Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 3938545, 20 pages
https://doi.org/10.1155/2017/3938545

*Research Article*

# Quantitative Comparison of the Efficiency and Scalability of the Current and Future LTE Network Architectures

**Morteza Karimzadeh,[1] Hans van den Berg,[1,2] Ricardo de O. Schmidt,[1,3] and Aiko Pras[1]**

[1] *Design and Analysis of Communication Systems (DACS), University of Twente, Enschede, Netherlands*
[2] *Netherlands Organization for Applied Scientific Research (TNO), The Hague, Netherlands*
[3] *SIDN Labs, Arnhem, Netherlands*

Correspondence should be addressed to Morteza Karimzadeh; mr.karimzadeh@gmail.com

Received 2 June 2017; Accepted 12 September 2017; Published 5 November 2017

Academic Editor: Luca Foschini

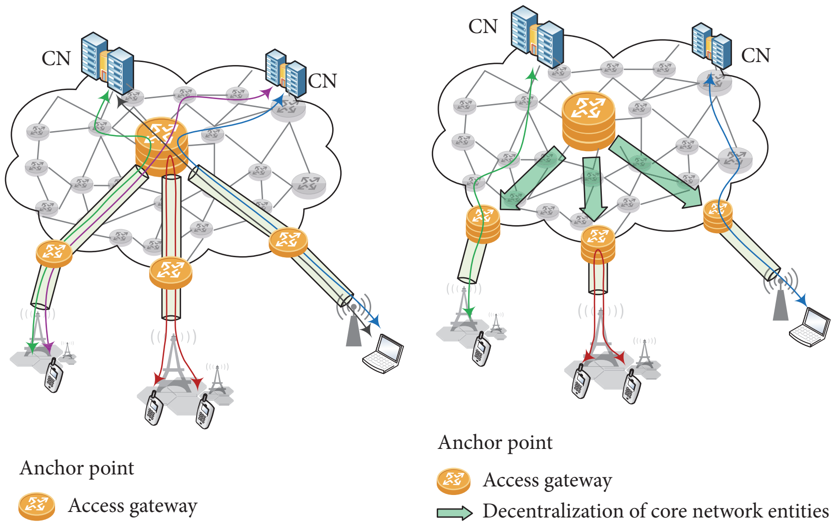
The core architecture of current mobile networks does not scale well to cope with future traffic demands owing to its highly centralized composition. Typically, it is believed that decentralization of the network architecture would be a sustainable approach to deal with ever growing amount of mobile data traffic. Nevertheless, the decentralization strategy of network architecture has not been properly examined through quantitative performance studies. Given that LTE will be the leading mobile networking technology in the coming 5–10 years, we conduct a hybrid study model to compare performance of current and future (decentralized) LTE network architectures. Particularly, our analysis presents numerical results quantifying impact of the number of attached nodes on the load at network routers and links, on the latency, and on the processing cost of the user's data and control planes. Analytical results demonstrate that decentralization of the LTE network architecture achieves higher performance compared to the current architecture and improves the latency and cost of data packet delivery more than 10 and 6 times, respectively. Furthermore, it is also observed that GTP outperforms PMIP for all studied performance metrics in the decentralized architecture and provides about twofold better latency and cost for data packet delivery and roughly 6 times lower data traffic load on the network routers.

## 1. Introduction

Over the last years, with the ubiquitous deployment and rapid evolution of mobile networks (e.g., 3GPP and WiMAX), the demand of accessing the Internet for mobile users has been soared dramatically. The mobile devices (e.g., smart-phones and tablets) become an integral part of everyone's daily live and generate a substantial part of the total Internet traffic, which is still increasing significantly.

It is forecasted that by 2021 there will be around 12 billion mobile devices worldwide, and 82% from these will be *smart mobile devices* generating up to 99% of all mobile data. Overall, mobile data is expected to increase from 7 EB per month, seen in 2016, to 49 EB per month in 2021 [1, 2]. Coping with such a demand in the current mobile networks is neither economically nor technically viable. The Radio Access Network (RAN) cannot be easily extended due to spectrum limitations. Furthermore, the core of mobile networks is highly centralized, which introduces scalability and reliability problems.

Mobile network operators increase RAN capacity by improving spectrum utilization in several ways, for example, deployment of small cells, selectively offloading traffic from cellular access to WiFi technology, and exploiting multicarrier techniques or multiple radio access technology approaches [3, 4]. The major challenge regarding the core networks (standardized by 3GPP, IETF) is related to the fact that a few high level network entities, entitled *anchor points*, manage both the *data plane* and the *control plane*. In such a centralized architecture, mobile node's (MN's) traffic must traverse the core anchor point and then go to the corresponding service node (CN); see Figure 1(a). This makes the network prone to several limitations, for example, suboptimal routing, low scalability, signaling overhead, and the lack of granularity on services [5, 6].



(a) (b)

Figure 1: A general view of the current (centralized) (a) and future (decentralized) (b) mobile network architectures.

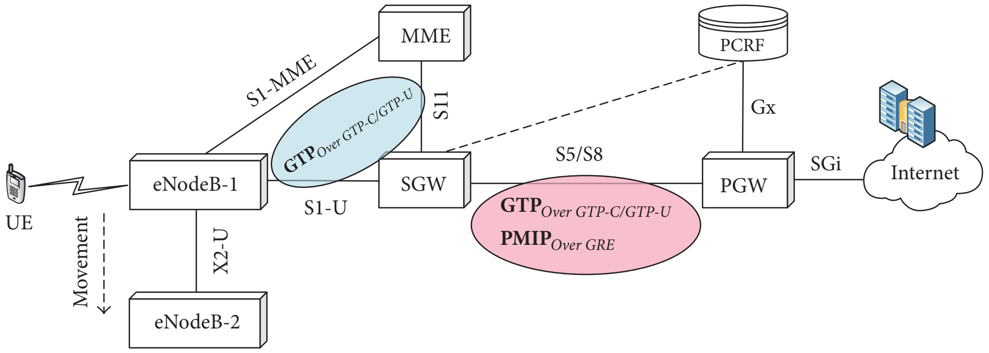


Figure 2: Current LTE architecture and its two mobility management protocols for 3GPP access.

