Efficient Solutions for Enhancing Data Traffic Management in 3GPP Networks

Tarik Taleb, *Senior Member, IEEE*, Yassine Hadjadj-Aoul, *Member, IEEE*, and Konstantinos Samdanis, *Member, IEEE*

***Abstract*—As mobile data traffic is continuously increasing and the average revenues per user are getting lower due to flat-rate changing, operators are in need of means to cope with higher-than-expected data volumes introducing minimal additional capital expenditures. This paper investigates network decentralization in conjunction with the selective IP traffic offload approaches to handle such increased data traffic. We first devise different approaches based on a per-destination-domain-name basis, which offer operators a fine-grained control to determine whether a new IP connection should be offloaded or accommodated via the core network. Two of our solutions are based on Network Address Translation (NAT) named simple-NATing and twice-NATing, whereas a third one employs simple tunneling, and a fourth adopts multiple Access Point Names. We also propose methods enabling user equipment (UE) devices, both in idle and active modes, and while being on the move, to always have efficient packet data network connections. A qualitative analysis and a simulation study compare the different approaches with respect to cost, complexity, service continuity, and network performance, demonstrating the significance of the proposed scheme for multimedia applications.**

***Index Terms*—Mobile networks, selective IP traffic offload (SIPTO), traffic management, traffic offload.**

I. Introduction

THE EVOLUTION of the mobile network infrastructure toward higher capacity radio, i.e., 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE), brings new horizons for user experience. New applications with increased network-based features are progressively launched, and bandwidth-intensive services, such as high-definition mobile video streaming, rapidly consume the offered capacity, increasing pressure on operators for further investments in their core networks [i.e., evolved packet core (EPC)]. Besides the increase in the average user demand and the ever-growing number of mobile users, the popular business models that introduce flat rates further reduce the average revenue per user. Hence, operators are faced with a dilemma of enhancing their network scalability while minimizing infrastructure investments [1], [2]. Effectively, to meet such an optimal resource utilization objective, operators should be capable of selectively offloading IP traffic as close to the edge of the network as possible, differentiating in this way "dumb," bit-pipe traffic, and value-added services. Such network paradigm is in line with evolving decentralized architectures [20], which aim to fulfill the quest for a network architecture flatter than what has been achieved with the EPC [3], [4].

Three fundamental concepts have been established in 3GPP's System Architecture working group (SA2), as specified in [5] and [6], including the following.

- **Local IP Access (LIPA)**: being supported by residential and corporate femtocell deployments, enabling camping user equipment (UE) devices to access local network services and applications directly without detouring via the operator's core network, ensuring higher quality of experience in addition to better network resource management.
- **Selected IP Traffic Offload (SIPTO)**: an approach that selectively offloads some IP traffic at both femtocell networks as well as at 3rd Generation (3G) and LTE macrocellular access networks, contributing toward a better distribution of traffic, avoiding congestion at the core network, and releasing core network resources for revenue-generating traffic.
- **IP Flow Mobility** and **Seamless Offload (IFOM)**: an approach that permits the establishment of different IP flows belonging to the same packet data network (PDN) connection via 3GPP macrocellular accesses and wireless local area networks.

Such solutions also provide the basis for further advanced, more comprehensive solutions, e.g., LIPA Mobility and SIPTO at the Local Network (LIMONET) [21], which combines selected features with further service requirements, particularly seamless service mobility. Data offloading solutions ensure efficient resource management by allocating different paths for selected services providing a means of regulation for the usage of the operator's scarce resources, i.e., PDN Gateways (PDN-GW) and Serving Gateways (S-GW). Operators may apply such a data offloading regulation on a per-UE basis and/or on a per-defined-IP-network basis, i.e., Access Point Name (APN), without compromising network security and user privacy. From the user perspective, data offloading should enhance the network experience introducing minimal complexity.

Currently, data offloading solutions have entered a mature phase with the initial architecture being completed. The focus of further research and standardization is toward decentralized network-based solutions, whereby local gateway (L-GW)

Manuscript received June 22, 2012; revised July 4, 2013; accepted September 16, 2013. Date of publication November 22, 2013; date of current version May 22, 2015.

T. Taleb and K. Samdanis are with NEC Laboratories Europe, 69115 Heidelberg, Germany (e-mail: Tarik.Taleb@neclab.eu; Konstantinos.Samdanis@neclab.eu).

Y. Hadjadj-Aoul is with IRISA Laboratory, University of Rennes 1, 35042 Rennes, France (e-mail: yassine.hadjadj-aoul@irisa.fr).

Digital Object Identifier 10.1109/JSYST.2013.2283957

selection and relocation is significant considering mobility and load balancing. This paper introduces a Domain Name System (DNS)-based solution to provide data offloading decisions for new and established session and analyzes the impact of L-GW selection considering service continuity.

An earlier version of the main contributions of this paper has been published in [26] and [27]. However, this paper attempts to integrate the different approaches proposed in [26] and [27] toward a comprehensive solution and to demonstrate their advanced performance through further simulations. Indeed, the contributions of this paper include mechanisms for traffic control using a DNS-based solution and a quantitative analysis of the various technical options to support service continuity of SIPTO traffic introducing minimal complexity with respect to both the core network and UE devices. Data offloading will not be efficient without decentralizing the mobile operator networks. For this purpose, small-scale core network nodes, e.g., PDN-GWs and S-GWs, are considered to be locally deployed to serve the local community of users, in a decentralized fashion [14]. In this paper, we also propose mechanisms to perform load balancing among L-GWs with the objective of improving user experience, verified via extensive simulations. We also performed a simulation study that demonstrates the impact of our gateway selection proposal compared with conventional proximity-based approaches for multimedia applications.

The remainder of this paper is organized as follows: Section II presents the state of the art. Section III presents our DNS-based SIPTO traffic control methods and our proposed gateway selection mechanisms considering UE devices in both idle and active modes. Section IV evaluates the overall approach and showcases its technical benefits. Finally, this paper concludes in Section V.

II. State of the Art

The EPC is designed to encompass different 3GPP accesses [i.e., 2nd Generation (2G), 3G, and LTE] and non-3GPP access [e.g., Worldwide Interoperability for Microwave Access (WiMAX), Code-Division Multiple Access 2000 (CDMA2000), and Single-Carrier Radio Transmission Technology (1xRTT)]. The richness of EPC accesses gave birth to a new 3GPP entity called Access Network Discovery Selection Function (ANDSF) that assists UE devices to find the best or most suitable access out of the many available ones [15]. It has also led to different interesting solutions, which can be classified into solutions for offloading IP traffic from the mobile core network and for exploiting the decentralization of the mobile network to optimize resources' usage.

