Geographical Location and Load Based Gateway Selection for Optimal Traffic Offload in Mobile Networks

Tarik Taleb, Yassine Hadjadj-Aoul, Stefan Schmid

Geographical Location and Load based Gateway Selection for Optimal Traffic Offload in Mobile Networks

Tarik Taleb[1], Yassine Hadjadj-Aoul[2], and Stefan Schmid[1]

[1] NEC Europe Ltd, Heidelberg, Germany
[2] University of Rennes-1, Rennes, France

[1] {tarik.taleb, stefan.schmid}@neclab.eu, [2] yhadjadj@irisa.fr

Abstract. To cope with the rapid increase of data traffic in mobile networks, operators are looking for efficient solutions that ensure the scalability of their systems. Decentralization of the network is one of the key solutions. With such solution, small-scale core network nodes (gateways) are locally deployed to serve the local community of users, in a decentralized fashion. In this paper, we devise methods that enable User Equipments (UEs), both in idle and active mode and while being on the move, to always have optimal Packet Data Network (PDN) connections in such decentralized networks. The proposed methods are compared based on their impacts on current 3GPP standards and the benefits of the overall approach are evaluated through simulations. Encouraging results are obtained.

Keywords: SIPTO, 3GPP Network, traffic offload, mobile network.

1 Introduction

Along with the ever-growing community of mobile users and the tremendous increase in the traffic associated with a wide plethora of emerging bandwidth-intensive mobile applications, mobile operators are facing a challenging task to accommodate such huge traffic volumes, beyond the original network capacities[1][2]. The challenge becomes more significant considering the fact that the Average Revenues per Users (ARPU) are getting lower given the trend towards flat rate business models. Operators are thus investigating cost-effective methods for accommodating such traffic with minimal investment to the existing infrastructure.

Decentralizing the mobile network is one of the key solutions. With such solution, small-scale core network nodes, e.g., Packet Data Network Gateways (PDN-GWs), and Serving GWs (S-GWs), are locally deployed to serve the local community of users, in a decentralized fashion[1]. The benefits of such decentralized mobile network are manifold. Indeed, with such decentralized networks, operators will be able to

[1] In this paper, the focus is on the Evolved Packet System (EPS) [4][5] but the general description can be equally applied to the General Packet Radio Service (GPRS). In this case, Serving GPRS Support Node (SGSN) would map on to S-GW and MME, and Gateway GPRS Support Node (GGSN) would map on to P-GW.

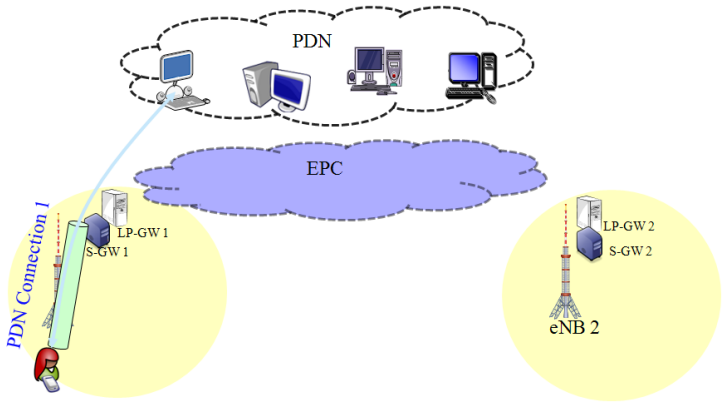
optimize usage of their resources by selectively breaking out IP traffic whenever and as near to the edge of operator's network as possible. This concept is in line with Selected IP Traffic Offload (SIPTO) [3], a topic currently under discussion within 3GPP. It also fits the changing paradigm for traffic routing, inspection and charging by operators (e.g., the trend towards flat rates and differentiation into "dumb", bit-pipe traffic and value-added services). It is also in line with the quest for a network architecture flatter than what has been achieved with the Evolved Packet System (EPS) [4][5]. Effectively, by breaking out selected traffic at entities close to the moving terminal and thus also at the edge of the network (e.g., Radio Network Controller (RNC) in 3G, eNodeB in the EPS), operators will be able to avoid overloading their scarce core network resources (i.e., GGSNs, SGSNs, and P/S-GWs).