The straightforward solution to cope with such an issue may consist of operators investment to upgrade the resources of the core network entities. Although this approach is technically feasible, network operators always prefer cost-effective and more sustainable solutions. Traffic offloading is an alternative approach to effectively reduce the traffic traversing through core network entities and to mitigate the traffic overhead on the limited resources of the core part. This can be achieved by placing small-scale anchor points in the proximity of the access network to locally handle MNs connections and traffic [7]. This essentially leads to a decentralized (flat) network architecture; see Figure 1(b). Even though decentralization also requires further investments for network architecture changes and management, it seems in the long term to be more cost-efficient than continuously extending capacity of the centralized architecture to cope with the future demands.

Long-Term Evolution (LTE) is expected to be the leading mobile networking technology in the next decade, handling substantial part of ever growing global mobile data traffic. It is predicted that the development of LTE will not keep up with the growth of mobile traffic, and while it supports almost 69% (out of 7 EB) of the current mobile traffic, it is estimated to handle about 79% (out of 49 EB) of the worldwide mobile data traffic by 2021 [1, 2].

Therefore, in this paper, we pay special attention to the LTE system and perform a hybrid study, including simulation and analytical models, to analyze in detail the performance and scalability of the current and a decentralized LTE network architecture. In particular, we analytically evaluate both network architectures to quantify how much the load on network resources (routers and links) and the latency and cost of the user's data plane (traffic forwarding procedures) and control plane (attachment and handover procedures) are affected by the increasing number of attached mobile nodes. We analyze performance of GPRS Tunneling Protocol (GTP) and Proxy Mobile IP protocol (PMIP), the two commonly used protocols in the LTE system to handle MNs data traffic and mobility in 3GPP access (Figure 2).

Summarizing, our main contributions in this paper are as follows:

(i) Develop a detailed model of the functions of GTP and PMIP protocols for 3GPP access on both current (centralized) and future (decentralized) LTE network architectures.

(ii) Carry out a hybrid study combining simulation and analytical modeling, capturing the most essential characteristics of the system while abstracting from the less important details, in order to evaluate various scenarios in feasible time.

(iii) Using the developed approach, derive various metrics to quantify and analyze the performance and scalability of the LTE network system (current architecture versus decentralized architecture).

(iv) Relying on the obtained numerical results, provide an intuition of the expected impact of the number of subscribers on the different LTE core network architectures.

The rest of this paper is organized as follows: Section 2 provides concisely the necessary background about the current LTE and its existing mobility management solutions for 3GPP access. Section 3 discusses our hybrid modeling study. Section 4 describes in detail the analytical calculation of the performance metrics and the evaluation procedure. The results to compare the performance and scalability of different network architectures are presented in Section 5. Section 6 reviews in brief the recent related works and specifies how our work differs from the literature. Finally, the paper ends up with the conclusion and discussion parts in Sections 7 and 8, respectively.

## 2. LTE Architecture

This section gives a brief overview for current LTE network architecture and the two existing mobility management protocols for 3GPP access, which are essential for perceiving the problem statement as well as the proposed model, in this work.

The LTE architecture is hierarchical and defines the Evolved Packet System (EPS) consisting of Evolved Universal Terrestrial Radio Access Network (E-UTRAN) and Evolved Packet Core (EPC). The E-UTRAN consists of a network of radio base stations (eNodeB—*evolved Node B*) that provide radio connectivity to User Equipment (UE). The EPC is a multiaccess IP-based network that allows for a common core network for 3GPP and non-3GPP radio access and fixed access.

The EPC consists of four main elements (Figure 2) that allow for the convergence of packet-based services [10]:

(i) Serving Gateway (SGW) is a user plane node that provides data paths and routes traffic between eNodeBs and PGW. It also acts as a local mobility anchor for UEs performing handover between eNodeBs.

(ii) PDN Gateway (PGW) provides the connection between the EPC and other external IP networks, as well as several additional functions, such as IP address anchoring and allocation, routing, packet filtering and monitoring, and policy control.

(iii) Mobility Management Entity (MME), whose key role is to handle UE mobility, also performs the control functions to access the LTE, assigns network resources, and supports roaming and handover procedure.

(iv) Policy and Charging Rule Function (PCRF) dynamically controls and manages all data sessions and determines quality of service (QoS) policies and charging rules to SGW and PGW.

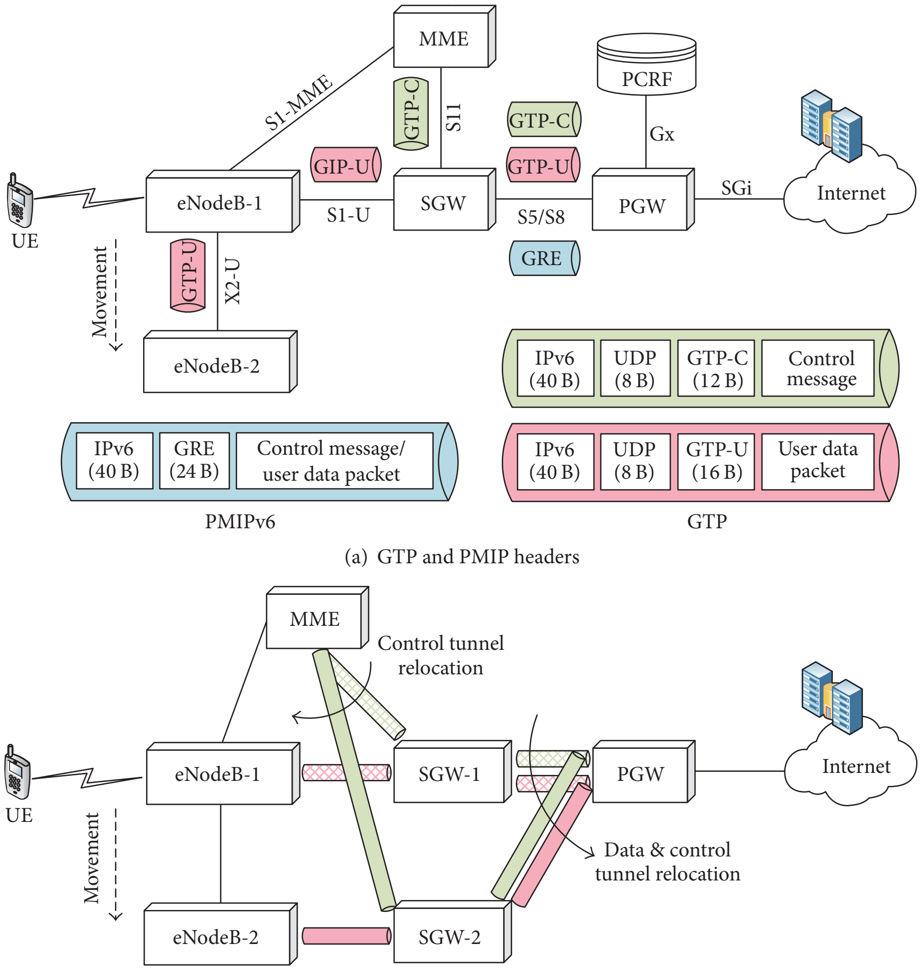
*2.1. Mobility Management Protocols.* Current EPC may use GTP or PMIP protocol to support UE's mobility for 3GPP access networks [11]. These protocols allow an uninterrupted handover for the UEs during internetwork mobility. To manage UE's mobility using GTP or PMIP, the PGW might connect to SGW via S5/S8 interfaces (Figure 2) for nonroaming and roaming scenarios, respectively [11].