A. Data Offload Solutions

Data offload solutions, including LIPA, SIPTO, and IFOM, have been widely studied within the 3GPP community. These 3GPP studies pertain to the specifications of architectural enhancements, mobility operations, and management functionalities. IFOM aims to address IPv6 mobility [16] with 3GPP suggesting the adoption of IETF RFC 5648 [17] to enable the support of multiple care-of address registrations and per-flow routing control. In IFOM, offloading decisions and mobility support are mainly based on the UE without imposing modifications by the core network. Consequently, IFOM is not providing resource control means to the mobile operators and is beyond the scope of the proposed DNS-based scheme, which enables operators to obtain a finer control and management of mobile traffic.

LIPA and SIPTO bridge such a control gap, allowing mobile operators to selectively offload traffic. They have therefore captivated 3GPP SA2 WG efforts toward LIPA/SIPTO specific enhancements in a Home (evolve) Node B (H(e)NB) subsystem and in macrocellular networks [7], [8]. A description of the different LIPA/SIPTO architectures is available in [5], which elaborates the main operations and requirements. A generic data offloading scheme beyond the conventional 3GPP framework focusing on local access networks and LIPA is described in [12]. It concentrates on UE devices with the support of multiple APNs and also on single-APN UE devices. However, the conventional LIPA/SIPTO solutions are based on predetermined policies or filters installed at local breakout points, lacking flexibility and dynamic quality-of-service (QoS) provisioning. A study that analyzes the QoS advancements in a mixed macrocellular and home network scenario is introduced in [11], quantifying the benefits of data offloading via LIPA and SIPTO services, considering the geographical position of micro- and macrocells as well as the radio characteristics of the system. Nevertheless, even this approach brings a static data offloading configuration.

Current efforts for data offloading solutions mainly focus on locating the offloading functionality on the operator's network and on providing methods for distinguishing SIPTO-related traffic. The perception of LIPA/SIPTO is that it is a service offered on a per-flow basis. Scalability improvements for the LIPA/SIPTO provision are proposed in [23], introducing an evolved packet system (EPS) bearer as the offload granularity considering one or more bearers within every PDN connection. Such an approach also adopts the conventional data offloading management mechanisms that rely on filters, which reside within each node and are used for regulating offloading decisions.

This paper advances the existing methods by introducing a flexible DNS-based approach for controlling SIPTO traffic handling, improving the performance of the predetermined filterbased schemes, considering also service continuity support for SIPTO traffic. Such DNS-based approach may easily reflect evolving traffic types and QoS demands, efficiently addressing issues related to static filter-based configurations, while ensuring a coordinated and unified policy on the entire network, enabling in this way efficient SIPTO mobility. The envisioned SIPTO service continuity is performed on a per-SIPTO-flow basis, similar to the IP flow mobility [13] and to the distributed IP flow mobility considered in [10], although the network environment and the means for achieving mobility support are different.

B. Solutions for the Reselection/Relocation of Data Anchor Gateways in Decentralized Mobile Networks

As elaborated above, the evolution of 3GPP radio access networks and their traffic characteristics have developed

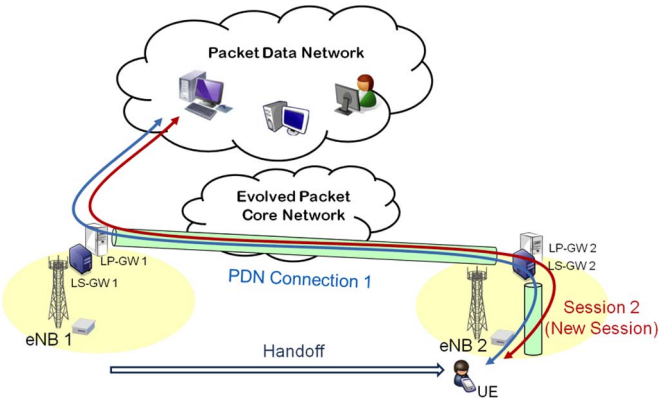

Fig. 1. Lack of mechanisms for PDN-GW selection/relocation per IP flow/session for UE devices in active mode in current 3GPP specifications.

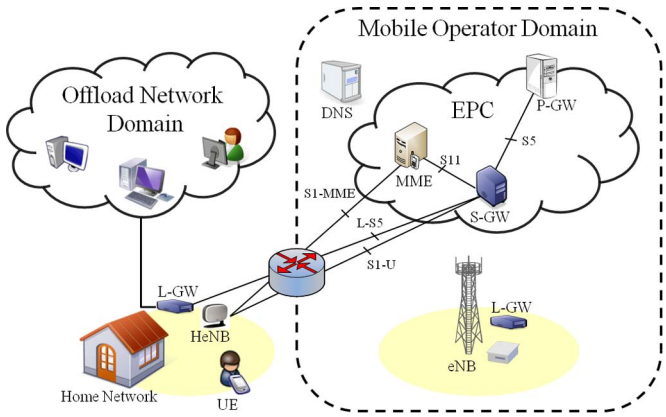

Fig. 2. Overall network architecture.

fundamental architecture and management features, leading the way towards the need for deploying a decentralized cellular infrastructure. Among these features, LIPA/SIPTO is of a significant role. In such a decentralized network, resource and session management, particularly when a UE changes its point of attachment to the network, is a process that needs to handle both already-established sessions and newly initiated ones. The key challenge is the fact that L-GWs and, hence, anchor points are distributed, so when a user moves, it may be more efficient to select or relocate sessions based on distance, load, and mobility patterns.

The network architecture of a decentralized EPS aiming for collocating gateway-related operations with evolve Node B (eNBs) towards the edge of the network is presented in [24]. Mobility and service continuity support is examined in [14], whereby a dynamic and distributed mobility management scheme is presented. The proposed approach merges the mobility anchors and base stations and uses tunneling to forward traffic upon a handover, allowing also the user to establish flows via different mobility anchors. While the proposed approach is proved to achieve efficient resource usage, its main drawback consists in the fact that it does not take into account neither load balancing nor resource management. Alleviating this drawback defines one of the main contributions of this work.

The study of the aforementioned solutions clearly allows routing different PDN connections through diverse access systems. Particularly, IFOM [6] allows a UE, equipped with multiple network interfaces, to establish multiple PDN connections to particular 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. While there is also another work, which enables UE devices 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 (see Fig. 1), a UE, supporting multiple APNs, may have different PDN connections, each associated with a different APN (e.g., Internet, IP Multimedia Subsystem). When a UE is using a PDN connection to access a particular APN, that PDN connection does not change until the UE becomes in idle mobility. In other words, PDN-GW relocation is recommended only during idle mobility, to avoid service disruption, because when the PDN-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 PDN-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 recommended. Its importance becomes further vital knowing the interest of operators to offload dump traffic as locally as possible to achieve the goals of SIPTO. In this paper, we also propose a set of mechanisms that enable a UE to know how and when to establish a new PDN connection for launching new IP sessions to a particular APN, above all in an efficient manner. This is done without impacting/compromising the ongoing (old) PDN connections to the same APN.

