Towards Elastic Application-oriented Bearer Management for enhancing QoE in LTE Networks

Tarik Taleb[ϕ], Konstantinos Samdanis[∂], Adlen Ksentini[*]
[ϕ]Aalto University, Espoo 02150, Findland
[∂] NEC Europe, Heidelberg 69115, Germany
[*]University of Rennes 1, Rennes 35065, France
tarik.taleb@aalto.fi, samdanis@neclab.eu, adlen.ksentini@irisa.fr

Abstract— **This paper introduces the concept of elastic bearer in Evolved Packet System (EPS), which allows, on one hand, the users to enhance on-demand the performance of certain applications and on the other hand, it permits the network to efficiently manage the resource allocation considering the application type. In particular, the paper introduces a set of mechanisms to trigger and support bearer elasticity in EPS based on Quality of Experience (QoE) perceived by users or based on feedback from Radio Access Network (RAN). Bearer elasticity can be attained through potential Packet Data Network/Serving Gateway (PDN/S-GW) relocation to eventually improve QoE within and beyond the mobile network operator premises. The paper also introduces a set of methods to identify and cope with a "storm" of requests for particular applications at densely populated areas.**

I. INTRODUCTION

As the mobile industry continues to advance at a high pace with additional 1 billion subscribers being expected by 2020, a thriving range of new and diverse applications that takes advantage of the higher speeds and enhanced network capacity creates an economic pressure on mobile network operators [1]. A simple and direct reaction from mobile network operators is to enhance the network infrastructure, but this is a costly process that may prove not to be enough considering the rapid evolution of data-intensive applications. Hence, other means that consider service differentiation, taking into account the application type, are desired, ensuring the awareness of mobile operators of perceived QoE.

This paper concentrates on application-aware network resource management, assuring for users their respective QoE desired in the 3GPP Evolved Packet Core (EPC). It introduces the concept of elastic bearer, which involves a set of procedures that augment and shrink network resources in a flexible manner, considering the impact on users' QoE. Indeed, the use of elastic bearers can enhance the efficiency of network resource utilization. Such concept of elastic bearers could be of vital importance for the forthcoming 5G systems considering the paradigm of network softwarization and slicing per verticals, wherein the deployment of EPC and mobile services can leverage the benefits of Network Function Virtualization (NFV) allowing a flexible and cost effective network arrangement based-on customer and business demands [2]. For mobile network operators, the adoption of elastic bearers can be also a way for increasing their revenue upon enhancing/assuring the QoE of selected applications, e.g. for specific Over-The-Top (OTT) providers or for particular users. Hence, the elastic bearer concept is an important enabler for mobile network operators to fine tune the QoE provisioning defining in this way new and fair business models with the OTT application providers. Effectively, this can bridge the gap between the increasing data traffic volumes and the decreasing average revenue per connection.

A practical realization of the elastic bearer is through the concept of turbo-boost, which enhances the QoE for end users upon request, with additional fees on top of the typical payment plan. Alternatively, the use of elastic bearers can enable advanced business schemes offering application-oriented services by allocating dedicated core network functions and network resources at particular locations that assure QoE. The proposed elastic bearer is fundamentally different from the Long Term Evolution (LTE) bearer modification [3]. The notion of elasticity is not simply changing the type of bearer, i.e. from default to dedicated, but can additionally guarantee delay, loss, and jitter by adjusting the routing and the selection of core network functions, i.e. performing PDN/S-GW relocation, taking into account the application type and the associated QoE.

In this paper, bearer elasticity can be explicitly requested by the end user for enhancing QoE of a particular application or can be triggered by the network based-on timers or particular events, e.g. congestion, considering also the application type. The main contributions of this paper are three fold: (i) to elaborate the new concept of elastic bearer, (ii) to introduce mechanisms to trigger and manage bearer elasticity and (iii) to handle storms of elastic service requests at particular areas. The main challenge consists in estimating the number of users that may request such elastic service, the associated overall traffic, the location and duration for the provision of an on-demand QoS enhancement or downgrading.