The GTP protocol is able to fully handle the control and data planes. It can forward the UE's downlink packets from source location to target place during handover. The PMIP however can only handle UE's mobility and perform data forwarding after handover procedure. Moreover, it is not able to control the bearers and QoS signaling. When PMIP is used over the S5/S8 interface, the GTP bearers are only defined between UE and the SGW. In this case, the SGW takes over the bearer binding operations, and an additional connection (dash line in Figure 2) needs to be created between the SGW and the PCRF to provide the required information on QoS policy [9, 10, 12].

UE may perform a handover in either idle or active mode. In idle mode, the UE stays in power consumption mode and does not inform the network about the location information. The network uses the *tracking* and *paging* procedures to discover position of UE. In active mode, UE's mobility is completely under control of the network. The decision to perform a handover and to choose the target cell is handled by the network, based on measurements performed by the eNodeB and the UE. During the handover procedure, GTP can locate the UE's position, even in the idle mode to establish the required data and control planes tunnels. However, PMIP does not support tracking and paging functions and the UE needs to be always in active mode. Therefore, GTP protocol is mostly used in the access network between SGW and eNodeBs to eliminate the aforementioned drawbacks of PMIP.

The decision for using GTP or PMIP over the S5/S8 interface depends on several parameters, such as technical support and existence of roaming scenarios among the 3GPP access and non-3GPP access networks. Note also that mobility management protocols for non-3GPP, such as Mobile IP (MIP) and Dual Stack MIP (DSMIP) [10], are out of the scope of this study.

*2.2. Data and Control Planes Tunneling.* When UE attaches to the LTE network, several control messages are exchanged



(a) GTP and PMIP headers

(b) Tunnel updating during handover

Figure 3: LTE data and control planes for 3GPP access.

between the eNodeB, MME, SGW, and PGW. Using GTP protocol, a successful attachment results in a tunnel for the user data plane (GTP-U) between the eNodeB and SGW and between the SGW and PGW. Another tunnel for the control plane (GTP-C) is established between the SGW and PGW and also between the MME and SGW (Figure 3(a)). While GTP-U simply transports data packets within the core and radio access networks, GTP-C tunnel is used to exchange the control messages for handling UE's mobility, as well as for path management and tunnel management (e.g., adjusting QoS parameters, updating sessions for roaming subscribers, and activating and deactivating subscriber sessions). These tunnels are created for each individual UE traffic flow (IP traffic). Each GTP tunnel has an identifier, entitled Tunnel End Point Identifier (TEID). Based on the TEID the network is able to choose the appropriate tunnels to transfer data packets and control messages between the end points.

In the case of PMIP protocol, basic IP connectivity over Generic Routing Encapsulation (GRE) tunneling is used between the SGW and PGW, and a GRE key is used to identify each tunnel.

As Figure 3(b) shows, during a handover between neighboring eNodeBs the GTP-U tunnel is updated, and if the handover is between neighboring SGWs both GTP-U and GTP-C (or GRE) are updated.

Note that tunneled packets might be further encapsulated by IPsec protocol in order to protect both control messages and data. In this study we only focus on the data and control planes' tunnels established over the transport network. Additional control signaling involved on UE's attachment, handover, and data delivery procedures with the other LTE components as well as IPsec is out of scope in this paper.

*2.3. Mobility Management Messages.* The signaling for UE mobility control is similar for GTP and PMIP in the access network (Figure 2). Differences depend on the mobility protocol in the core network over the S5/S8 interfaces.

Figure 4 shows the GTP and PMIP mobility messages in the core network during UE attachment and handover procedures.

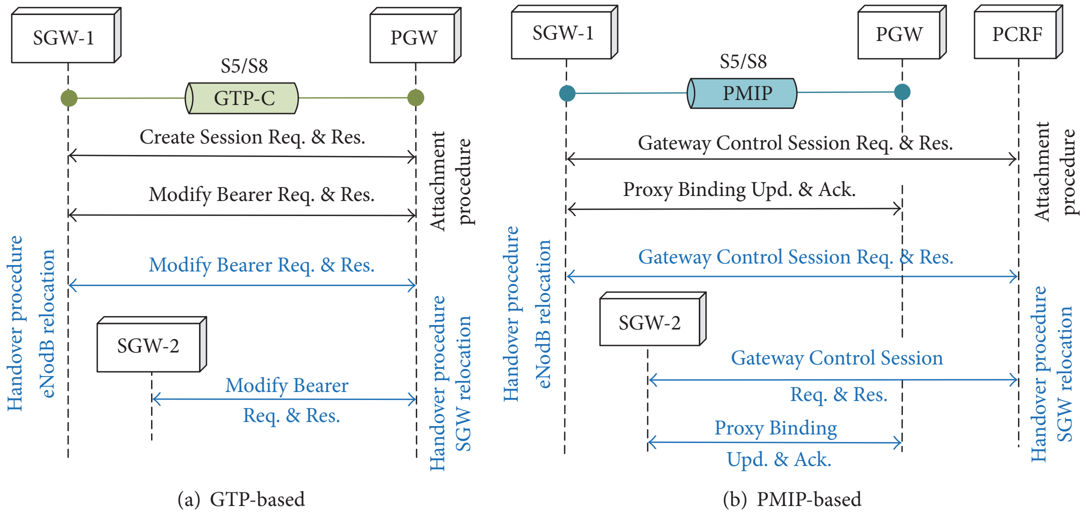


Figure 4: UE attachment and handover procedures messages.