III. Enhanced Data Traffic Management

A. Envisioned Network Architecture

In order to develop a clear understanding of the proposed solutions, we initially describe the base network architecture and elaborate the role of the main network components. Fig. 2 illustrates the overall network architecture, which contains the individual network components, including a core DNS server, a Mobility Management Entity (MME), an S-GW, a PDN-GW, an eNB, a Home eNB (HeNB)/Security-GW, and a HeNB with a collocated local PDN gateway (L-GW). Small-scale PDN gateways and serving gateways are also locally deployed nearby eNBs in the macro network, resulting in a decentralized network architecture. The L-GWs are directly connected to an offload-enabled network domain, functioning as a local small-scale PDN-GW in case of UE devices supporting multiple APNs or as a simple L-GW including some necessary PDN-GW functions [5]. It should be noted that a DNS proxy could be potentially positioned at the local GW. For the sake of completeness, two types of UE devices are considered in this paper, namely, UE devices with a single PDN connection/IP address for both SIPTO and non-SIPTO traffic and UE devices

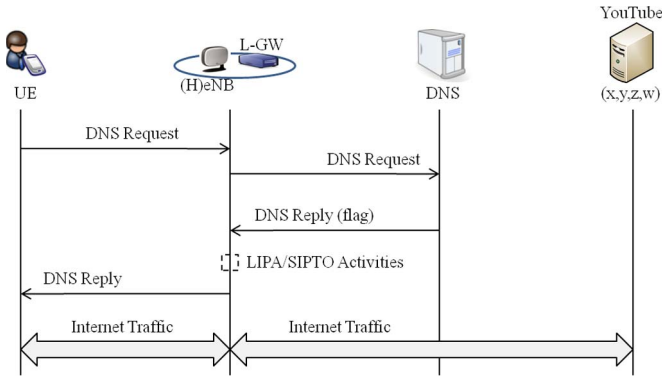

Fig. 3. SIPTO traffic handling/control based on explicit indications from operator's DNS.

capable of supporting multiple PDN connections/IP addresses, with at least one dedicated for SIPTO traffic.

B. Traffic Offload Handling and Control

The provision of SIPTO traffic control, i.e., separation of SIPTO traffic and core network data traffic, is typically performed using a set of predefined IP flow-based routing policies, managed by the Policy and Charging Rules Function (PCRF) or by the Home NodeB Management System (HMS) in case of small-cell deployments. Such a set of policies can be in the form of "IP flow filters," which are communicated and dynamically updated on eNBs and HeNBs to reflect evolving traffic policies. Hence, a new interface between PCRF and (H)eNBs should be defined, and the notion of IP flow-based routing policies for SIPTO should be established in PCRF or HMS.

"IP flow filters" specify traffic flows subject to SIPTO based on source/destination IP address/port number (or subnetwork address), protocol version, and optionally, the transport protocol type. The provision of "IP flow filters" to particular radio access points may appear as a scalable and responsive solution due to its distributed nature; however, major concerns are raised against the increased complexity associated with the dynamics and the proactive provision of evolving traffic policies [25]. Significant signaling cost is expected with the dynamic updates of "IP flow filters," while synchronization problems among neighboring eNBs or HeNBs may cause service inconsistency and problems to support traffic continuity for mobile users.

Hence, additional intelligence is required to support SIPTO and traffic continuity, while enabling operators with full control. Such a tight control is achieved by a centralized traffic management system, in charge of providing explicit indications regarding offload decisions to (H)eNBs during an IP flow setup by UE devices. In this vein, in [22], we proposed a DNS-based solution and analyzed a number of variants. According to our proposed DNS-based SIPTO control solution, as illustrated in Fig. 3, a UE that needs to establish a new IP connection contacts the DNS server following current standards. The DNS server then indicates in its reply the peer IP address and some information on the routing profile, i.e., how the traffic of the particular IP connection should be handled. Following the DNS reply, the L-GW takes appropriate actions. Depending on these DNS actions, four different SIPTO handling variants can be envisioned considering different UE capabilities with respect to multiple APNs support and different functionalities on the (H)eNBs as well as on the core network. In all variants, the DNS server indicates SIPTO support via a flag inside the DNS response message. With no specific purpose in mind regarding the order, the envisioned SIPTO handling variants are as follows.

1) **Simple Source NATing Solution** considering UE devices with a single APN, whereby an L-GW provides Network Address Translation (NAT) services for SIPTO traffic. The DNS indicates SIPTO support and the IP address of the peer, which is stored at the HeNB and is used for traffic offload enforcement. It should be noted that NAT-related limitations may restrict applications that do not work through NAT.
2) **Twice-NATing Solution** whereby both source and destination addresses are translated as offload traffic across the L-GW, which performs twice-NATing on SIPTO traffic. Traffic offload decisions are again indicated in the DNS reply message, whereas the IP address of the peer is stored in the L-GW, which associates it with the local DestNAT. Certain limitations subject to DestNAT apply, although IPv6 support as described in [9] can be envisioned.
3) **Simple-Tunneling Solution** whereby UE devices forward SIPTO traffic by establishing a tunnel toward the L-GW, which performs simple source address translation. Upon receiving a SIPTO indication, the L-GW adds to the DNS response toward the UE its IP address. UE devices realize that a particular connection is subject to SIPTO and tunnels the uplink (UL) traffic to the L-GW address. Source routing via IP header options could replace tunneling for IPv6 traffic.
4) **Multiple-APN UE-oriented Solution** considering UE devices that support multiple PDN connections to different APNs, with at least one dedicated for SIPTO. The DNS indicates to the UE which APN to use for SIPTO traffic, provided that the UE is aware of the configured APNs. UE devices, in turn, bind new IP flows to the UE's IP address associated with the recommended APN.

Further details on the mechanisms of these DNS-based SIPTO control variants along with a discussion on their advantages and pitfalls are available in [22].