The remainder of this paper is organized as follows. Section II presents the related work. Section III describes the elastic bearer management mechanism focusing on the core network and RAN, coping also with elastic bearer "storms". Section IV provides an analytical model of the QoE-aware bearer elasticity concept and discusses the results. Finally, the paper concludes in Section V.

II. RELATED WORK

Service elasticity was originally introduced for Digital Subscriber Line (DSL) networks as turbo-boost: a DSL user triggers a mechanism that instantly increases the offered capacity. Such an on-demand service enhancement has been

also considered in the context of mobile networks as a feature to improve QoE of specific users [4]. The notion of service elasticity has gained momentum with the introduction of 5G systems and mobile cloud networking, whereby virtual network functions are instantiated/relocated on-demand providing the opportunity for users to trigger in a flexible way potential service enhancements based-on the perceived QoE [2]. Users may issue a request to enhance QoE of a particular service via a dedicated Application Program Interface (API) or as a direct indication towards the application provider. Alternatively, the network can provide such a service autonomously on behalf of the user, in response to data traffic variations including congestion events.

In 3GPP LTE, the adopted QoS framework is network-con rolled. Furthermore, the bearer modification procedure that is initiated by the user is equally controlled by the network. In addition, LTE enables application-based signaling as an alternative means to perform a bearer modification, wherein a User Equipment (UE) signals towards the Application Function (AF) a request for QoS adjustments. AF, in turn, provides the specific QoS parameters to the Policy and Charging Rules Function (PCRF), which establishes the desired bearers for that particular application. In either case, a UE-initiated bearer modification request aims to adjust the capacity and/or QoS for a single traffic flow or change the packet filters related with an active flow without modifying the QoS [6].

Besides the conventional QoS framework, 3GPP has also explored alternative means to enhance QoS provisioning, including the User Plane Congestion Management (UPCON) [7] and data offloading [8]. In particular, UPCON helps the network to become aware of the applications, aiming for efficient traffic management and QoS provisioning. However, UPCON is considered as a complex solution lacking scalability. Data offloading, on the other hand, associates and forwards data traffic flows to either local networks or via offloading points directly to the Internet, allocating different paths for selected services, reducing in this way the traffic load in the operator's scarce EPC resources. A Domain Name System (DNS)-based solution for performing data offloading is presented in [9], while [10] builds-on the top of such an approach providing service optimization and load balancing among the local gateways considering user mobility. Mobile users ensure service continuity for ongoing sessions by maintaining constant the established connectivity, while for new sessions a new gateway can be selected considering load optimization and geographical proximity.

This paper adopts the same philosophy for gateway selection/relocation, but instead of concentrating on network resource optimization, it takes into account the user QoE perspective and application type. In particular, it allows the UE to notify the Mobility Management Entity (MME) with the application type and that is in the bearer modification signaling message, which is then used to trigger an advanced bearer modification including PDN/S-GW relocation based on user QoE, and/or policy settings. In addition, it allows the network to adjust a UE bearer within the core and RAN based-on e.g. congestion indications. Finally, our proposed schemes can also cope with storms of elastic bearer requests at particular areas assuring prioritization.

III. Elastic Bearer Management Mechanisms

This section introduces the bearer management mechanisms that support service elasticity following the current specifications of the 3GPP network architecture considering core network-based or RAN-triggered solutions as well as mechanisms to cope with "storms" of elastic bearer requests.

A. Network-based Bearer Elasticity

This subsection introduces three core network-based mechanisms that support bearer elasticity in EPC. In the first solution, a UE triggers the elasticity via an explicit signaling to the core network specifying the application type and the desired QoS. In the second solution, S-GW relocation and RAN handovers may be enforced by the core network if deemed adequate for improving a user's QoE. The third solution may enforce PDN-GW relocation to control QoE towards specified content providers outside the operator's network domain considering an end-to-end perspective.

As stated earlier, the conventional bearer modification procedure is not enough to capture certain parameters such as service duration, desired delay, and desired bandwidth. These parameters can be explicitly communicated to the MME using unused fields in specific signaling messages such as the UE-initiated bearer resource modification request message [3]. On the user side, the UE needs to be aware of the QoE of an ongoing session and be notified by the network for potential alternations as the user moves. This allows the UE to make the best use of the network resources and enjoy his privileges. The proposed mechanisms capture two distinct approaches that can help accomplish bearer elasticity focusing on the:

- UE, with respect to specific applications considering measures of critical QoS/QoE parameters, e.g. packet inter-arrival time, delay, packet loss, etc [12][13].
- network, i.e. MME assisted, whereby the network notifies the user of imminent QoS degradation based on the current/forecasted traffic, e.g., upon a handover to avoid a potential congestion experience.