Table 1: Mobility messages for GTP [8].

| Message | Size | Headers |
|---|---|---|
| Create Session Request (C.S.Req) | 335 B | 60 B |
| Create Session Response (C.S.Res) | 224 B | 60 B |
| Modify Bearer Request (M.B.Req) | 67 B | 60 B |
| Modify Bearer Response (M.B.Res) | 81 B | 60 B |

Table 2: Mobility messages for PMIP [9].

| Message | Size | Headers |
|---|---|---|
| Gateway Control Session Request (G.C.S.Req) | 336 B | 68 B |
| Gateway Control Session Response (G.C.S.Res) | 972 B | 68 B |
| Proxy Binding Update (P.B.U) | 104 B | 64 B |
| Proxy Binding Acknowledge (P.B.A) | 104 B | 64 B |

*2.3.1. For the GTP Protocol.* Table 1 lists the messages used in the procedures described in the following as well as their respective sizes. Note that the size information does not include the extra GTP-C tunnel header size, which is presented in a separated column.

For GTP protocol, when UE's switch is turned on, an *Attach Request* message is sent to the MME through the eNodeB. The MME then sends a C.S.Req message to the SGW, which is forwarded to the PGW. This request is meant for setting up the UE's default bearer and also for requesting a Packet Data Network (PDN) connectivity. The reply from the PGW, containing an IP address for the UE and a default bearer ID, is also forwarded by the SGW to the eNodeB and the MME. With this information the attachment procedure is concluded and the traffic from the UE can flow from the eNodeB to the SGW via the S1-U interface (Figure 3(a)). The MME still sends a M.B.Req message to the SGW, which in turn is forwarded to the PGW, containing the TEID assigned to the eNodeB. The PGW finally replies to this request with a M.B.Res message, and the user data plane is then set up for traffic flow in the core network.

During a handover between eNodeBs, the target eNodeB sends a *Path Switch Request* to the MME, informing that the UE has changed its physical location. The MME then sends to the SGW a M.B.Req message with the address of the new eNodeB and the TEID of the user plane for downlink. This information is forwarded by the SGW to the PGW. The PGW replies with a M.B.Res message to the SGW that starts forwarding downlink packets to the target eNodeB (current UE location).

On realizing that the SGW has been relocated (from the *Path Switch Request* message sent by the target eNodeB), the MME sends a C.S.Req message to the target SGW. This message contains the PGW addresses, the TEIDs used for uplink traffic, the address of the target eNodeB, and the protocol type used over S5 or S8 interface. The target SGW assigns the addresses and TEIDs for downlink traffic from the PGW and sends a M.B.Req message to the PGW informing about the changes. The PGW then updates its context field and replies to the SGW with a M.B.Res message including its address and TEIDs information [11].

*2.3.2. For the PMIP Protocol.* Table 2 lists the messages used in the procedures explained in this section as well as their respective sizes. Note that the size information does not include additional header sizes, which can differ and are presented in a separated column.

For PMIP the initial attachment procedure of UE is similar to that of the GTP. The only difference is that the SGW has to establish a control session towards the PCRF to obtain the QoS policy information needed to perform the bearer binding. All these are obtained through a G.C.S.Res message sent by the PCRF to the SGW in reply to a G.C.S.Req message. This message exchange is done over the Stream Control Transmission Protocol (SCTP).

To establish the default bearers, the SGW sends to the PGW a P.B.U message containing, among others, the GRE key

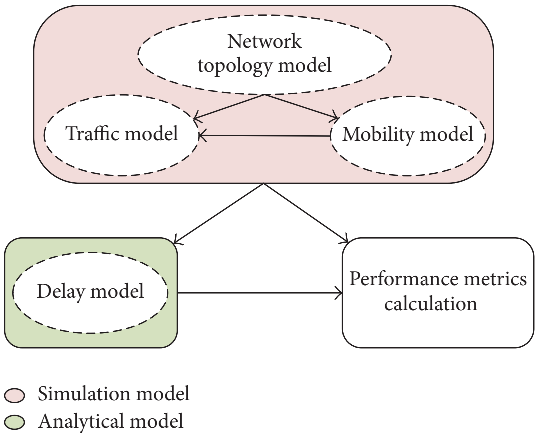
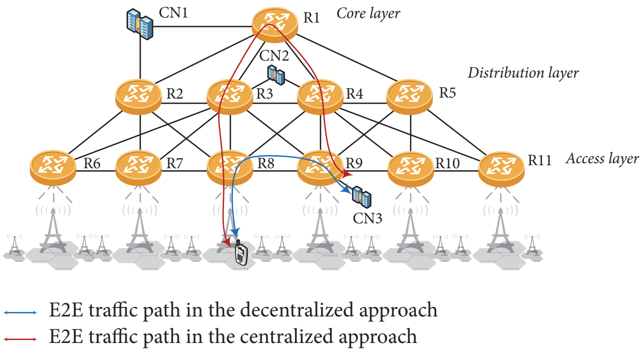


Figure 5: The hybrid modeling study.


CN1
R1
Core layer
CN2
Distribution layer
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
Access layer
CN3
E2E traffic path in the decentralized approach
E2E traffic path in the centralized approach

| | R1 function | R2–R5 function | R6–R1 function |
|---|---|---|---|
| Centralized approach | PGW | Router | SGW |
| Decentralized approach | Router | Router | S/PGW |

Figure 6: The network topology used in the analytical model.

for downlink traffic, address information of the UE to request an IPv6 prefix, and charging characteristics. The PGW replies to the SGW with a P.B.A message containing, among others, the UE's address, the GRE key for uplink traffic, and the charging ID. These two messages are exchanged over a GRE tunnel, and after this exchange the SGW and the PGW set up an additional bidirectional GRE tunnel for forwarding of UE's data flows.

