (Activation/Deactivation).

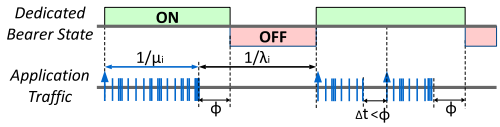

Fig. 3: Behaviour of a Dedicated Bearer

F. Mobility Model

Let C_{e} be the total number of eNBs in the evaluated region, A_{c} be the coverage area of a eNB and ρ be the UE density (number of UEs per unit area). For the sake of simplicity, each eNB is represented by a cell, which is assumed uniform (disk). Let C_{ta} be the number of eNBs per TAL. In order to compute the messages load due to handover events we use the fluid-flow model [16] to determine the mobile crossing rate out of an enclosed region with perimeter length l. We assume that UEs have an average speed V. Based in [17], we estimate the rate of border crossings per UE for a given eNB coverage area as follows:

$$\upsilon = \frac{Vl}{A_{\mathrm{c}}\pi} \tag{15}$$

Let P_{rel} be the S-GW relocation probability. We further assume that each S-GW serves a TAL; then P_{rel} can be well approximated by $1/\sqrt{C_{\mathrm{ta}}}$.

Let λ_{ue} be the arrival rate of UEs in an eNB area. Each eNB area can be seen as an infinite capacity system with random arrival of customers (UEs) with rate λ_{ue} and service rate υ. Hence, using Little's law, we can write:

$$\lambda_{\mathrm{ue}} = \upsilon \mathrm{E}(N) \tag{16}$$

where $\mathrm{E}(N)$ is the average number of customers in the eNB area and can be computed as $\mathrm{E}(N)$=ρA_{c}. In steady-state, the arrival rate of UEs in the eNB area is also the departure rate of UEs from the eNB area. The probability that a UE has at least one new session (created in current eNB) at any time is $(1-\pi_{\mathrm{idle}})$. This is the same probability with which a UE carries at least one new session while departing form the eNB area. Let σ_{c} be the arrival rate of a context on an eNB. Therefore, from (7) and (16), σ_{c} is computed as follows:

$$\sigma_{\mathrm{c}} = \lambda_{\mathrm{ue}}(1 - \pi_{\mathrm{idle}}) \tag{17}$$

V. Cost Analysis

A. Context Load Evaluation

Let $\overline{T}_i^{\mathrm{on}}$ be the average time where a dedicated bearer used by type-*i* application is active. From equation (7), (9), (10) and (14) we can figure out that:

$$\overline{T}_i^{\mathrm{on}} = \frac{e^{\lambda_i \phi}}{\mu_i} + \frac{e^{\lambda_i \phi} - 1}{\lambda_i} \tag{18}$$

Let A_x be the area served by the LTE/EPC equipment x. Let n be the total number of applications running in a UE and m the number of applications using the default bearer. Therefore, the Context Load for an LTE/EPC equipment x can be computed as follows:

$$D_x = A_x \rho \left[\beta \overline{T}_{\mathrm{connected}} + \sum_{i=m+1}^{n} \Omega_i \overline{T}_i^{\mathrm{on}} \right] \tag{19}$$

For the PGW and MME A_x=$A_{\mathrm{c}}C_{\mathrm{e}}$, for the S-GW A_x=$A_{\mathrm{c}}C_{\mathrm{ta}}$ and for the eNB A_x=A_{c}.

B. Processing Load and Memory Access Rate Evaluation

In order to evaluate the *Processing Load* (S), we extend the analysis proposed in [8] by adding procedures related to mobility. This analysis is based on some elements described in [2], which have been enriched introducing our state machine model described above. We take into consideration mechanisms described in section III, whose signaling call flows are detailed in [10]–[12], [18] and are summarized in Table I.

Let M_x^y be the number of incoming signaling messages and I_x^y be the number of context creations, modifications or releases addressed to element x (e.g. MME, S-GW, eNB) triggered by the procedure y. The number of incoming signaling messages and context creations, modifications or releases are summarized in Table I.