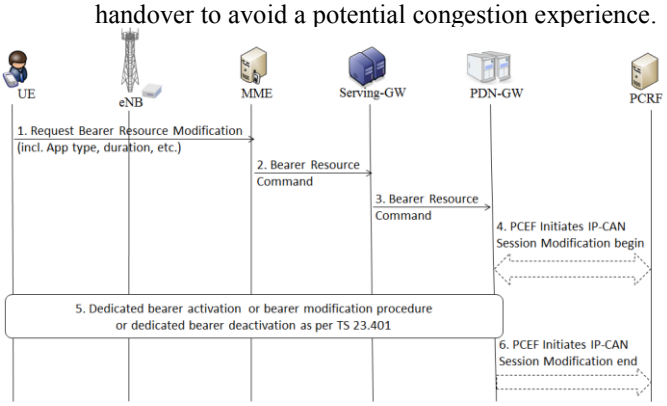

Fig. 1: UE requested bearer resource modification [6].

For the sake of improving its QoE, a UE may issue a bearer resource modification request according to [6] as illustrated in Fig. 1. In addition to the information elements standardized so far (e.g., Linked Bearer Identity, Procedure Transaction Identity, EPS Bearer Identity, QoS, Traffic Aggregate Descriptor, and Protocol Configuration Options), in the

proposed mechanism, the UE also specifies information about the application type, its duration, the corresponding application server and target performance parameters (e.g., delay, throughput, goodput, packet drops, and jitter). Once a user issues a QoE enhancement request, the associated MME then assesses the resource availability within the mobile network and the congestion levels towards the content provider. In particular, when a MME receives a QoE enhancement request, related to an on-going session, it assesses the requested resources and compares them against the ones already allocated. If the QoE enhancement request enquires resources higher than the allocated ones, the network tries to accommodate the request, provided that sufficient resources are available by e.g., triggering PDN/S-GW relocation or even enforcing a RAN handover. Otherwise, the request is rejected or the user is requested to downgrade the desired QoE. If the QoE enhancement request enquires resources smaller or equal to the ones currently allocated, the network then deems the application is inefficiently utilizing the allocated resources. The reasons behind this could be:

(i) the bearer from the PDN-GW to the UE has enough resources, but the link from the PDN-GW to the corresponding server associated with the application is congested.

(ii) the user adopted a new application with a lower QoE demand utilizing a bearer a priori allocated for a different application with a higher QoE.

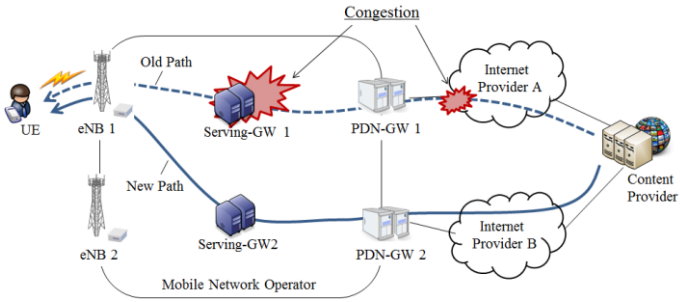

Fig.2: Overview of the potential pain points within the EPC.

Fig.2 illustrates an example where a user experiences service degradation due to congestion at S-GW and/or along the path from PDN-GW to the corresponding server. The figure also shows how the proposed bearer elasticity scheme copes with the service degradation by performing PDN/S-GW relocation. Typically, a MME has knowledge on the geographical locations and loads of S-GWs, which is used for gateway selection and relocation [6]. However, current specifications do not allow a MME to identify congestion between a PDN-GW and a content provider (i.e., as the fixed network domain is generally outside the visibility of the mobile operator), nor about the impact of the current load of an S-GW on a particular application received by a UE. One of the key contributions of this paper is that a request for service elasticity could be an indicator of such. In particular, if a UE places a QoE enhancement request for an application despite the fact that the requested resources are already allocated, the network can infer this as an indication of congestion over the path linking the current PDN-GW and the corresponding application server. As a solution, the network triggers PDN-GW relocation selecting another PDN-GW that provides a better path towards the corresponding server.