During handover between eNodeBs, the SGW sends to the PCRF a G.C.S.Req message informing about the change of UE's location, which was received from the MME. The PCRF replies to the SGW with a G.C.S.Res message providing, among others, the updated QoS policy and charging rules. In case of a handover with SGW relocation, in addition to the PCRF messages, the new SGW have to exchange the P.B.U and P.B.A messages with the PGW [10].

In the following, Section 3 describes our hybrid modeling approach to quantify the gains on performance and scalability in four different scenarios: the current (centralized) and a decentralized EPC architecture with GTP and PMIP mobility management protocols. Next in Section 4, we present the calculation of various performance metrics in detail.

## 3. Hybrid Simulation and Analytical Modeling

This section describes our hybrid study containing simulation and analytical modeling. As described in the Introduction, the simulation model captures the dynamic behaviour of the system at mobile node and connection level, delivering information about, for example, link load and load of routers in the different network layers. This information, together with other network and traffic parameters, is used in the analytical delay model to derive queuing delay for data packet and control messages processing in the nodes as well as queuing delays for transmitting this information on the network links. Eventually, these intermediate results are used to calculate end-to-end data packet delivery delay and cost, control plane, and data plane loads.

Figure 5 shows schematically how the various parts of our hybrid model relate to each other.

It is important to remark that, in our approach on one hand, we capture the most essential characteristics of the system and on the other hand abstract from the less important details to set up an straightforward environment to perform the modeling and analyses in a feasible time. We use MATLAB as the environment to implement the models and perform the analysis. Using a normal desktop computer, it takes only a couple of minutes to carry out the complete tasks for all scenarios.

Table 3 presents the notation used in the hybrid model as well as in the analytical calculation to compute different performance metrics (Section 4).

*3.1. Network Topology Model.* Figure 6 shows the network topology used in our modeling. This topology follows the Cisco 3-layer hierarchical model consisting of the core, distribution, and access layers. The core layer handles traffic transferred to and from the routers at the distribution layer. The distribution layer enables the communication between routers from the core and access layers. The access layer mainly controls the attachment of end users and devices to the network.

We define two network scenarios in our models: the current (centralized) and future (decentralized) network architectures in context of the LTE system. In each of these, the routers in the core and the access layers play different roles.

In the *first scenario*, for the centralized architecture, the PGW is placed in the top of the topology (R1 in Figure 6), and the SGWs are placed at the edge routers (R6 to R11). In this scenario the PGW is the anchoring point of the network and handles all data and control plane operations. The SGWs provide the data paths towards the core, and route UE packets between the access network and the PGW.

In the *second scenario*, for the decentralized architecture, we define S/PGWs, which are physical nodes combining functions of the SGW and PGW elements. In this scenario the S/PGWs operate as the distributed anchoring points and handle the data and control plane operations locally. These are

Table 3: The notations used in the hybrid model and analytical calculation.

| Symbol | Definition | Value |
|---|---|---|
| $H_{x\text{-}y}$ | Number of hops between arbitrary nodes $x$ and $y$ | |
| $d_{(x)_l}$ | Queuing delay of $x$ in the network link $l$, due to traffic load | $x \in \{p, (p + t_u), (m + t_c)\}$ |
| $\delta_{(x)_l}$ | Transmission delay of $x$ in the network link $l$ | $x \in \{p, (p + t_u), (m + t_c)\}$ |
| $\tau_{p_{(p/m)_j}}$ | Packet ($p$) or message ($m$) processing delay in router $j$ | |
| $\tau_{r_{(p/m)_j}}$ | $p$ or $m$ routing delay in router $j$ | |
| $\tau_{t_{(p/m)_x}}$ | $p$ or $m$ tunneling delay in the node $x$ | $x \in \{\text{SGW}, \text{PGW}, \text{S/PGW}\}$ |
| $d_{(p/m)_j}$ | $p$ or $m$ queuing delay in router $j$ due to processing load | |
| $N_{h_i}$ | Number of handovers ($h$) for node $i$ during simulation time | |
| $N_{k_i}$ | Number of paths ($k$) created for node $i$ due to mobility | |
| $\text{MNAP}_{(x)}$ | Messages number of initial attachment procedure | $x \in \{\text{GTP}, \text{PMIP}\}$ |
| $\text{MNHP}_{(x)}$ | Messages number of handover procedure | $x \in \{\text{GTP}, \text{PMIP}\}$ |
| $\text{ALDPD}_i$ | Average latency of packet delivery for node $i$ | |
| $\text{ACDPD}_i$ | Average processing cost of packet delivery for node $i$ | |
| $\text{LDPD}_k$ | Latency of packet delivery in path $k$ | |
| $\text{CDPD}_k$ | Processing cost of packet delivery in path $k$ | |
| $\text{LAP}_i$ | Latency of attachment procedure for node $i$ | |
| $\text{CAP}_i$ | Processing cost of attachment procedure for node $i$ | |
| $\text{LHP}_i$ | Latency of handover procedure for node $i$ | |
| $\text{CHP}_i$ | Processing cost of handover procedure for node $i$ | |
| $\text{CML}_{(N)_j}$ | Control message lode (number) over router $j$ during simulation time | |
| $\text{CML}_{(\text{KBs})_j}$ | Control message load (size) over router $j$ during simulation time | |

placed in the access layer (R6 to R11). Herein, the core router (R1) is only used when data has to be routed through it.

In both scenarios, routers on the distribution layer perform as normal L3 routers and enable the communication between the core and access layers.

We consider three static CNs located at each layer. These represent the data centers that in reality could be geographically distributed. The data traffic to UE is transmitted with the shortest path through (CN → PGW → SGW → UE) or (CN → S/PGW → UE) for the centralized and decentralized scenarios, respectively. It is important to mention that we ignore all the detailed functions of the routing mechanism (e.g., load balancing) in our model. As in this paper we perform a *comparative analysis* for the different network architectures, using a plain routing solution having no severe effect on the final outcome of the comparison.