Therefore, let S_x^y be the number of incoming messages per unit of time due to y procedure in the element x of LTE/EPC network. Let L_x^y be the number of context creations, modifications or releases per time unit due to y procedure in the x element of LTE/EPC network.

Procedures	Events	eNB	MME	SGW	PGW
Service Request *(sr-net/ue)*	Context Creation	1			
	Context Release				
	Context Modification		1	1	
	Incoming Messages	3/4*	3/4*	1/3*	0
Switch to IDLE state *(ci)*	Context Creation				
	Context Release	1			
	Context Modification		1	1	
	Incoming Messages	2	3	1	0
Dedicated Bearer Activation *(db-net/ue)*	Context Creation	1	1	1	1
	Context Release				
	Context Modification				
	Incoming Messages	3/4*	3/4*	2/3*	1/2*
Dedicated Bearer Deactivation *(db-net/ue)*	Context Creation				
	Context Release	1	1	1	1
	Context Modification				
	Incoming Messages	3/4*	3/4*	2/3*	1/2*
Handover without S-GW relocation *(ho-nsr)*	Context Creation	k			
	Context Release	k			
	Context Modification		k		
	Incoming Messages	7	2	2	0
Handover with S-GW relocation *(ho-sr)*	Context Creation	k			
	Context Release	k			
	Context Modification		k	k	
	Incoming Messages	7	3	3	1
Tracking Area Update *(tau)*	Context Creation				
	Context Release				
	Context Modification	1	1		
	Incoming Messages	1	2	0	0

k: number of active bearers in current time

** If communication is initiated by the UE*

TABLE I: Summary of Relevant LTE/EPC Procedures

1) Service Request procedure: The *Processing Load* due to Service Request procedure for x can be computed as follows:

$$S_x^{\mathrm{sr}} = \beta A_x \rho \left[M_{\mathrm{x}}^{\text{sr-ue}} P_{\mathrm{ue}} + M_x^{\text{sr-net}} (1 - P_{\mathrm{ue}}) + M_x^{\mathrm{ci}} \right] \tag{20}$$

Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{sr}} = \beta A_x \rho \Big(I_x^{\text{sr}} + I_x^{\text{ci}} \Big) \tag{21}$$

2) Dedicated Bearer: when a data transmission through a dedicated bearer is finished, the release procedure is triggered after an inactivity time ϕ. Therefore, the *Processing Load* due to Dedicated Bearer Activation and Deactivation requested by the type-*i* application for x can be computed as follows:

$$S_x^{\text{db}}(i) = \Omega_i A_x \rho \Big[M_x^{\text{db-ue}} \chi_i + M_x^{\text{db-net}} (1 - \chi_i) \Big] \tag{22}$$

Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{db}}(i) = \Omega_i A_x \rho \Big(I_x^{\text{db}} \Big) \tag{23}$$

3) Handover: We use the mobility fluid-flow model described previously. Let N_x^{enb} be the number of eNBs served by an equipment x. The *Processing Load* due to handover events for x is given by:

$$S_x^{\text{h}} = \sigma_{\text{c}} N_x^{\text{enb}} \left[M_x^{\text{h-sr}} (1 - P_{\text{rel}}) + M_x^{\text{h-nsr}} P_{\text{rel}} \right] \tag{24}$$

For the MME and P-GW N_x^{enb}=C_e, for the S-GW N_x^{enb}=C_{ta} and for the eNB N_x^{enb}=1. Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{h}} = \sigma_{\text{c}} N_x^{\text{enb}} \left[I_x^{\text{h-sr}} (1 - P_{\text{rel}}) + I_x^{\text{h-nsr}} P_{\text{rel}} \right] \tag{25}$$