The discussions and analysis in the 3GPP standardization group currently focus on the definition of the architecture, i.e., on where the point of local breakout/traffic offload should be placed. Issues regarding security, charging, mobility, traffic control/handling, and optimal gateway selection are yet to be investigated. In this paper, we focus on finding adequate solutions that address the latter issue.

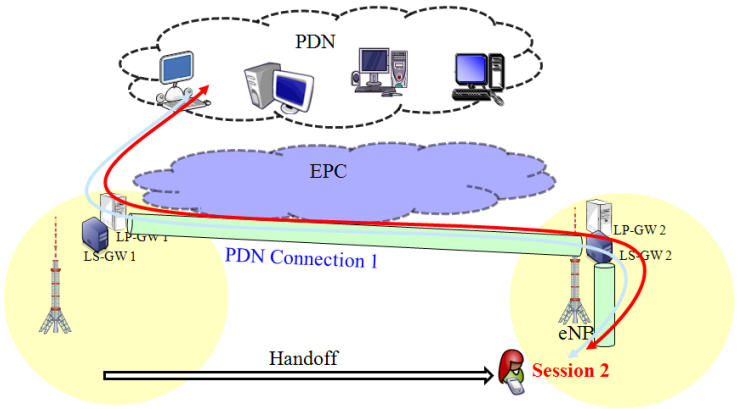
The optimal gateway selection problem is illustrated in Fig. 1. As will be discussed later, the issue can be solved relatively easily in case of UEs being in ECM-Idle (EPS Connection Management) state (i.e. idle mode). However, in this paper, we will also present a solution for optimal gateway reselection for UEs being in ECM-connected state (i.e., active mode). Here, we assume UEs to have the capability to support multiple PDN connections to multiple Access Point Names (APNs). Fig. 1(a) depicts the case of a UE initiating a PDN connection to a particular P-GW via a Long Term Evolution (LTE) radio cell (i.e., eNB1). At a later instant, and after performing handoff to a distant eNB (i.e., eNB2 in Fig. 1(b)), the UE initiates a new IP session to the same APN while the old IP session is still active. Following the current standards [6], the UE cannot have multiple IP sessions via different PDN connections to the same APN over the same access. Therefore, the new IP session (i.e., session 2) will be established via the old P-GW (i.e., P-GW1) as depicted in Fig. 1(b). This is clearly not an optimal decision given the fact that another more optimal P-GW (P-GW 2) is available in the visited area. The optimality of the new PDN gateway can be assessed in terms of both geographical proximity to the UE as well as expected load. A radical solution to this issue could be to set up the new IP sessions via the optimal gateway and to migrate the existing sessions to the new gateway. However, this solution would have significant impact on the user experience due to service disruption. As an alternative, we suggest that the UE would keep the old IP sessions via the old P-GWs but sets up new IP sessions (to the same APN) via the currently most optimal gateway. We propose and compare three methods whereby the setup of new IP sessions via a newly selected gateway is initiated by UE, MME, or P-GW.

Indeed, when a UE, accessing a particular APN using a given PDN connection, performs handoff to a new radio cell, and/or a more optimized P-GW (e.g., in terms of load, geographical proximity relative to UE, etc.) becomes available, the UE shall be able to set up a new PDN connection to the more optimal P-GW when it initiates a new IP session. This paper is about devising a mechanism that enables UEs to establish another optimized PDN connection (e.g., upon a trigger from the network or when judged appropriate by the UE) to the same APN, and a mechanism with which a UE would bind IP flows/connections/sessions (when they are established) to their

corresponding PDN connections. In this way, the UE always remembers which IP flow/connection/session uses which PDN connection. In should be noted that in this paper a session is defined based on the IP address of the peer node, application type, and underlying protocol types, and that a session can be associated with more than one protocol type (e.g., SIP, RTP, and RTCP).



(a) PDN Connection 1 initiated at time T when UE is connected to eNB1.



(b) After performing handoff to eNB2, UE initiates a new session via the old P-GW.

Fig. 1. Limitation of current standards.