For performing S-GW relocation, MME can be configured with a table or a tool that can assess the impact of the load of a S-GW on particular application types. If such load impacts the QoE, the MME selects another S-GW with optimal geographical proximity and with sufficiently low load not to impact the QoE of the corresponding application. It shall be noted that load is used as an example metric and can be replaced by other metric or a set of metrics that reflect the state of the S-GW with respect to a particular application, e.g. a combination of CPU usage, delay and load. The MME can be notified of the S-GW state in near real-time, in the form of actual resource measurements or range of resource usage (e.g., 10% to 20% of load could define a single state, whereas 20% to 50% of load defines another state of the S-GW) [14]. The aforementioned MME table can be then accordingly configured.

In current specifications, MME selects PDN-GWs based on load and geographical proximity. As an additional constraint, MME can also consider the end-to-end QoS by introducing a special weight metric that represents the congestion level of each path from selected PDN-GWs towards the desired content provider. Acquiring QoS parameters on-demand can be achieved by examining the performance of ongoing applications offered by particular content providers using in-band mechanisms, e.g. Congestion Exposure (ConEx) [11], or by injecting special packets, e.g., ping, towards the content provider to test performance parameters such as delay, jitter, and throughput. Alternatively, UE can assess the perceived QoE and provide specific performance information to the network on-demand or upon experiencing a performance degradation using the bearer modification request or a dedicated message [12]. The PDN-GW that is deemed optimal is selected and the corresponding UE is instructed to relocate all or a subset of its flows, or only flows pertaining to the particular application. In case none of the selected PDN/S-GWs have adequate resources, the MME may enforce the mobility of some ongoing sessions to other gateways in order to secure the desired QoE at the selected PDN/S-GWs. Flows of applications with no strict QoS requirements can be easily moved, while those with strict QoS constrains should be avoided. Such a process assures flexibility in resource allocation, since the network can control the load associated with certain PDN/S-GWs, taking also into account the end-to-end QoS. The selection of sessions to relocate can follow the steps described below:

- select always sessions with the maximum congestion impact based-on the congestion volume, i.e. loss percentage.
- relocate such session towards the PDN/S-GW with either maximum resource availability or towards the PDN/S-GW dedicated for best effort traffic.

This operation assumes that the application type is known to the MME, in order to trigger the corresponding PDN/S-GW relocations. Effectively, the MME should be able to retrieve

the application type from base stations, i.e. evolved NodeBs (eNBs), or PDN-GWs for particular flows.

B. RAN enabled Bearer Elasticity

Besides core network-based mechanisms where the application type is known to the MME, this paper also considers RAN-triggered bearer elasticity, wherein eNBs are capable to categorize UEs considering the application type. It shall be noted that trends in 3GPP standards head towards carrying out user plane congestion avoidance. Hence, eNBs are expected to retain knowledge of different application types and to accordingly categorize traffic flows. One way to realize this is by marking packets of a particular application type at the PDN-GW as in [7] in order to make the application type known at the eNB. In case more resources are required for a particular service, e.g. for High Definition Video (HDV), following an elastic bearer request from a UE or an Application Function (AF), eNBs can categorize the attached UEs using the application type, e.g., considering QoS sensitivity and the portion of resources. Alternatively, eNBs can trigger an elastic bearer request upon experiencing congestion according to a predefined threshold policy, considering certain performance parameters, i.e. measuring delay, loss, and jitter. To enable bearer elasticity, eNBs then notify the corresponding MME and/or the Access Network Discovery and Selection Function (ANDSF) of the related UEs indicating the associated application type. Based on this information and by consulting the subscriber profile in the Home Subscriber Server (HSS), the MME may create a data offload policy, which is then propagated towards particular eNBs. Alternatively, the MME or ANDSF can instruct the corresponding UEs to freeze selected applications or enforce a handover to a different Radio Access Technology (RAT such as WLAN) provided that a heterogeneous environment exist with cell overlapping, e.g. enforce flow mobility to a different wireless interface or a different Access Point Name (APN) [12]. In this way, the MME or ANDSF creates adequate resource capacity to accommodate specific QoE requests by pausing or shifting a selected set of application flows within or outside the mobile network. Such instructions from MME or ANDSF towards selected UEs or eNBs can be performed using dedicated messages or using available fields in existing messages.