4) Tracking Area Update: We assume a centralized MME architecture which only involves intra-MME TAU. Let N_x^{ta} be the number of TAs served by an equipment x and $\lambda_{\text{ue}} \sqrt{C_{ta}}$ be the crossing rate out of a TA. The *Processing Load* due to TAU events can be approximated by:

$$S_x^{\text{tau}} = N_x^{\text{ta}} M_x^{\text{tau}} \lambda_{\text{ue}} \sqrt{C_{\text{ta}}} \tag{26}$$

Where, in case of the MME and P-GW N_x^{ta}=$C_{\text{e}}/C_{\text{ta}}$, for the S-GW N_x^{ta}=1 and for the eNB N_x^{ta}=$1/C_{\text{ta}}$. Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{tau}} = N_x^{\text{ta}} I_x^{\text{tau}} \lambda_{\text{ue}} \sqrt{C_{\text{ta}}} \tag{27}$$

5) Summary: Finally, let n be the total number of applications running on a UE, m the number of applications using the default bearer and $n-m$ the number of applications supported by a dedicated bearer. From equations (20) to (27) the total *Processing Load* (S) of the element x can be computed as follows:

$$S_x = S_x^{\text{sr}} + S_x^{\text{h}} + S_x^{\text{tau}} + \sum_{i=m+1}^{n} S_x^{\text{db}}(i)$$

And the total *Memory Access Rate* (L) of the element x can be computed as follows:

$$L_x = L_x^{\text{sr}} + L_x^{\text{h}} + L_x^{\text{tau}} + \sum_{i=m+1}^{n} L_x^{\text{db}}(i)$$

VI. Numerical Results and Analysis

In this section, we present numerical results and a performance evaluation of different scenarios. The proposed scenarios are based on real statistics from an high user density area in Paris region and its suburbs which are presented in Table II.

Parameter	Value
Area Size (A_t)	$1300km^2$
User Density (ρ)	2300 UEs/km^2
Mean User Speed (V) [19]	$5km/h$
Total of eNBs in the region (C_e)	2800
Total of eNBs in TAL	300 eNBs
Overlapping Factor 20% (γ)	1.2

TABLE II: Scenario Parameters

Assuming uniform circular cells with an overlapping factor γ, the required cell radius, r, to cover the entire area is $r = \gamma\sqrt{A_t/(C_{\text{e}}\pi)}$ km. Based on Orange statistics we propose four main application types: voice, media streaming (i.e. YouTube, Dailymotion, Deezer), social networks (i.e. Facebook, Tweeter) and Background with their associated busy-hour parameters, which are detailed in Table III.

Application	Session arrival rate per hour (α_n)	Session duration (s) (μ_n^{-1})	χ_i
(A) Voice	0.67	180	0.5
(B) Streaming	5	180	1
(C) Social Network	20	30	0.5
(D) Background	40	10	0.8

TABLE III: Traffic Parameters

Table IV shows the five analysed scenarios; the first one is a "Best Effort" deployment, which is the most frequent case currently. The second one assumes a Voice over LTE (VoLTE) offer, which is currently being deployed by some operators (we assume that a unique PDN is used by all services including VoLTE). The last three scenarios represent multi-level QoS deployments. In the third scenario, 10% of streaming traffic is supported by dedicated bearers in addition to VoLTE. In the fourth scenario, 5% of Social Network traffic is supported by dedicated bearers in addition to VoLTE and 10% of Streaming traffic. In the fifth scenario, all streaming and VoLTE traffic is supported by dedicated bearers (worst case).

Scenarios	App. Using Default Bearer	App. Using a Dedicated Bearer
BE	A + B + C + D	-
QoS_1	B + C + D	A
QoS_2	90% B + C + D	A , 10% B
QoS_3	90% B + 95% C + D	A , 10% B , 5% C
QoS_4	C + D	A , B

TABLE IV: Scenarios

The bearer-inactivity timer is usually around 20 seconds [20]. VoLTE dedicated bearers are deactivated via the IP Multimedia Subsystem (IMS) signaling procedure described in [21], [22]. This means that the bearer-inactivity timer for VoLTE dedicated bearer is 0.