The two network architecture scenarios described above are later (in Section 4) combined with both the GTP and PMIP protocols to define the four scenarios, being analyzed in this work.

*3.2. Mobility Model.* A straightforward approach to model mobility is by obtaining the time spent by a mobile node (also known as *the residence time* or *the dwell time*) in each SGW (or S/PGW) during movement. The Fluid-Flow mobility model is a simple approach to drive the mobile node's dwell time in cellular networks, which has been extensively used in previous work [13–16]. By applying the Fluid-Flow model, where $E(v)$ is the average speed of a mobile node, the average dwell time of the node in each SGW is given by

$$E\left[T_{\text{dwell}}\right] = \frac{\pi \times A}{E(v) \times P} = \frac{\pi \times \left(\pi R^2\right)}{E(v) \times (2\pi R)} = \frac{\pi R}{2E(v)}, \qquad (1)$$

where $R$, $A$, and $P$ denote the radius, coverage area, and perimeter of each SGW, respectively. Note that, for each mobile node, $E(v)$ is arbitrarily chosen from the predefined values listed in Table 5.

We assume that a mobile node randomly attaches to one of the SGWs and after passing the dwell time it starts to move towards the neighbor SGW. This movement is modeled by randomly choosing one of the neighbor SGWs and staying in it for the dwell time. This procedure is continuously repeated during the simulation time. Therefore, the number of handovers for each mobile node ($N_{h_i}$) can be easily derived using its dwell time and the simulation time. For both scenarios, mobile nodes choose the same trajectories, affecting the paths created for the related control and data planes during the simulation time. In our model we assume that $E(v)$ for each mobile node is constant during the simulation time. It is also assumed that each SGW covers a circular domain consisting of three eNodeBs, and therefore for a SGW relocation there will be three handovers between neighboring eNodeBs.

*3.3. Traffic Model.* We assume that each mobile node randomly attaches to one of the SGWs (or S/PGWs) and starts to download data from one of the CNs, also randomly chosen.

Every mobile node has a single active session to one of the CNs during the whole simulation time. Every node follows the mobility model described in the previous section and after staying at each access layer entity (SGWs or S/PGWs) for its dwell time moves to one of the neighboring entities.

A Poisson traffic stream with average rate of 100 packets per second (CNTR) is generated from the CNs towards the connected mobile nodes, simulating the download of data by the attached nodes. For the sake of simplicity, this model ignores the packet level details (e.g., packet loss and loss recovery mechanism). To avoid IP packet fragmentation as a result of tunneling overhead at the core network, we set the size of packets from the CNs to 1200 Bytes ([17] advises a default MTU size of 1280 Bytes).

*3.4. Delay Model.* During a transmission between two endpoints, data packets or control messages may be delayed due to, for example, link congestion and queues. In our model, the network link delay in each hop includes the transmission delay ($\delta_{(x)_l}$) and the queuing delay ($d_{(x)_l}$), where $x$ is either a pure or a tunneled data packet or control message (Table 3). Applying an M/M/1 queuing model, the average delay ($T_{(p)_l}$) of a data packet of size $p$ transmitted in the network link $l$, with transmission rate TR and traffic load CNTR per mobile node, is given by

$$T_{(p)_l} = \left[\delta_{(p)_l} + d_{(p)_l}\right] = \frac{\delta_{(p)_l}}{1 - L_{(p)_l}}, \quad L_{(p)_l} = \frac{\lambda_l \times p}{\mathrm{TR}}. \tag{2}$$

$\lambda_l$ denotes the network link traffic, derived from the simulation by keeping track of all paths that are established using link $l$, as well as their duration, and taking into account the packet rate.

For a data packet $p$ or a control message $m$, and a node $j$, the router delay consists of the processing delay ($\tau_{p_{(p/m)_j}}$), the data packet encapsulation/decapsulation or control message construction/extraction delay ($\tau_{t(p/m)_j}$) during a tunneling, the routing delay ($\tau_{r_{(p/m)_j}}$), and the queuing delay ($d_{(p/m)_j}$) at the router (Table 3). For matters of simplicity we assume that $\tau_{p_{(p/m)_j}} = \tau_{t_{(p/m)}} = \tau_{r_{(p/m)}}$. Similar to the network link delay, for a network router $j$ with the processing rate PR, the average delay ($T_{(p)_j}$) for a data packet $p$ is defined by

$$T_{(p)_j} = \left[\tau_{p_{(p)_j}} + d_{(p)_j}\right] = \frac{\tau_{p_{(p)_j}}}{1 - L_{(p)_j}}, \quad L_{(p)_j} = \frac{\lambda_j \times p}{\mathrm{PR}}. \tag{3}$$

Herein, $\lambda_j$ signifies the network router traffic, obtained from the simulation by keeping track of all paths that crossed router $j$, as well as their duration.

## 4. Calculation of the Performance Metrics

This section presents the analytical calculation of the performance metrics (Table 4), defined in our model to quantify the impact of the number of mobile nodes on the performance of the EPC current and decentralized network architectures, with the GTP and PMIP protocols.

Table 4: Performance metrics.

| Symbol | Definition |
|---|---|
| ALDPD | Average latency of data packet delivery |
| ACDPD | Average processing cost of data packet delivery |
| ALAP | Average latency of initial attachment procedure |
| CAP | Processing cost of initial attachment procedure |
| ALHP | Average latency of handover procedure |
| ACHP | Average processing cost of handover procedure |
| LR | Load of routers |
| LNL | Load of network links |
| $\mathrm{CML}_{(N)}$ | Control messages load (number) |
| $\mathrm{CML}_{(\mathrm{KBs})}$ | Control messages load (size) |

In current EPC architecture, PGW handles centrally the MN's data traffic and mobility through the whole network. However, in decentralized architecture, the S/PGWs, being distributed closer to edge of the network, manage traffic of the MNs attached locally and handle their mobility when moving between the eNodeBs. Therefore, an additional mechanism needs to be implemented on top of the S/PGWs to keep ongoing traffic sessions active for the MNs performing handovers with S/PGW relocation. Different approaches may demand additional components and modifications in the network topology as well as impose further signaling efforts in the network, which must also be taken into account; see [18–22] as examples.