Another approach is to alleviate congestion associated with particular applications by specifying an Average Maximum Bit Rate (AMBR) per application type per UE. The concept of AMBR has been originally used for a single UE, but here, we suggest adopting it per application. For example, a user of an application "X" that contains different flows (e.g., a webpage could embed text, image and video transmitted on multiple flows) shall have an AMBR per APP for that application. The aggregate resources used by the UE for that particular application shall not exceed the value of AMBR-APP. This shall ensure a certain level of control and fairness.

C. Coping with Storms of Elastic Bearer Requests

When the network gets congested in a crowded area (e.g. at a train station during peak-times, traffic jam or open air festival) where a potential number of users connect to the network using diverse applications at the same time, multiple requests for improving QoE may be simultaneously issued, resulting in "storms" of requests. Prioritizing such QoE enhancement requests in these cases assures efficient usage of bearer elasticity. In case MME receives multiple QoE requests from a number of UEs for different applications, the MME can prioritize these requests based on a set of parameters such as the application type, the IP address or any other service identifier, the duration for which the requested QoS shall be granted, the current locations of UEs, and the requested amount of resources. Such a prioritization may also consider the revenues that each application generates for the operator or based on the QoE reported by the UE.

Based on this prioritization and given the available network resources, the network, i.e. MME, can then accept or reject specific QoE requests or can downgrade the resources of other users following the preferences and policies of the mobile operator. In case the total number of QoE requests received for a particular application during a specific time interval exceeds a predefined threshold, the network may use this as an indicator that a "storm" of requests with respect to that particular application may come and shall accordingly arrange for resources in advance. "Storm" of requests can also be predicted based-on traffic forecasting information that considers a-priori traffic patterns and the number of users being located within the vicinity of the affected area. When a "storm" of QoE requests is anticipated in an early stage, the MME can accordingly plan ahead the release and allocation of corresponding resources.

IV. PERFORMANCE EVALUATION

A. Model for PDN-GW Elastic Bear Requests

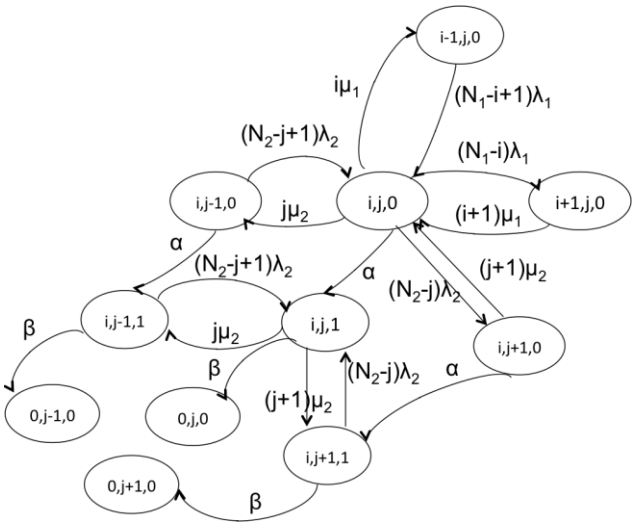

Fig.3: Example of state transition.