We also vary the inactivity timer τ from 20 to 100 seconds. A common value used by mobile operators for dense areas [23] is an inactivity timer (τ) equal to 40 seconds.

For the sake of clarity, only relative values are presented hereafther. Fig. 4, 5 and 6 show the impact of QoS depending on the inactivity timer (τ) on MME, S-GW and eNB. Fig. 4

shows the *Context Load* values normalized using the min/max values of the "Best Effort" scenario (BE). Fig. 5 and 6 show the percentage increase in *Processing Load* and *Memory Access Rate* respectively, relatively to the BE scenario.

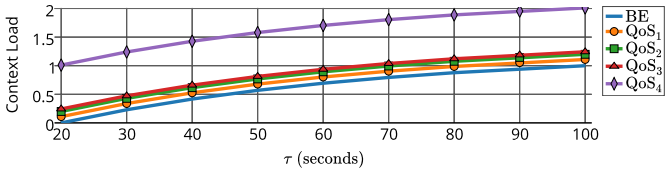

Fig. 4: Normalized Context Load (eNB, S-GW, MME and P-GW)

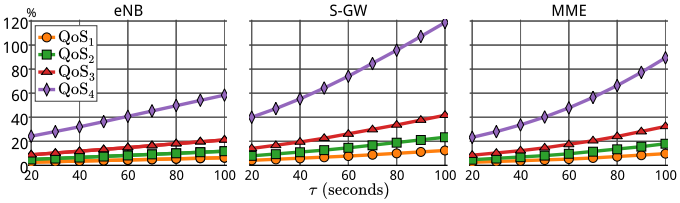

Fig. 5: Signaling Load compared to BE scenario

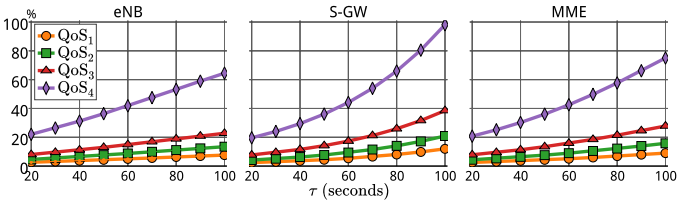

Fig. 6: Memory Access Rate compared to BE scenario

The normalized *Context Loads* increase with τ, as the number of average active contexts. On the contrary, in [8] we showed that the *Processing Load* decreases exponentially with τ reflecting the decreasing number of transitions from CONNECTED to IDLE states. The total *Processing Load* an *Memory Access Rate* are obviously increased compared to the BE scenario due to QoS deployment.

Simulations show that the increase of *Processing Load* and *Memory Access Rate* due to multi-bearer deployment for QoS management is relatively moderate, less than 60% in usual configuration (τ=40s). However, it is more perceptible in rather centralized equipments such as the MME and the S-GW. Nevertheless, scenario QoS_4 shows a major impact in *Context Load*, which is increased by around 200%.

VII. Conclusions

In this paper, we have presented a novel analytical model to evaluate the impact of the standard LTE/EPC QoS model in terms of *Context Load*, *Processing Load* and *Memory Access Rate*. The deployment of the VoLTE/ViLTE and other premium services using dedicated bearers could have a significant impact on the performances of LTE/EPC nodes as shown above. In order to avoid this, a proper network dimensioning in terms of equipments processing/memory capacity and appropriate engineering rules (i.e. τ value) are therefore essential. It is also important to take into account the traffic behaviour (ON-OFF cycles) of premium services, since it could be detrimental to the performances of LTE/EPC elements.

Another alternative to avoid the negative impact of current QoS model is the so-called "IP-centric" approach [24], [25], which has already drawn interest amongst some major actors of the mobile industry. In this model, the QoS is managed at IP level and the dedicated bearers are not necessary.

References

[1] Ericsson. (2015, June) White paper: "Mobility Report".