In our model we only study the parameters related to the core network. That is because the structures and characteristics for the access and wireless networks are the same for both the centralized and decentralized LTE architectures.

In the following, Section 4.1 defines the performance metrics listed in Table 4. Next we detail the latency and processing cost related metrics for both architectures in Sections 4.2 and 4.3, respectively.

*4.1. Definition of Performance Metrics*

*(i) The Average Latency of Data Packet Delivery (ALDPD).* The ALDPD is obtained using the average latency of data packet delivery for each mobile node $i$:

$$\mathrm{ALDPD} = \frac{\sum_{i=1}^{N_{\mathrm{node}}} \mathrm{ALDPD}_i}{N_{\mathrm{node}}}. \tag{4}$$

The individual ALDPD for a mobile node $i$ is given by

$$\mathrm{ALDPD}_i = \mathrm{LAP}_i + \overline{\mathrm{LHP}_i} + \overline{\mathrm{LDPD}_i}, \tag{5}$$

where

$$\overline{\mathrm{LHP}_i} = \frac{\sum_{h_i=1}^{N_{h_i}} \mathrm{LHP}_{h_i}}{N_{h_i}},$$
$$\overline{\mathrm{LDPD}_i} = \frac{\sum_{k_i=1}^{N_{k_i}} \mathrm{LDPD}_{k_i}}{N_{k_i}}. \tag{6}$$

Table 5: Input parameters *(assumptions)*.

| Symbol | Definition | Value |
|---|---|---|
| $N_{\text{node}}$ | Number of mobile nodes | $[100 \cdots 1000]$ |
| $E(v)$ | Average velocity of mobile nodes | 25, 50 or 75 km/h |
| $R$ | Radius of the coverage area of the access routers | 1200 m |
| $p$ | Average data packet size | 1200 Bytes |
| CNTR | CNs generating traffic (a Poisson traffic stream) | With average 100 pps |
| PR | Routers processing rate | 1000K pps |
| TR | Network links transmission rate | 100K pps |
| $C_{\text{PGW}}$ | Cost of processing $C_p$, routing $C_r$, or tunneling $C_t$ at PGW | 1 unit per KB |
| $C_{\text{SGW}}$ | Cost of processing $C_p$, routing $C_r$, or tunneling $C_t$ at SGW | 1/3 unit per KB |
| $C_{r_j}$ | Cost of routing at the distribution router $j$ | 1/4 unit per KB |

Note that $\text{LAP}_i$ is the latency of the initial attachment procedure for mobile node $i$. The $\overline{\text{LHP}_i}$ and $\overline{\text{LDPD}_i}$ define the average handover latency and the average latency for data packet delivery over the created paths (due to mobility) for mobile node $i$, respectively.

*(ii) Average Processing Cost of Data Packet Delivery (ACDPD).* The ACDPD is obtained by

$$\text{ACDPD} = \sum_{i=1}^{N_{\text{node}}} \text{ACDPD}_i. \tag{7}$$

The individual ACDPD for a mobile node $i$ is given by

$$\text{ACDPD}_i = \text{CAP}_i + \overline{\text{CHP}_i} + \overline{\text{CDPD}_i}, \tag{8}$$

where

$$\begin{aligned} \overline{\text{CHP}_i} &= \frac{\sum_{h_i=1}^{N_{h_i}} \text{CHP}_{h_i}}{N_{h_i}}, \\ \overline{\text{CDPD}_i} &= \frac{\sum_{k_i=1}^{N_{k_i}} \text{CDPD}_{k_i}}{N_{k_i}}. \end{aligned} \tag{9}$$

Herein, $\text{CAP}_i$ is the processing cost of the initial attachment procedure for the mobile node $i$. The $\overline{\text{CHP}_i}$ and $\overline{\text{CDPD}_i}$ define the average handover processing cost and the average processing cost for data packet delivery over the created paths (due to mobility) for mobile node $i$, respectively.

One may note that $\text{LAP}_i$ and $\text{CAP}_i$ may slightly affect the overall amount of the ALDPD and ACDPD, respectively. However, we would like to discuss all parameters involved during the data packet delivery procedure to acquire more precise results.

*(iii) Average Latency of Initial Attachment Procedure (ALAP).* It is given by

$$\text{ALAP} = \frac{\sum_{i=1}^{N_{\text{node}}} \text{LAP}_i}{N_{\text{node}}}. \tag{10}$$

*(iv) Processing Cost of Initial Attachment Procedure (CAP).* The CAP is obtained by

$$\text{CAP} = \sum_{i=1}^{N_{\text{node}}} \text{CAP}_i. \tag{11}$$

*(v) Average Latency for Handover Procedure (ALHP).* The ALHP is given by

$$\text{ALHP} = \frac{\sum_{i=1}^{N_{\text{node}}} \overline{\text{LHP}_i}}{N_{\text{node}}}. \tag{12}$$

*(vi) Average Processing Cost for Handover Procedure (ACHP).* It is given by

$$\text{ACHP} = \sum_{i=1}^{N_{\text{node}}} \overline{\text{CHP}_i}. \tag{13}$$

*(vii) Load of the Network Routers (LR) and Load of the Network Links (LNL).* The LR and LNL define the data plane loads, implying how many times MNs traffic passes through the network routers and links, respectively.

*(viii) The Control Message Load in Terms of Number ($CML_{(N)}$) and Size of Messages ($CML_{(KBs)}$).* The $\text{CML}_{(N)}$ defines the load of control plane at the network routers in terms of the number of messages. It is obtained by

$$\begin{aligned} \text{CML}_{(N)_j} = \sum_{i=1}^{N_{\text{node}}} \Big( & \text{MNAP}_{(\text{GTP/PMIP})} \\ & + \text{MNHP}_{(\text{GTP/PMIP})} \times N_{h_i} \Big). \end{aligned} \tag{14}$$

We count the $\text{CML}_{(N)}$ for both EPC network architectures using the GTP and PMIP protocols.

The $\text{CML}_{(\text{KBs})}$ is another interpretation of $\text{CML}_{(N)}$ that takes into account also the size of messages and represents a better view of control plane related load at the network routers. Considering that the control messages type and size for the GTP and PMIP protocols are different, making the expression of $\mathrm{CML}_{(\mathrm{KBs})}$ more complex, we avoid to present it here.