The remainder of this paper is organized as follows. Section 2 presents the state of the art. Section 3 presents our proposed gateway selection mechanisms and that is for UEs in both ECM-idle mode and ECM-connected mode, respectively. The implementation issues of the proposed solutions are also discussed. Section 4 evaluates the overall approach and showcases its technical benefits. The paper concludes in Section 5.

2 State of the Art

The Evolved Packet Core (EPC) is designed to encompass different 3GPP accesses (i.e., 2G, 3G and LTE) as well as non-3GPP access (e.g., WiMAX, CDMA2000 ©, 1xRTT, etc.). The richness of EPC accesses gave birth to a new 3GPP entity called ANDSF (Access Network Discovery Selection Function) that assists UEs to find the best or most suitable access out of the many available ones [7]. It has also led to different interesting 3GPP study items whereby UEs are allowed to have simultaneous accesses to different networks using different access technologies. In [8], the 3GPP SA2 group started a study item to investigate different possibilities for dynamic IP flow mobility between 3GPP access and one, and only one, non-3GPP access. The study of solutions to support routing of different PDN connections through different access systems is also in scope. Some of the solutions are being standardized in the technical specifications [9]. In [8], a work item is proposed to allow a UE, equipped with multiple network interfaces, to establish multiple PDN connections to different APNs via different access systems and to selectively transfer PDN connections between the accesses with the restriction that multiple PDN connections to the same APN shall be kept in one access. Whilst there is also another work, impacting different technical specifications, that aims for enabling UEs to establish and disconnect multiple PDN connections to the same APN uniformly across the EPS, a mechanism that indicates to the network or to the UE when it is beneficial to set up a new IP session over a new PDN connection via the same access is overlooked. Indeed, in current solutions, a UE, supporting multiple APNs, may have different PDN connections, each associated with a different APN (e.g., Internet, IMS, WLAN, etc). When a UE is using a PDN connection to access a particular APN, that PDN connection (to the APN in question) does not change until the UE becomes in idle mobility. In other words, P-GW relocation is recommended only during idle mobility, to avoid service disruption, because when the P-GW changes, the old PDN connection is simply torn down. Additionally, as long as the UE has a PDN connection (to access a particular APN) and is in active mode, the UE will always use the same P-GW to set up any new IP sessions to the same APN.

As stated earlier, mobile operators are aiming for the decentralization of their networks. In this context, a mechanism that indicates to the network when it is beneficial for a UE to set up a new IP session via a new PDN connection will be highly required. Its importance becomes further vital knowing the interest of operators in offloading “dump” traffic as locally as possible to achieve the goals of SIPTO [3]. In this paper, we propose a set of mechanisms that enable a UE to know how and when to establish a new optimized PDN connection for launching new IP sessions to a particular APN. This is done without impacting/compromising the on-going (old) PDN connections to the same APN.

3 Local GW Selection

In this paper, two states of UEs are separately discussed: ECM-idle mode and ECM-connected mode. In the latter, we specifically consider the case of UEs supporting

multiple PDN connections. The objective is to trigger UEs to re-establish PDN connections (e.g., those subject to SIPTO) when it is beneficial for both the network and the UEs to reselect another nearby local P-GW, e.g., in case the UE traveled a significant distance from the original P-GW.

3.1 Triggering local GW selection for UEs in idle mode

In the EPS the concept of "always-on" was adopted. This means that once a UE has established a PDN connection, it remains configured in the network even in case the UE goes idle. This feature has the great advantage that the UE can communicate right-away when it becomes active and does not require re-establishment of a PDN connection (incl. allocation of a new IP address) at the time when the user wants to communicate. As stated earlier, the problem with this in the context of SIPTO or decentralized networks in general is the following:

1. Once a UE has established a PDN connection to a well-positioned Local P-GW (in the core or close to the radio station), it maintains this connection – and therefore also the associated "local" GW – until it explicitly disconnects from this PDN.

2. This implies that if the user moves a significant distance away from the initial (H)eNB, the chosen local P-GW providing the access towards the service network (i.e. PDN) may not be optimal anymore.

Fig. 2 illustrates this deficiency for a user that first connects to a service via a macro cell (eNB1) and a "near-by" P-GW in the core network (step 1), and then moves a significant distance away while in idle mode. Without the optimization, the UE remains connected to the P-GW originally chosen, although a local P-GW (LP-GW1) could offer a more optimal access to the service.