C. L-GW Selection

Upon the setup of a new IP flow/session, a UE knows that the relevant IP traffic needs to be serviced via an L-GW using the aforementioned DNS-based SIPTO traffic handling. Nevertheless, the following issue is raised in case the UE has already some IP flows running on an L-GW: shall the UE establish the new IP flow/session via the currently used L-GW or shall it search for another L-GW that could be optimal and better than the currently used L-GW in terms of both load and geographical proximity. The remainder of this section introduces a number of L-GW selection mechanisms and compares them considering UE devices in one of the following two states: 1) EPS Connection Management (ECM)-idle mode and 2) ECM-connected mode. For the ECM-connected mode, the L-GW

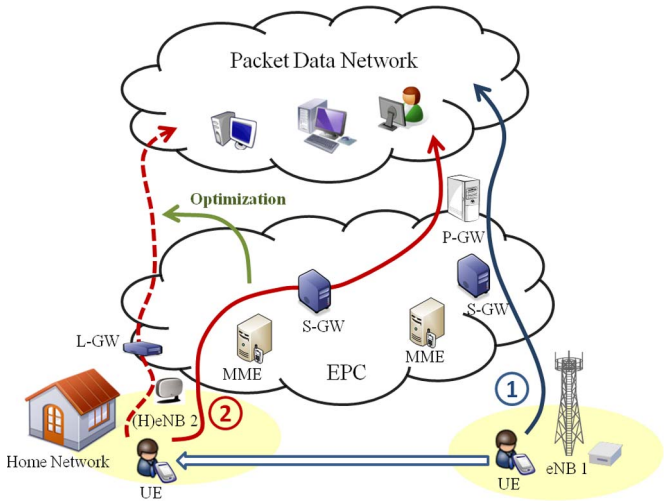

Fig. 4. Expected enhancement for UE devices in idle mode.

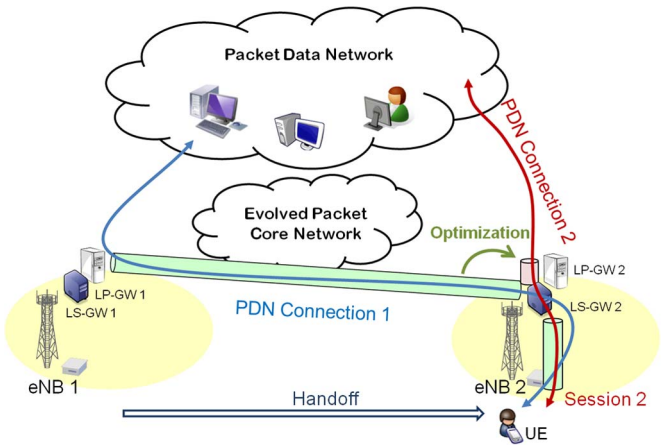

Fig. 5. Setting up new IP sessions via optimal PDN connections to a particular APN while old IP sessions (to the same APN using the old PDN connection) are not compromised.

selection analysis concentrates on UE devices with multiple PDN connections that traveled a significant distance from the original L-GW and need to reestablish the PDN connections, for optimization reasons.

In the EPS, once a UE establishes a PDN connection to a certain APN (e.g., Internet) via a specific PDN-GW, the PDN connection remains established even in case the UE goes idle. Such an "always on" feature is beneficial for UE devices that become active and need to communicate without requiring the reestablishment of a PDN connection.

Despite its responsiveness, "always on" connectivity has major drawbacks for supporting SIPTO in particular and in case of decentralized networks in general with the support of service continuity. Once a UE establishes a PDN connection with a local PDN-GW to an APN based on certain optimization constraints, such a local PDN-GW is maintained until the UE explicitly disconnects from this PDN. Hence, service continuity may result in suboptimal local PDN-GW usage, particularly if the user moves a significant distance away from the initial (H)eNB (see Fig. 1).

A simple example to demonstrate such deficiency is depicted in Fig. 4, considering a UE initially connected via the macrocell eNB 1 to a suitable PDN-GW (step 1) and then moving a significant distance while being idle, ending up at its home network where connectivity is provided by (H)eNB 2. The UE, during the entire movement phase, remains connected to the originally selected PDN-GW, even at its home network where another L-GW option offers better access. To overcome such a problem, this paper proposes mechanisms that assist UE devices to always establish PDN connections to the same APN via a suitable PDN-GW. Current PDN-GW selection mechanisms may associate UE devices with a suitable PDN-GW based on users' geographical/topological proximity to the network and/or load of gateways.

The process of reestablishing a PDN connection to the same APN may be either handled by the UE after receiving a notification from the network or enforced by the network itself via simply disconnecting the PDN connection based on certain optimization policies. In the former case, particularly in the case of a UE in idle mode, the UE may be notified during a tracking area update (TAU) procedure [3], [20]. In the latter case, the UE requests the network for the establishment of a new PDN connection according to one of the following solutions [20].

- **Option 1:** Periodically after a configurable time period; with the main issue being the increased signaling of stationary UE devices without any gain for the operator.
- **Option 2:** Upon TAU; i.e., whenever the UE changes tracking area ensuring that the UE has actually moved away from the original location.
- **Option 3:** Upon indication from the network based on global topology knowledge and resource availability, which may be provided as a part of idle-mode mobility procedures or with a special cause for the PDN disconnection.

Intuitively, depending on the application and service type, these procedures may have some impact on user experience as extra delay in the connection setup may be incurred.

As for UE devices in active mode, triggering L-GW selection for the accommodation of new IP sessions is a more complicated, yet a highly important procedure. A simple example that demonstrates this concept is illustrated in Fig. 5, whereby a UE that performs a handover to eNB2 establishes a PDN connection via the optimal LP-GW 2 to initiate new IP sessions to the same PDN/APN, while maintaining ongoing sessions (to the same APN) constant using the old PDN connection. Although in this example the proposed solution is elaborated for a UE under a handover scenario where another optimal PDN-GW (from the geographical proximity point of view) becomes available, it is also possible to apply the same solution for UE devices that remain camping in the same cell. In such a case, the purpose is to optimize PDN-GW association considering congestion, i.e., if the current PDN-GW becomes highly loaded, or optimize the overall network load balancing distributing the IP sessions of UE devices toward less loaded PDN-GWs.

While the need for PDN-GW reselection is clear for optimization reasons, an open issue yet remains regarding the way of establishing a new PDN connection, while also keeping the old ones uncompromised. Depending on the node that initiates

the establishment of new PDN connections, we envision the following methods.

- **MME-initiated:** whereby MMEs check the efficiency status of PDN-GW and its resource availability considering network and gateway load information and/or UE mobility prediction during the handover process. Once another more suitable PDN-GW becomes available, the used MME indicates via the existing handover signaling that for new IP sessions to a particular APN, the UE should consider requesting for a new PDN connection. Such an indication can be in the form of a flag or can explicitly indicate the IP address of the best PDN-GW to use, e.g., as a part of the TAU procedure.
- **UE-initiated:** whereby a UE queries the associated MME either when the UE sets up a new IP session to an APN with which it has an ongoing PDN connection, upon a number of handovers/topological distance from the current P-GW, or once a UE enters a new tracking area. Compared with the MME-initiated approach, this solution needs extra regulations since UE devices may easily generate unnecessary queries to the MME.
- **PDN-GW initiated:** whereby the serving PDN-GW simply rejects potential requests for any new IP sessions or notifies a selected set of UE devices to establish new PDN connections with other PDN-GWs when the current PDN-GW is suboptimal. Rejecting established IP sessions can be done via a new error message once the serving PDN-GW realizes by filtering traffic per IP flow/session that a UE has better serving PDN-GW options. Such an approach assumes that serving PDN-GWs have knowledge of a set of neighboring PDN-GWs and their load status. Notifying UE devices to change PDN-GW is performed under certain circumstances, e.g., when the serving PDN-GW load exceeds a certain threshold. It can be performed by introducing a specific message using S5/8 and S11 interfaces, or by a flag in either data packets or in existing signaling messages between PDN-GWs and UE devices (e.g., protocol configuration options [18]).