In this section, we mainly focus on evaluating the performance of the case where a UE requests an elastic bearer due to QoE degradation. We assume that this degradation is caused mainly by the link between the PDN-GW and the remote servers. Each UE can run two types of applications; application requiring high QoE (type 1) and best effort application (type 2). UEs using the first type trigger an elastic

bearer request if the PDN-GW connection quality degrades, while UEs having the second type of application remain using the same bearer (i.e. same PDN-GW) regardless of the connection quality. The key objective of the envisioned model is to estimate the number of UEs requesting an elastic bearer according the PDN-GW connection quality. We assume that each UE can use either application type 1 or 2. The numbers of UEs using application types 1 and 2 are denoted as N_1 and N_2, respectively. We assume that the arrival rates of UEs using application types 1 and 2 follow a Poisson distribution with rate λ_1 and λ_2, respectively. Furthermore, the service rates of UEs using application types 1 and 2 follow an exponential distribution with rate μ_1 and μ_2, respectively. In the same way, the time duration before a link quality change occurs and the time duration for restoring the link quality are exponentially distributed with respective rates α and β. These assumptions lead us to model the system using a Markov chain $X = \{X_t, t \geq 0\}$ on the state space S defined by $S = \{(i,j,k) \mid i = 0, \ldots, N_1, j = 0, \ldots, N_2$ and $k=0,1\}$. In this model, $X_t = (i,j,k)$ indicates that, at time t, there are *i* active UEs of type 1, *j* active UEs of type 2 connected to a PDN-GW, and the link quality is in state k. While k=0 indicates that the link quality is good, k=1 represents the fact that the link quality between the PDN-GW and the servers is degraded. Fig. 3 illustrates an example (part) of the transitions graph of the envisioned system. The different transitions are as follows:

- If a UE of type 1 arrives, while already i ($0 \leq i \leq N_1-1$) UEs of type 1 as well as j ($0 \leq j \leq N_2$) UEs of type 2 are connected to the PDN-GW, and the link quality is good, then there is a transition from state (i,j,0) to state (i+1,j, 0) with rate $(N_1 - i)\lambda_1$.
- If a UE of type 2 arrives, while already i ($0 \leq i \leq N_1$) UEs of type 1 as well as j ($0 \leq j \leq N_1-1$) UEs of type 2 are connected to the PDN-GW, and regardless the link quality, there is a transition from state (i,j,k) to state (i,j+1, k) with rate $(N_2 - j)\lambda_2$. The difference with the precedent transition consists of the fact that UE of type 1 will ask for an elastic bearer, while UE of type 2 remains connected to the PDN-GW.
- If the service of a UE of type 1 ends while already i ($1 \leq i \leq N_1$) UEs of type 1 as well as j ($0 \leq j \leq N_2$) UEs of type 2 are connected to the PDN-GW, and the link quality is good, then there is a transition from state (i,j,0) to state (i−1,j,0) with rate $i\mu_1$.
- If the service of a UE of type 2 ends while already i ($0 \leq i \leq N_1$) UEs of type 1 as well as j ($1 \leq j \leq N_2$) UEs of type 2 are connected to the PDN-GW, and regardless the link quality, there is a transition from state (i,j,k) to state (i,j-1,k) with rate $i\mu_2$.
- If the link quality degrades while i ($0 \leq i \leq N_1$) UEs of type 1 as well as j ($0 \leq j \leq N_2$) UEs of type 2 are connected to the PDN-GW, then there is a transition from state (i,j,0) to (i,j,1) with rate α.
- If the links quality improves while j ($0 \leq j \leq N_2$) UEs of type 2 are connected to the PDN-GW, then there is a transition from state (i,j,1) to (0,j,0) with rate β. Here, the system should come to a state where i=0 as all UEs of type 1 have been relocated to another PDN-GW or rejected if no resources are available.

The Markov chain X being irreducible and with finite state space, it has a limiting distribution that we denote by π. Due to the lack of space, we will not develop the balance equations. To solve this chain, we used matrix computation technique. The key objective of the model is to obtain expected number of UEs of type 1 that are requesting elastic bearers in steady state. We denote by E[Req] the expected number of requests, which is obtained as follows.

$$E[Req] = \sum_{i=1}^{N_1} \sum_{j=0}^{N_2} i\, \pi_{i,j,1}$$

B. Results Analysis

For all the following results, we considered that the maximum number of sessions supported by a PDN-GW is 2000 active bearers regardless the type of application. We denote by $\rho = \lambda_1/\mu_1$ and $\tau = \alpha/\beta$ the traffic intensity of UE of type 1 and the proportion of time where the link is good, respectively. Note that the higher values of ρ are, the higher the time that UEs spend connected to the PDN-GW is. Then, we fix the number of UEs of type 2 and vary the number of UEs of type 1 from 100 to 2000, for different values of ρ and τ. It shall be reiterated that the purpose of this model is to focus on the number of UEs (type 1) requesting an elastic bearer. For this reason, no results are shown on the performance of the second type of application.