To resolve this problem, this paper suggests that the UE simply re-establishes the PDN Connections (e.g., those subject to SIPTO traffic) during idle mode mobility when it has moved away a significant distance from the originally selected "local" P-GW. By simply re-establishing the PDN Connection, the default P-GW selection mechanism would ensure that the UE will again be connected to a local P-GW that is geographically/topologically close to the user's location. As a result, the traffic subject to SIPTO would be again "broken out" or "offloaded" at the most suitable location from the user and operator point of view.

To re-establish a PDN connection, a UE can be either triggered by the network via, for example, a flag during the Tracking Area Update (TAU) procedure [5] or the network could simply disconnect the PDN connection. However, in the latter case, the UE would first have to establish a new PDN connection (incl. IP address allocation, tunnel configuration, etc.) before it can become active. This would break the "always on" concept and also negatively impact the user experience as extra delay is incurred. Nevertheless, since the proposed PDN connection re-establishment procedure is proposed to take place only during idle-mode mobility, no degradation of the service quality or service disruption would be introduced.

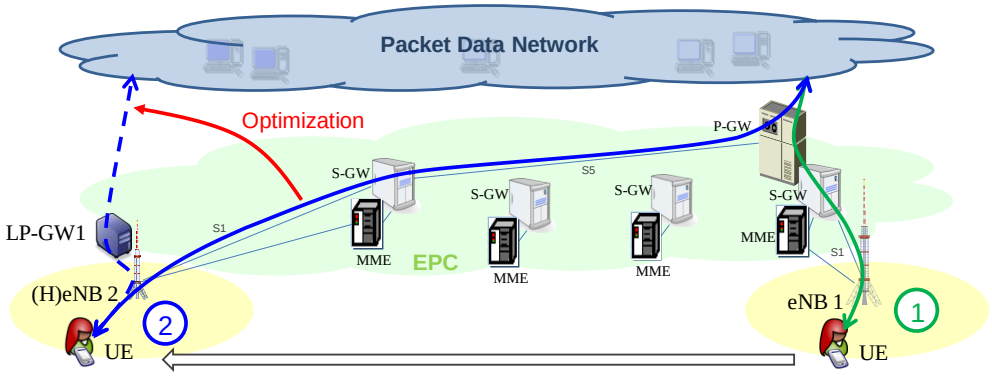


Fig. 2. Expected optimization for UEs in idle mode.

The open issue is to identify when a UE should re-establish a PDN connection. The following solutions are feasible:

- **Option 1** - Periodically – after a configurable time period: The issue with this solution is that stationary UE will introduce a lot of extra signaling overhead without any gain for the operator.
- **Option 2** - Upon Tracking Area Update – whenever the UE changes tracking area: The change of tracking area ensures that the UE has actually moved away from the original location. It is, however, not clear whether the re-establishment would actually lead to a different Local GW.
- **Option 3** - Upon indication from the network – whenever the network considers it beneficial: The network indicates to the UE (e.g., as part of idle-mode mobility procedures or with a special cause for the PDN disconnection) when a particular PDN connection should be re-established. Since the operator has the knowledge of the network topology and the available local P-GWs, it can indicate to the UE when the re-establishment of a PDN connection is worthwhile – i.e., when this leads to a better P-GW selection and thus to a more optimal path.

3.2 Triggering local GW selection to accommodate new IP sessions of UEs in active mode

As discussed earlier, in current standards, as long as a UE has a PDN connection to access a particular APN through a particular access type, the UE will maintain the same P-GW to set up any new IP sessions to the same APN. Motivated by the example of Fig. 1, discussed in the introduction, we argue that when a UE, accessing a particular APN using a given PDN connection, performs handoff to a new base station, and a more optimal P-GW (e.g., in terms of load, geographical/topological proximity relative to UE) becomes available, the UE should be able to set up a new

PDN connection to the more optimized P-GW when the UE initiates new IP sessions to the same APN.