It should be noted that the MME-initiated solution and the UE-initiated solution require certain modifications for allowing multiple PDN connections to the same APN through the same access type, but only minimal extensions of the standard signaling interfaces, i.e., a new indicator (flag) between UE and MME. In contrast, the PDN-GW-initiated solution significantly increases the PDN-GW complexity and related standardization efforts, introducing also some minor modifications to UE devices.

Finally, to support the establishment of a new (most suitable) PDN connection for new IP sessions, UE devices need the ability to bind new IP sessions (when they are established) to a specific PDN connection or PDN-GW. A straightforward solution for such binding would be a mapping based on the destination IP address of the peer, application type, and protocol type. This way, a UE always knows which IP flow/session uses a particular 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.

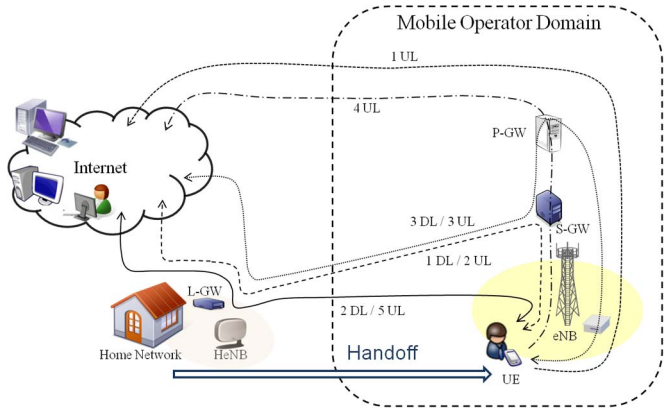

Fig. 6. DL/UL potential paths for an offloaded traffic after handoff.

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.

IV. Performance Evaluation

Here, we evaluate the proposed solutions and highlight their technical benefits considering first a quantitative analysis of the DNS-based traffic offload handling mechanisms and then a simulation study that elaborates the local PDN-GW selection and relocation considering the QoS impact and resource utilization.

A. Evaluating Service Continuity for the Proposed SIPTO Approach

This section quantitatively analyzes the performance of the proposed DNS-based SIPTO approaches with respect to service continuity support considering the UL and downlink (DL) potential paths, as illustrated in Fig. 6. In particular, three candidate paths are identified for DL traffic, namely, 1DL–3DL, and five possible paths are identified for UL traffic, namely, 1UL–5UL.

In case of IP flow filters, the UL SIPTO breakout point is located at the target (H)eNB using 1UL, as depicted in Fig. 6, without the option of service continuity support in case of handover. Similarly, the DNS-based Simple Source NATing solution cannot support service continuity since the correspondent peer will, upon a handover, receive the associated UE traffic with a different source IP address, namely, the UE's global IP address. Service continuity for ongoing SIPTO traffic can be supported only if the breakout point for ongoing connections remains constant at the L-GW of the source (H)eNB, which shall act as a data anchor point.

In such a case, traffic may traverse the core network before reaching the new location of the UE; hence, some further functionality is needed at selected network elements. For instance, paths that transverse the S-GW (e.g., 2UL in Fig. 6) require some extra functionality at S-GW to distinguish SIPTO from non-SIPTO traffic, break it out, and route it to the L-GW at the source (H)eNB. Directly connecting the L-GWs at the source and target (H)eNBs, by supporting data forwarding over the X2

TABLE I
COMPARISON AMONG THE DNS-BASED SIPTO TRAFFIC HANDLING VARIANTS

	Single APN				Multiple APN DNS-based
	IP flow filter	DNS-based			
		Simple NATing	Twice NATing	IP-in-IP Tunneling	
System complexity & cost	High	Low	Moderate	Moderate	Low
Transparency to UE	Yes	Yes	Yes	Transparent to application layer but network layer involved	Transparent to application layer but network layer involved
Service continuity support	No	No	Yes	Yes	Yes
Changes to DNS resolution	None	SIPTO flag	SIPTO flag & Dest-NAT	SIPTO flag & L-GW address	SIPTO flag or APN
Flushing of DNS caching at UE	No impact	Requires DNS cache flush upon (H)eNB change	Caching not possible (unless IPv6/IPv4 NAPT supported)	Requires DNS cache flush upon (H)eNB change	No impact
Packet inspection processing	Yes	Yes	Yes	Yes	No
Further issues	-	-	IP address space of LP-GW	-	APN priorization

interface as shown in the 2DL/5UL path in Fig. 6, is an efficient solution but may need certain (H)eNB/L-GW enhancements, which are outside the scope of this paper.

In case of the Twice-NATing solution, service continuity of SIPTO traffic is performed allowing the DL and UL traffic to follow paths 1DL or 3DL and 2UL or 3UL, as shown in Fig. 6, respectively. Considering the UL, path 3UL can be easily established since it merely requires the Twice-NATing functionality in the L-GW to intercept packets sent to the DestNAT address. However, the use of the alternative path 2UL requires some extra functionality in the S-GW in order to detect traffic targeted to the L-GW based on the DestNAT address range. Considering the DL, path 3DL follows the conventional standardized path, whereas the optimized 1DL requires further functionality alternations at the S-GW.

In the Simple-Tunneling-based solution, the IP address of the L-GW, used for the IP-in-IP tunnel, is routable within the operator network toward the source (H)eNB, and therefore, service continuity of SIPTO traffic can be supported by enforcing the DL and UL traffic to follow paths 1DL or 3DL and 2UL or 3UL, as shown in Fig. 6, respectively. In the UL, path 3UL can be easily established as this merely requires the support of the Simple Tunneling functionality in the L-GW, which needs to terminate the tunnel and route the traffic toward the destination. Path 2UL requires some extra functionality in the S-GW to detect traffic targeted to the L-GW based on the L-GW address range, whereas the DL path 3DL follows the conventional standardized path. Again, the optimized 1DL path is also supported but requires additional S-GW functionalities.