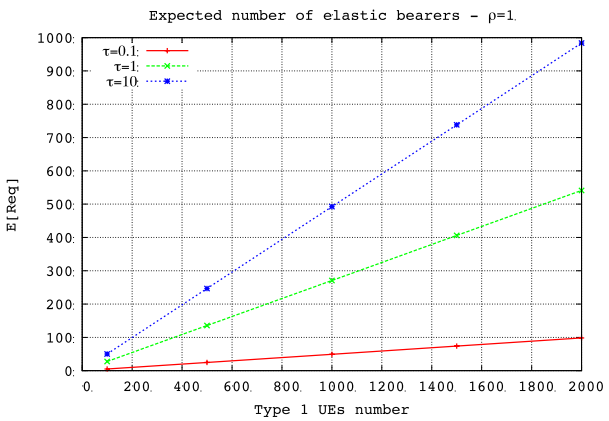

Fig.4: The expected number of requests – case ρ=1

Figs. 4 and 5 represent the expected number of requests for different numbers of UEs of type 1 (N_1) and for different link conditions. While Fig.4 represents the scenario of medium traffic intensity (ρ=1), Fig.5 illustrates the scenario of high traffic intensity (ρ=10). As it is expected, the channel condition highly impacts the number of UEs requesting an elastic bearer. Bad channel condition (τ=10) results in high number of requests, irrespective of the traffic load. Furthermore, we remark that the traffic load also impacts the number of requests. The higher the load, the higher the number of requests. We argue this by the fact that UEs, in this scenario, stay connected to the PDN-GW for a longer duration; hence increasing the probability to experience a degradation of channel quality. But, definitely the parameter with the higher impact in this system is the link quality. Indeed, we remark that when the link condition is good (τ=0.1), the number of requests for elastic bearer is very low, roughly 200 and 100 requests are

received, for N_1=2000 and traffic load is 10 and 1, respectively. Meanwhile, it merely reaches zero when N_1=100, irrespective of the traffic intensity.

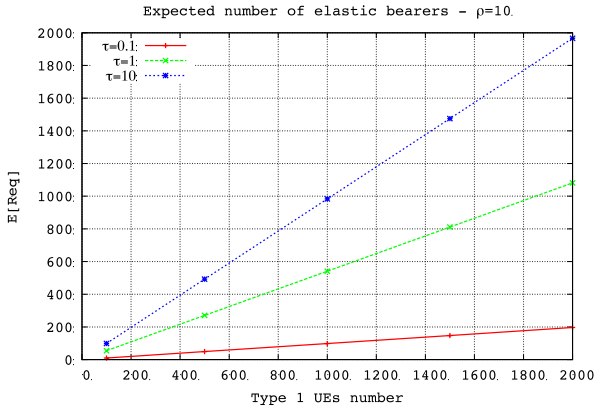

Fig.5: The expected number of requests – case ρ=1

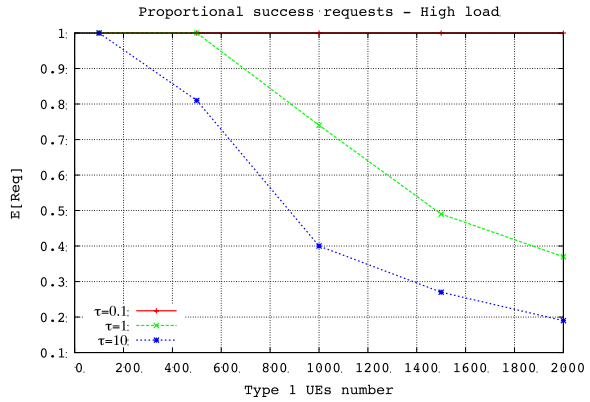

Fig.6: Proportion of successful relocations.