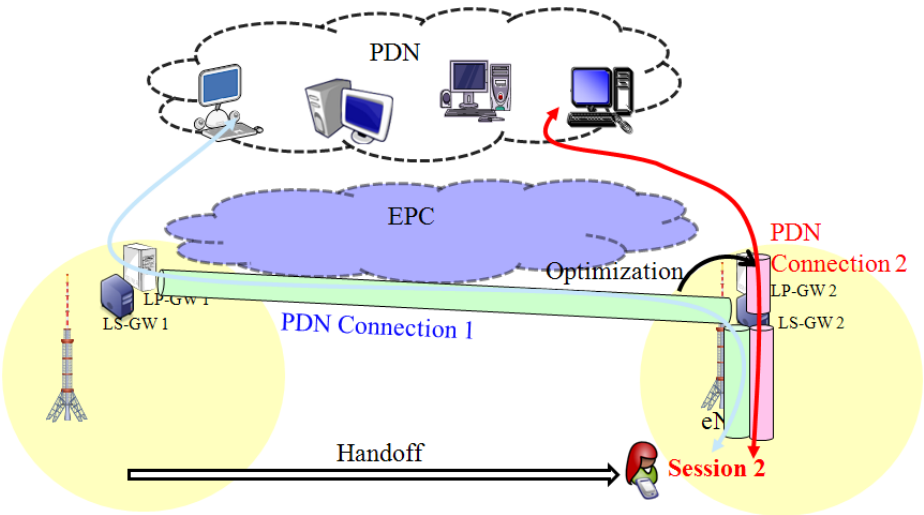


Fig. 3. Setting up new IP sessions via optimal PDN connections while old IP sessions using the old PDN connection are not compromised.

Fig. 3 depicts the envisioned solution. Indeed, after handoff, and when a new, optimal P-GW becomes available, a UE establishes a PDN connection to the new P-GW to establish any new IP session. The on-going sessions keep using the old PDN connection. In the figure, we apply our solution to a scenario whereby the UE performs handoff to an area where another optimal P-GW becomes available. However, the solution can be also applied even if the UE remains in the same area (i.e., cell) and the P-GW it is connected to becomes non-optimal (e.g., because it is currently highly loaded) and another P-GW (e.g., a less loaded one) becomes optimal.

The key question is how to trigger the UEs to establish a new PDN connection for new IP sessions while keeping the old ones on an already available PDN connection. To cope with this issue, we propose the following three methods and qualitatively compare them.

MME-initiated: The MME may apply different mechanisms to check if there are any more optimal P-GWs available (e.g., take into account network and GW load information or UE mobility prediction). When a more optimal P-GW is available, the MME could use existing signaling message during handover to indicate to the UE that for new IP sessions to the APN in question, it should consider establishment of a new PDN connection. This indication can be in the form of a flag, based on which the UE establishes a new optimized PDN connection when it wants to initiate a new IP session, or it can be in the form of an explicit indication of the IP address of the optimal P-GW (e.g., as part of the TAU procedure).

UE-initiated: In this solution, when a UE wants to set up a new IP session to an APN with which it has an ongoing PDN connection, it queries MME (e.g., using NAS

signaling) if it should use the existing PDN connection or consider a new one. Querying MME can be also done based on other events, such as when a UE performs a number of handoffs, after a particular period of time, after the UE moves for a certain distance based on location information, after the UE enters into a new tracking area and/or a specific area during a specific time, etc. Compared to the MME-initiated approach, the UE-initiated solution may generate unnecessary queries to the MME. However, both these solutions require, apart from the modifications for allowing multiple PDN connections to the same APN through the same access type, only very minimal extensions of the standard signaling interfaces, i.e. a new indicator (flag) between UE and MME.

P-GW initiated: For this solution, two alternatives can be envisioned. In the first one, when the current P-GW realizes that the UE is to be better serviced by another P-GW, it simply rejects any requests for any new IP sessions. The rejection can be done via a new error message. This operation intuitively requires that P-GWs have the ability to filter traffic per IP flow/session. It also requires that P-GWs have knowledge on the optimality of a set of other P-GWs (e.g., only neighboring ones).