In the multiple-APN UE-oriented solution, service continuity for SIPTO traffic is supported as the standard mobility procedures ensure that the PDN connections are maintained during handover. The DL and UL traffic follow paths 1DL and 2UL, as shown in Fig. 6, respectively. It should be noted that unlike the other schemes, the multiple-APN approach avoids cashing problems related with DNS results and peer addressing in the service continuation process. A summary of qualitative comparisons among all introduced solutions is presented in Table I.

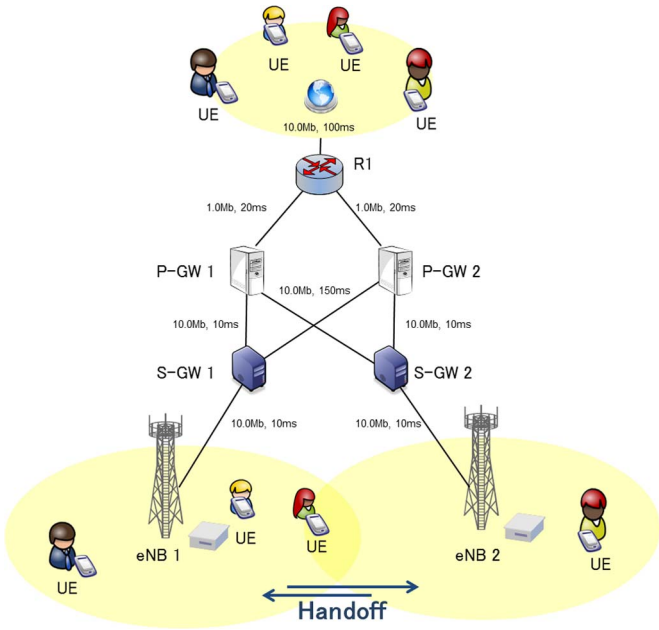

Fig. 7. Considered network topology.

B. Performance Evaluation of the L-GW Selection

Having described the details of our proposed L-GW selection mechanism, we now direct our focus to evaluating its performance via a simulation-based study using the Network Simulator 2 (NS2) [19]. We performed a series of simulations comparing our approach and the conventional one, which does not consider any parameters to enhance the efficiency of L-GW selection. The detailed criterion for the efficiency of an L-GW is defined by an operator policy, but typically, it may be derived from geographical proximity of a gateway (to the UE) or load. The focus, in the following, concerns only active UE devices.

The conducted simulations were run over the topology depicted in Fig. 7 for 900 s, which is a duration long enough to ensure that the system has reached its stable state. The network topology was simulated with care so that the selected simulation parameters would have no impact on the fundamental

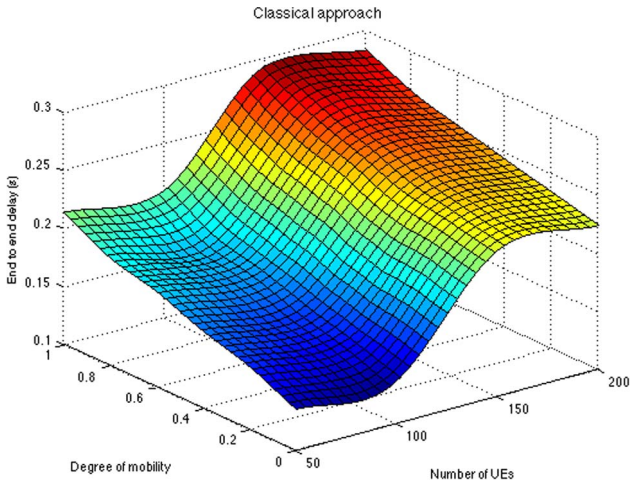

Fig. 8. Impact of the number of UE devices and the degree of mobility on the end-to-end latencies for the classical approach.

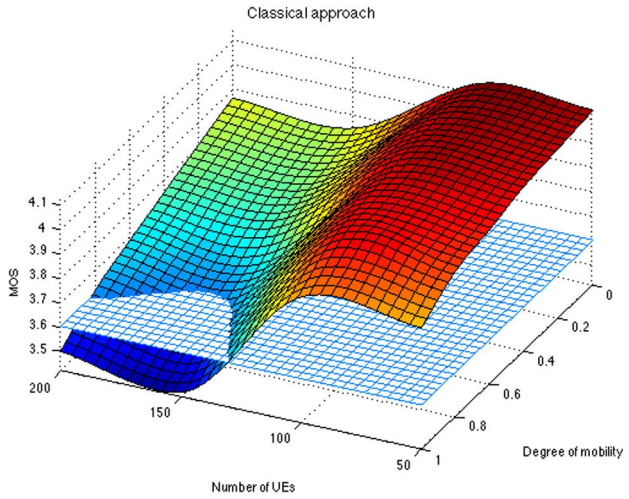

Fig. 9. Impact of the number of UE devices and the degree of mobility on the MOS for the classical approach.

observations made about the proposed approach. The simulations focus on the abilities of our mechanism to reduce packets' latencies while reducing the risk of packet drops. We simulate a variable number of active UE devices, within the range of 50–200 UE devices, being uniformly distributed around a surface of 20×20 km^2. All simulated UE devices are moving over the coverage areas of eNB1 and eNB2, changing their point of attachment to the network once the signal of another eNB becomes stronger than that of the source eNB. We use a random waypoint mobility model, where each mobile node selects a random destination and moves with a random speed, with a maximum speed limit equal to 20 m/s. At the beginning of the simulation, each UE initiates a Voice-over-IP session, starting between the interval of [0:100] seconds, using the ITU-T G.726 codec with a constant bit rate equal to 16 Kb/s. Upon moving to a new cell, UE devices also initiate new IP sessions that share the same characteristics as described above.

To truly assess the effectiveness of the proposed approach, we consider different degrees of mobility. A particular degree of mobility P means that each UE has probability P to be mobile, following the waypoint mobility model, described above. Otherwise, it is static.

We considered two metrics to evaluate the efficiency of the proposed mechanism: the average end-to-end delay of the different sessions and the perceived users' quality, which is measured using the E-model [28]. It is worth noting that packet loss caused by the conventional and proposed approach is similar, since the same amount of data is transmitted in both cases. Thus, the main factor considered for the evaluation of the perceived quality is the delay impairment factor I_d [28].