[2] I. Widjaja, P. Bosch, and H. La Roche, "Comparison of MME signaling loads for long-term-evolution architectures," in *IEEE VTC Fall*, 2009.

[3] C.-K. Han, H.-K. Choi, J. W. Baek, and H. W. Lee, "Evaluation of authentication signaling loads in 3GPP LTE/SAE networks," in *IEEE LCN*, 2009.

[4] C.-K. Han and H.-K. Choi, "Security analysis of handover key management in 4G LTE/SAE networks," *Mobile Computing, IEEE Transactions on*, vol. 13, no. 2, 2014.

[5] M. Wang, M. Georgiades, and R. Tafazolli, "Signalling Cost evaluation of mobility management schemes for different core network architectural arrangements in 3GPP LTE/SAE," in *IEEE VTC Spring*, 2008.

[6] D. S. Tonesi, L. Salgarelli, Y. Sun, and T. F. La Porta, "Evaluation of Signaling Loads in 3GPP networks," *IEEE Wireless Communications*, vol. 15, 2008.

[7] I. Sato, A. Bouabdallah, and X. Lagrange, "Improving LTE/EPC signaling for sporadic data with a control-plane based transmission procedure," in *IEEE WPMC*, 2011.

[8] W. Diego, I. Hamchaoui, and X. Lagrange, "The Cost of QoS in LTE/EPC Mobile Networks Evaluation of Processing Load," in *IEEE VTC-Fall*, 2015.

[9] 3GPP, "QoS Concept and Architecture," TS 23.107 version 8.2.0 Release 8, 2011.

[10] ——, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS)," TS 24.301 version 10.15.0, 2014.

[11] ——, "GPRS enhancements for E-UTRAN access," TS 23.401 version 8.18.0 Release 8, 2013.

[12] ——, "User Equipment (UE) procedures in idle mode," TS 36.304 version 8.10.0 Release 8, 2011.

[13] J. Rodriguez, *Fundamentals of 5G Mobile Networks*. John Wiley & Sons, 2015.

[14] 3GPP, "E-UTRA and E-UTRAN Overall description," TS 36.300 version 8.12.0 Release 8, 2010.

[15] G. P. Deivasigamani, "Dynamic configuration of inactivity timeouts for data radio bearers," 2013, US Patent App. 14/070,827.

[16] R. Thomas, H. Gilbert, and G. Mazziotto, "Influence of the movement of the mobile station on the performance of a radio cellular network," in *Proc. 3rd Nordic Seminar*, 1988, pp. 9–4.

[17] G. Morales-Andres and M. Villen-Altamirano, "An approach to modelling subscriber mobility in cellular radio networks," in *Proceedings of the Forum Telecom*, 1987, pp. 185–189.

[18] 3GPP, "E-UTRAN; X2 Application Protocol (X2AP)," TS 23.107 version 8.9.0 Release 8, 2011.

[19] M. H. Bornstein and H. G. Bornstein, "The pace of life," 19 Feb. 1976.

[20] R. Bassil, A. Chehab, I. Elhajj, and A. Kayssi, "Signaling oriented denial of service on LTE networks," in *ACM MSWiM*, 2012.

[21] 3GPP, "IP Multimedia Subsystem (IMS)," TS 23.228 version 10.9.0 Release 10, 2015.

[22] GSMA PRD N2020.01, "VoLTE Service Description and Implementation Guidelines", version 1.0, December 2014.

[23] 3GPP, "LTE Radio Access Network (RAN) enhancements for diverse data applications," TR 36.822 version 11.0.0, 2012.

[24] I. Hamchaoui, W. Diego, and S. Jobert, "IP centric QoS model for mobile networks - Packet based QoS management for Intra-bearer arrangements," *In IEEE WCNC 2014*.

[25] P. Szilgyi and C. Vulkn, "Application Aware Mechanisms in HSPA Systems," *In IARIA ICWMC 2012*.