In the second alternative, when a particular P-GW starts running under specific conditions (e.g., at a load exceeding a certain threshold), it notifies a selected set of UEs (with ongoing connections to the P-GW) to establish new PDN connections with other P-GWs to accommodate new IP sessions. This can be done by designing new and specific signaling messages using S5/8 and S11 interfaces, in addition to NAS signaling from the MME to the UE, by introducing a flag in data packets, or including a flag in existing signaling messages between PGW and UE (e.g., PCO – Protocol Configuration Options [10]). Whilst achieving these requirements is not impossible, they admittedly add some level of complexity to P-GWs and some minor ones to UEs.

Finally, to support the establishment of a new (more optimal) PDN connection for new IP sessions, UEs need the ability to bind new IP sessions (when they are established) to a specific PDN connection or P-GW. A straightforward solution for such binding would be a mapping based on the destination IP address of the peer, application type, and protocol types. In this way, a UE always knows which IP flow/session uses which PDN connection. When all IP flows/sessions associated with a particular PDN connection are finished, the UE can trigger the release of the corresponding PDN connection. For this purpose, a timer based solution can be adopted, which simply tears down a PDN connection when no packets are sent for some time.

4 Performance Evaluation

In this section, we evaluate the proposed solutions and highlight their technical benefits. In the performance evaluation, the focus is on the case of active UEs. The conducted simulations were run for 100s; a duration long enough to ensure that the system has reached its stable state. They are based on the network simulator ns-3 [11] using a network topology as depicted in Fig. 4. We simulate 20 UEs, distributed uniformly around eNB1 over a surface of 2000 x 2000 m2. All simulated nodes are

moving over the coverage areas of eNB1 and eNB2, changing their point of attachment to the network once the signal of the target eNB becomes stronger than that of the source eNB.

At the beginning of the simulation, each UE initiates an ON-OFF application, with ON time set to one second, sending data packets at a rate randomly selected from within the interval [75:150] Kbps, simulating applications ranging from VoIP to video streaming (e.g., YouTube). The packet payload length is set to 256 bytes. Upon moving to a new cell, UEs initiate new IP sessions with the same characteristics as described above.

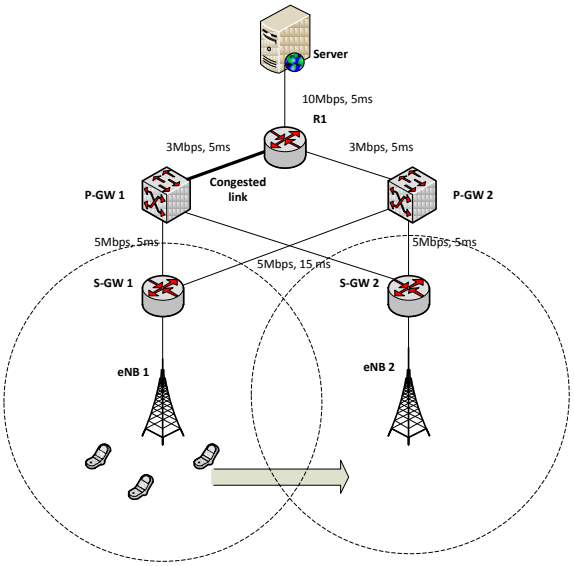


Fig. 4. The considered network topology.

The first metric used to evaluate the efficiency of the proposed solution consists in the transmission buffer length of P-GW1. Fig. 5 demonstrates that the conventional approach, which uses the old P-GW for the new IP sessions, experiences increased buffer lengths, resulting in several buffer overflows and also increasing the flows' latencies as depicted in Fig.7. Indeed, with the conventional approach, UEs do not consider nearby P-GWs when establishing new IP sessions. Instead, they establish their new IP sessions via the old gateway, which increases its load, resulting in buffer overflows and longer latencies.

As a direct consequence of buffer overflows, the aggregated packets loss, shown in Fig. 6, increases significantly which may degrade the quality of the different services. In contrast, by exploiting local P-GWs to accommodate new IP sessions, the proposed approach distributes better the traffic among the available P-GWs. This feature helps in avoiding congestion at the old P-GW, as demonstrated in Figs. 5 and 6. Indeed, the proposed approach exhibits no buffer overflows and almost null packet losses.