Fig. 8 demonstrates that the conventional/classical approach, which uses the old PDN-GW for the new IP sessions, experiences increased latencies compared with the proposed approach in Fig. 10. Indeed, with the conventional/classical approach, UE devices do not consider nearby PDN-GWs when establishing new IP sessions. Instead, they establish their new IP sessions via the old gateway, which increases the end-to-end delays as longer paths are considered. This degrades the mean opinion score (MOS), as illustrated in Fig. 9. The average quality deteriorates even further as the number of UE devices increases,

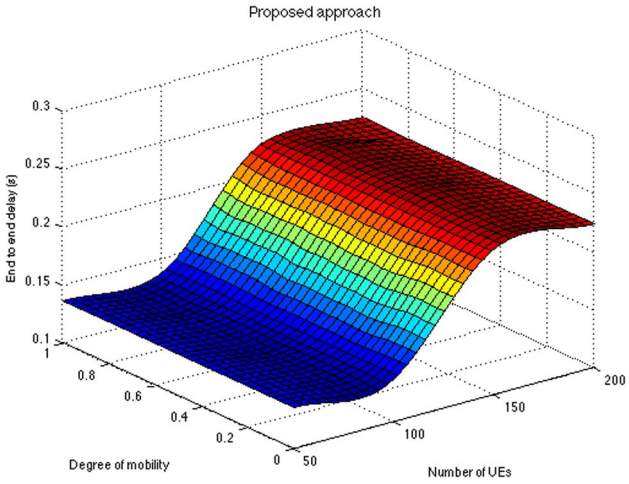

Fig. 10. Impact of the number of UE devices and the degree of mobility on the end-to-end latencies for the proposed approach.

because of congestion, and as the terminal mobility increases due to enhanced latency. Thus, Fig. 9 shows that many users are unsatisfied (i.e., MOS below 3.6) as the number of users becomes greater than 130 and the degree of user mobility becomes greater than 0.6, which is a value that represents relatively high mobility. In contrast, by exploiting local PDN-GWs to accommodate new IP sessions, which consists the proposed approach, better paths are always selected, thus guaranteeing shorter delays. In fact, this is shown in Fig. 10, in where the end-to-end delay remains approximately stable independent of user mobility, in contrast to the conventional/classical approach (see Fig. 8). As the delay remains approximately stable for variable degree of mobility, the proposed approach may accommodate a larger number of users before degrading the session quality. Indeed, the average MOS remains higher than 3.85, as shown in Fig. 11, which reflects that a majority of users are satisfied. For the worse cases, the proposed approach presents an improvement of about 11.8%.

It is also shown in Figs. 8 and 10 that the latencies significantly increase when the number of UE devices exceeds 150. This is principally due to buffering at the intermediate nodes,

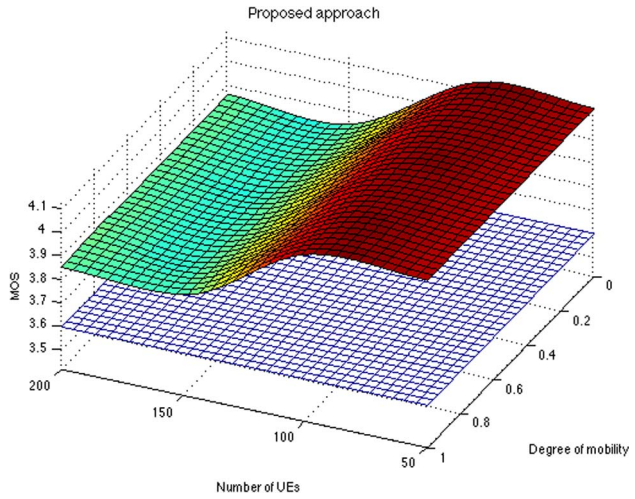

Fig. 11. Impact of the number of UE devices and the degree of mobility on the MOS for the proposed approach.

as 150 terminals induce more traffic than the one typically expected within the proposed topology. We can clearly see that the latencies increase faster in the case of the conventional approach than in the case of the proposed approach. This is also confirmed when there is no mobility as both schemes exhibit similar performance. In the case of the conventional scheme, the degree of mobility severely impacts the end-to-end latencies, whereas in our approach, the mobility does not have a serious impact on the system performance in terms of connection setup latency. In fact, when using the proposed approach, the system always selects the best gateways, which significantly decreases the average delay for each flow. It shall be recalled that this good performance is obtained for new IP sessions without disrupting the service of already-established applications via the old PDN-GW.

It is also shown in Figs. 8 and 10 that the latencies significantly increase when the number of UE devices exceeds 150. This is principally due to buffering at the intermediate nodes, as 150 terminals induce more traffic than the one typically expected within the proposed topology. We can clearly see that the latencies increase faster in the case of the conventional approach than in the case of the proposed approach. This is also confirmed when there is no mobility as both schemes exhibit similar performance. In the case of the conventional scheme, the degree of mobility severely impacts the end-to-end latencies, whereas in our approach, the mobility does not have a serious impact on the system performance in terms of connection setup latency. In fact, when using the proposed approach, the system always selects the best gateways, which significantly decreases the average delay for each flow. It shall be recalled that this good performance is obtained for new IP sessions without disrupting the service of already-established applications via the old PDN-GW.

V. Conclusion

In this paper, we have proposed solutions addressing the increased data traffic in 3GPP networks. We have proposed a set of DNS-based solutions for providing flexible and fine-grained traffic offload control, considering UE devices supporting only a single PDN connection and UE devices supporting multiple concurrent PDN connections. The service continuity support of SIPTO traffic during handover is achieved by enforcing both DL and UL traffic to traverse the L-GW at the (H)eNB, which anchors the IP flow/connection. For UE devices supporting only a single PDP Context/PDN connection, service continuity for SIPTO traffic can be supported either with no additional complexity in the core network (when the traffic tunneled to the PDN-GW and then routed to the local GW based on normal IP routing) or with little additional complexity in the S-GW (when the traffic is directly routed to the L-GW by the S-GW) through either Twice-NATing or Simple Tunneling. For UE devices supporting multiple PDN connections, an operator can simply control the traffic offload based on DNS replies that indicate to the UE which PDN connection to use for a particular IP flow/connection. For this, UE devices require minimal extra functionality but could also be involved in the decision process. Service continuity for SIPTO traffic is also supported in this solution based on the standard handover support.

When the offload is not possible or not desirable, we have proposed to maintain an always-best path selection between the mobile terminals or UE devices and their corresponding anchor gateways. Adequate methods were devised for UE devices 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 suitable PDN connection before creating new IP sessions, without compromising the old IP sessions. The suitability of a PDN connection may be defined by an operator policy, but typically, it may be derived from geographical proximity of a gateway to a UE or load. 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. While 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. Future research directions involve the consideration of further user-centric metrics in the process of selecting the best PDN connection based on user mobility history statistics and user activity.

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017*, Feb. 2013, White Paper.
[2] *MOBILE TRAFFIC GROWTH + COST PRESSURES = NEW SOLUTIONS?* Jan. 2010, New Mobile.
[3] "General packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access," 3GPP, Valbonne, France, TS 23.401, Rel.12, Jun. 2013.
[4] "Architecture enhancements for non-3GPP accesses," 3GPP, Valbonne, France, TS 23.402, Rel.12, Jun. 2013.
[5] "Local IP access and selected IP traffic offload," 3GPP, Valbonne, France, TR 23.829, Rel.10, Oct. 2011.
[6] "IP flow mobility and seamless wireless local area network (WLAN) offload; Stage 2," 3GPP, Valbonne, France, TR 23.261, Rel.11, Sep. 2012.