In Fig. 6 we show the proportion of successful relocations for high load scenario; i.e. most other PDN-GWs are highly loaded (i.e., they have the capacity to accept only 400 new UEs) and the traffic intensity is high (ρ=10). We obtained the shown values by using the following formula:

$$Success = \begin{cases} \frac{C_{pdngw}}{E[Req]} & if\ E[Req] > C_{pdngw} \\ 1 & else \end{cases}$$

Here, we considered that the MME has information on the target PDN-GW's current load. We remark that a proportion of successful relocations depends on the number of requests (i.e. link condition). The lower the quality of the link is, the higher is the probability to be rejected, and hence loosing connection. We observe that the worst case (only 20% of requests is accepted) is experienced when the link quality is bad (τ=10) and N1=2000. We explain this by the fact that the number of requests highly exceeds the capacity of the targeted PDN-GW. As mentioned before, one solution could be through prioritizing the requests. It should be noted that except the case of high load, results obtained for all other configurations (i.e. ρ=1 and highly loaded PDN-GW) indicate that 100% of the relocation requests are accepted.

V. CONCLUSIONS

This paper introduced and analyzed a number of different bearer elasticity mechanisms for improving the resource utilization and the users' QoE. Bearer elasticity may provide a means for monetizing QoE assurance allowing mobile operators to bridge the gap between the increasing data volumes and decreasing revenue per connection. The logic of providing an elastic bearer goes beyond simple load balancing, because an end-to-end approach is considered and the purpose is not to balance resources but to also manage efficiently the resource allocation considering the application type. Hence, instead of distributing the load among PDN/S-GWs, our proposed schemes provide a selection process that considers the QoE provision associated with particular applications.

ACKNOWLEDGEMENT

This research work is partially supported by the TAKE 5 project funded by the Finnish Funding Agency for Technology and Innovation (TEKES), a part of the Finnish Ministry of Employment and the Economy.

REFERENCES

[1] GSMA, The Mobile Economy, 2015.

[2] T. Taleb, "Towards Carrier Cloud: Potential, Challenges and Solutions", IEEE Wireless Communications, Vol.21, No.3, Jun 2014.

[3] 3GPP TS 24.301, Technical Specification Group Core Network and Terminals, Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS), stage 3, Rel.13, Jun 2015.

[4] K. Samdanis, F.G. Mir, D. Kutscher, T. Taleb, "Service Boost: Towards on-demand QoS Enhancements for OTT Apps in LTE", IEEE ICNP, Gottingen, Oct. 2013.

[5] A. Bradai, K. Singh, T. Ahmed, T. Rasheed, "Cellular Software Defined Networking: A Framework", IEEE ComMag, Communications Standards Supplement, Vo.53, No.6, Jun. 2015.

[6] 3GPP TS 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, Rel. 13, Jun 2015.

[7] 3GPP TR 22.805, Feasibility Study on User Plane Congestion Management, Rel 12, Jun 2012

[8] K. Samdanis, T. Taleb, S. Schmid, "Traffic Offload Enhacements for eUTRAN", IEEE communication Surveys and Tutorials, Vol.14, No.3, 3rd Quarter 2012.

[9] T. Taleb, K. Samdanis, S. Schmid, "DNS-based Solution for Operaotr control of Selected IP Traffic Offload", IEEE ICC, Kyoto, Jun 2011.

[10] T. Taleb, Y.H. Aoul, K. Samdanis, "Efficient Solutions for Enhancing Data Traffic Management in 3GPP NEtworks", IEEE System Journal Vol.9, No.2, Jun. 2015.

[11] M. Mathis, B. Briscoe, Congestion Exposure (ConEx) Concepts, Abstract Machanisms and Requirements, Internet-Draft, Oct. 2014.

[12] A. Ksentini, T. Taleb, and K. Benletaief, "QoE-based Flow Admission Control in Small Cell Networks", to appear in IEEE Trans. on Wireless Communications.

[13] T. Taleb and A. Ksentini, "QoS/QoE Predictions-based Admission Control for Femto Communications," in IEEE ICC 2012, Ottawa, Canada, Jun. 2012

[14] F.Z. Yousaf and T. Taleb, "Fine Granular Resource-Aware Virtual Network Function Management for 5G Carrier Cloud," to appear in IEEE Network Magazine .