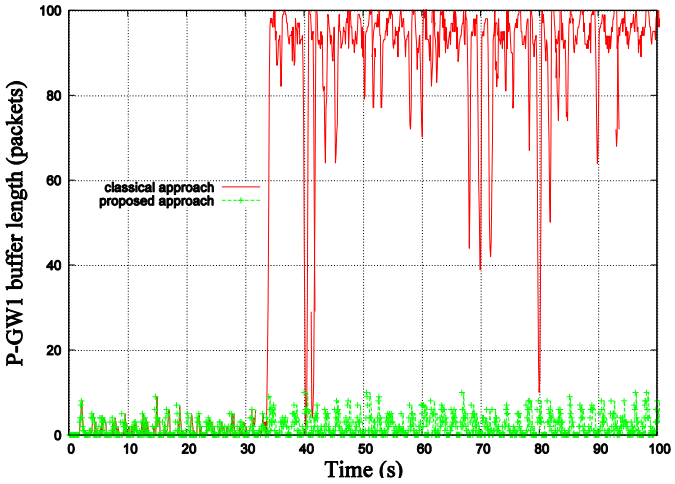


Fig. 5. The transmission buffer length at P-GW1.

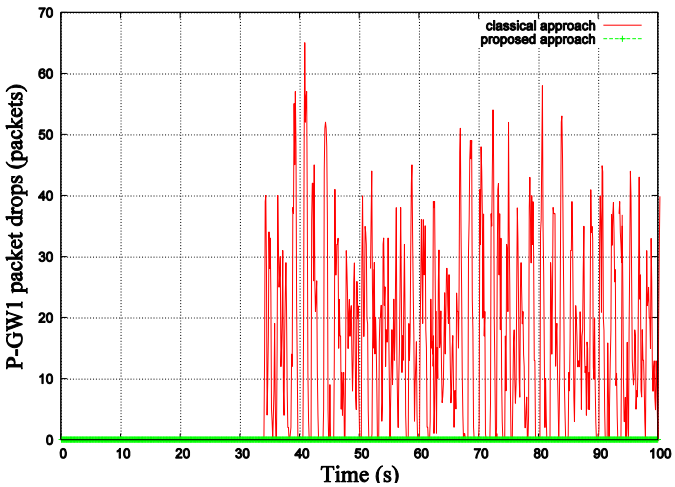


Fig. 6. The packet drops at P-GW1.

Fig. 7 depicts the difference between the average latencies experienced by each application in case of the conventional and the proposed approaches. The delay differences exhibited by the first 20 flows are approximately 30ms and correspond mainly to the buffering delay as the application packets traverse the same path to the server. Indeed, in both approaches, the first 20 flows are not impacted by gateway selection: they are created before the UEs' movement to the other eNB. Flows ranging from 21 to 40 are newly created after the handoff of UEs. They are therefore impacted by gateway selection. The flows 21-40 exhibit higher delay differences, in the order of approximately 50ms. In fact, when using the proposed approach, the system always selects optimal gateways, which significantly decreases the average delay for each flow. It should be recalled that already-established applications continue using the old P-GW.

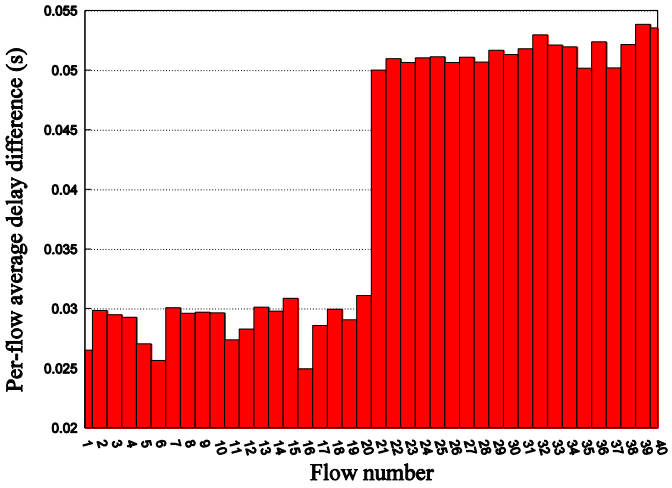


Fig. 7. The difference between the latencies of the classical and the proposed approaches.