[7] "Service aspects; Service principles," 3GPP, Valbonne, France, TS 22.101, Rel.12, Mar. 2013.
[8] "Service requirements for home NodeBs and home eNodeBs," 3GPP, Valbonne, France, TS 22.220, Rel.11, Sep. 2012.
[9] K. Nishida, S. Ata, H. Kitamura, and M. Murata, "An Unified Multiplex Communication Architecture for Simple Security Enhancements in IPv6 Communications," in *Proc. EuroView*, Aug. 2010.
[10] D. Liu, J. C. Zuniga, P. Seite, H. Chan, and C. J. Bernardos, "Distributed mobility management: Current practices and gap analysis," Network Working Group, IETF Draft, Work in Progress, Jun. 2013.
[11] D. Calin, H. Claussen, and H. Uzunalioglu, "On Femto deployment architectures and marcocell offloading benefits in joint marco-femto deployments," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 26–32, Jan. 2010.
[12] P. Tinnakornsrisuphap, F. Ulupinar, J. W. Nasielski, J. Wang, P. A. Agashe, R. Gupta, and R. Rezaiifar, "Local IP access scheme," U.S. 2009/0268668 A1, Oct. 2009.
[13] "IP flow mobility and seamless wireless local area network (WLAN) offload," 3GPP, Valbonne, France, TS 23.261, Rel.11, Sep. 2012.
[14] C. B. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 46–53, Jun. 2012.
[15] "Access network discovery and selection function (ANDSF) management object (MO)," 3GPP, Valbonne, France, TS 24.312, Rel.12, Jun. 2013.
[16] H. Soliman, "Mobile IPv6 support for dual stack hosts and routers," IETF RFC 5555, Jun. 2009.
[17] R. Wakikawa, V. Devarapalli, G. Tsirtsis, T. Ernst, and K. Nagami, "Multiple care-of addresses registration," IETF RFC 5648, Oct. 2009.
[18] "Mobile radio interface Layer 3 specification; Core network protocols; Stage 3," 3GPP, Valbonne, France, TS 24.008, Rel.12, Jun. 2013.
[19] *The NS-2 Network Simulator*. [Online]. Available: http://www.isi.edu/nsnam/ns/
[20] T. Taleb, K. Samdanis, and F. Filali, "Towards supporting highly mobile nodes in decentralized mobile operator networks," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 5398–5402.
[21] "LIPA mobility and SIPTO at the local network," 3GPP, Valbonne, France, TR 23.859, Rel.12, Apr. 2013.
[22] K. Samdanis, T. Taleb, and S. Schmid, "Traffic offload enhancements for eUTRAN," *IEEE Commun. Surveys Tutorials*, vol. 14, no. 3, pp. 884–896, Third Quarter, 2012.
[23] L. Ma, W. Li, and X. Qiu, "Policy based Traffic Offload Management Mechanisms in H(e)NB Subsystem," in *Proc. 13th Asia-Pacific Netw. Oper. Manage. Symp.*, Taipei, Taiwan, Sep. 2011, pp. 1–6.
[24] Z. Yan, L. Lei, and M. Chen, "WIISE: A completely flat and distributed architecture for future wireless communication systems," in *Proc. WWRF 21*, Stockholm, Sweden, Oct. 2008, pp. 1–5.
[25] P. Bertin, S. Bonjour, and J.-M. Bonnin, "Distributed or centralized mobility?" in *Proc. IEEE GLOBECOM*, Honolulu, HI, USA, Dec. 2009, pp. 1–6.
[26] T. Taleb, K. Samdanis, and S. Schmid, "DNS-based solution for operator control of selected IP traffic offload," in *Proc. IEEE ICC*, Kyoto, Japan, Jun. 2011, pp. 1–5.
[27] T. Taleb, Y. Hadjadj-Aoul, and S. Schmid, "Geographical location and load based gateway selection for optimal traffic offload in mobile networks," in *Proc. IEEE/IFIP Networking*, Valencia, Spain, May 2011, pp. 331–342.
[28] *The E-model: A Computational Model for Use in Transmission Planning*, ITU-T Recommendation G.107, Dec. 2011, ITU-T Recommendation.

Tarik Taleb (M'03–SM'10) received the B.E. degree in information engineering (with distinction) and the M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively.

He is currently a Senior Researcher and a 3GPP standards expert with NEC Laboratories Europe, Heidelberg, Germany. He previously worked as Faculty Staff with Tohoku University, in a laboratory fully funded by KDDI Corporation. He is a member of 3GPP's System Architecture working group and, as such, was directly engaged in the development and standardization of the evolved packet system. His research interests include architectural enhancements to mobile core networks, mobile cloud networking, mobile multimedia streaming, congestion control, and social media networking.

Dr. Taleb has been on the editorial board of many Wiley and IEEE journals. He is serving as a Vice Chair of the Wireless Communications Technical Committee. He was a recipient of many awards, including the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher award.

Yassine Hadjadj-Aoul (M'05) received the B.Sc. degree in computer engineering (with high honors) from Mohamed Boudiaf University, Oran, Algeria, in 1999 and the Master's and Ph.D. degrees in computer science from the University of Versailles, Versailles, France, in 2002 and 2007, respectively.

He is currently an Associate Professor with the University of Rennes 1, Rennes, France, where he is also a member of the IRISA Laboratory and the INRIA project-team Dionysos. He was also a Postdoctoral Fellow with the University of Lille 1, Lille, France, and a Research Fellow, under the EUFP6 EIF Marie Curie Action, with University College Dublin—National University of Ireland, Dublin, Ireland. His work on multimedia and wireless communications has led to more than 40 technical papers in journals and international conference proceedings. His research interests include wireless networking, multimedia streaming, congestion control, and quality-of-service provisioning.

Dr. Hadjadj-Aoul has been on the technical program committee of many IEEE conferences.

Konstantinos Samdanis (M'04) received the Ph.D. degree in mobile communications from King's College London, London, U.K.

He is a Senior Researcher and a backhaul standardization specialist with NEC Laboratories Europe, Heidelberg, Germany. He is leading a research project on LTE network management, and he is the leader of the network virtualization Work Package in FP7 ITN CROSSFIRE MC. His standardization activities include SDN and Cloud for broadband multiservice networks.

Dr. Samdanis is the Editor of the Energy Efficient Mobile Backhaul work item in Broadband Forum. He has served as the Editor for IEEE COMMUNICATIONS MAGAZINE and MMTC E-Letters. He is also the Next Generation Networking Symposium Cochair at the 2014 IEEE International Conference on Communications.