Fig. 8 depicts the packet loss experienced by the different flows in case of the conventional approach (i.e., as almost no loss was experienced in case of the proposed approach). The loss is mainly a result of buffer overflows experienced at congested P-GWs. The loss may become more significant along with the increase in the number of mobile terminals, terminals' mobility, and/or the application's data transmission rates.

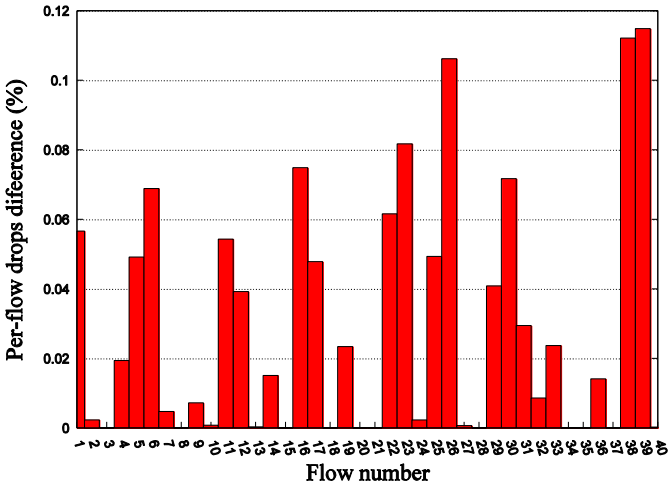


Fig. 8. The difference between the drops of the classical and the proposed approaches.

5 Concluding Remarks

In this paper, we highlighted the need and benefits for always optimizing the path of mobile terminals (UEs) to their corresponding anchor gateways, namely P-GWs in

EPS, in case of decentralized mobile operator networks and/or networks that adopt traffic offload strategies. Adequate methods were devised for UEs in ECM-idle mode and those in ECM-connected mode. In the latter case, we compared three methods that trigger a UE to first establish a new and more optimal PDN connection before creating new IP sessions, without compromising the old IP sessions. Admittedly, the devised methods involve some additional complexity at different core network nodes (depending on the solution). However, the benefits for operators, verified through simulations, clearly justify the required enhancements.

Whilst the main motivation behind this work consists in supporting the decentralization of future mobile operator networks and the envisioned traffic offload strategies, the devised solutions can also assist in energy saving and efficient load balancing. Tailoring our proposed methods to such objectives forms the future directions of our research work in this area.

Acknowledgment

The work described in this paper is partially supported by the national French project ANR VERSO ViPeer.

References

1. "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009-2014," White Paper, Feb. 2010.
2. "MOBILE TRAFFIC GROWTH + COST PRESSURES = NEW SOLUTIONS?," New Mobile, Jan. 2010.
3. "TS Group Services ad System Aspects; Local IP Access and Selected IP Traffic offload (Rel. 10)", 3GPP TR 23.829 V1.1.0. May 2010.
4. 3rd Generation Partnership Project, "Architecture enhancements for non-3GPP accesses Rel 10," 3GPP TS 23.402 V10.0.0, Jun. 2010.
5. 3rd Generation Partnership Project, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," 3GPP TS 23.401 V10.0.0, Jun. 2010.
6. 3rd Generation Partnership Project, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3," 3GPP TS 24.301 V9.3.0, Jun. 2010.
7. 3rd Generation Partnership Project, "Access Network Discovery and Selection Function (ANDSF) Management Object (MO)," 3GPP TS 24.312 V9.1.0, Mar. 2010.
8. 3rd Generation Partnership Project, "Multi Access PDN connectivity and IP flow mobility," 3GPP TR 23.861 V1.3.0, Feb. 2010.
9. 3rd Generation Partnership Project, "IP flow mobility and seamless Wireless Local Area Network (WLAN) offload; Stage 2," 3GPP TS 23.261, Jun. 2010.
10. 3rd Generation Partnership Project, "Mobile radio interface Layer 3 specification; Core network protocols; Stage 3," 3GPP TS 24.008 V9.3.0, Jun. 2010.
11. ns3, "The ns-3 Network Simulator"; http://www.nsnam.org/