In the following, we elaborate on the $\mathrm{ALDPD}_i$ and $\mathrm{ACDPD}_i$, covering also the latency of the initial attachment ($\mathrm{LAP}_i$) and the average handover ($\overline{\mathrm{LHP}_i}$) procedures as well as the related costs ($\mathrm{CAP}_i$ and $\overline{\mathrm{CHP}_i}$), respectively. Abbreviations of the control messages used in the following sections are listed in Tables 1 and 2. Moreover, the notations for the elements used in the analytical calculation of the performance metrics as well as the input parameters are listed in Tables 3, 5, and 6, respectively.

*4.2. ALDPD for Individual Mobile Node.* $\mathrm{ALDPD}_i$ from (5) consists of the latency of the initial attachment procedure ($\mathrm{LAP}_i$), the average latency of handover procedure ($\overline{\mathrm{LHP}_i}$), and the mean latency of data packet delivery ($\overline{\mathrm{LDPD}_i}$) for node $i$, in the paths created from CN to the access layer routers. In the following we detail these items for the centralized and decentralized LTE architectures, considering both GTP and PMIP protocols.

### 4.2.1. The Centralized Architecture

*The GTP-Based Approach.* Referring to Figure 4(a), $\mathrm{LAP}_i$ defines the latency of mobile node initial attachment procedure, caused by exchanging C.S.Req/Res and M.B.Req/Res messages between the attached SGW and PGW. That is the delay of tunneling (constructing/extracting) of the messages in the attached SGW and PGW and also the delay of routing those messages (over the GTP-C) in the path between them.

$\overline{\mathrm{LHP}_i}$ is calculated using the latency of handover in each path ($\mathrm{LHP}_{h_i}$) during the simulation time. For both eNodeB and SGW relocations M.B.Req/Res messages are exchanged between the attached SGW and PGW. Therefore, $\mathrm{LHP}_{h_i}$ includes (i) the delay of tunneling of messages at the attached SGW (for eNodeB relocation) or at the second SGW (for SGW relocation) and PGW and (ii) the delay of routing the tunneled messages in the path built between them.

*The PMIP-Based Approach.* In the PMIP-based approach the initial attachment latency includes (i) the delay for exchanging G.C.S.Req/Res messages (over SCTP protocol) between the attached SGW and PCRF; (ii) the delay for tunneling the P.B.U/P.B.A messages at the attached SGW and PGW; and (iii) the delay of routing the messages (over the GRE) in the created path (Figure 4(b)).

$\mathrm{LHP}_{h_i}$ defines the delay for exchanging G.C.S.Req/Res messages (over the SCTP protocol) between the first SGW (for eNodeB relocation) or the second SGW (for SGW relocation) and PCRF. In case of SGW relocation, it also includes the delay of tunneling of P.B.U/P.B.A messages in the second SGW and PGW and the delay of routing the messages in the path between them.

Finally, $\mathrm{LDPD}_{k_i}$ specifies the delay of data packet delivery from CN to SGW in path $k$. In both GTP and PMIP approaches the $\mathrm{LDPD}_{k_i}$ includes (i) the delay of routing IP packets between CN and PGW; (ii) the delay of data packet encapsulation (over GTP-U or GRE) on PGW; (iii) the delay of routing the tunneled packets in the path between PGW and SGW; and (iv) the delay of data packet decapsulation in SGW.

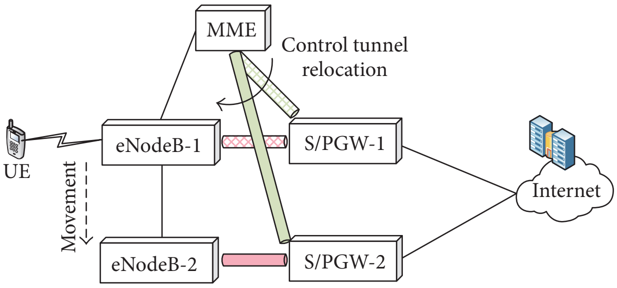


Figure 7: A general view of LTE decentralized architecture.

If the MN experiences a handover with SGW relocation during the session time, data packets have to be tunneled between two SGWs. The delay caused by this procedure must also be taken into account in $\mathrm{LDPD}_{k_i}$.

The detailed derivation of $\mathrm{ALDPD}_i$ for the centralized architecture is given in Appendix A.1.

*4.2.2. The Decentralized Architecture.* In the decentralized architecture, S/PGW performs both SGW and PGW functionalities. The control messages are not tunneled during the initial attachment and handover procedures, and neither the data packets are forwarded between PGW and SGW. The only GTP tunneling is between S/PGW and eNodeBs, which is also the case for the centralized architecture (Figure 7).

*The GTP-Based Approach.* $\mathrm{LAP}_i$ only includes the delay of processing the C.S.Req/Res and M.B.Req/Res messages on S/PGW. Similarly, the delay of $\mathrm{LHP}_h$ is for processing the M.B.Req/Res messages in the first attached S/PGW during eNodeB relocation and in the second one for S/PGW relocation.

*The PMIP-Based Approach.* The delay of $\mathrm{LAP}_i$ is for processing the G.C.S.Req/Res, including the SCTP and IPv6 protocol headers, and P.B.U/P.B.A messages on S/PGW. $\mathrm{LHP}_{h_i}$ for eNodeB relocation only includes the processing delay of G.C.S.Req/Res messages in S/PGW. In the case of S/PGW relocation the delay is due to processing the G.C.S.Req/Res and P.B.U/P.B.A messages in target S/PGW.

In the decentralized architecture, regular IP data packets with no tunneling are forwarded through CNs to S/PGWs. Therefore, $\mathrm{LDPD}_{k_i}$ only includes the delay for routing the packets in the path between an arbitrary CN and S/PGW.

Similar to the centralized architecture in the case of a S/PGW relocation during node session time the delay for forwarding the tunneled data packet between two S/PGWs must be considered.

The detailed derivation of $\mathrm{ALDPD}_i$ for the decentralized architecture is given in Appendix A.2.

*4.3. ACDPD for Individual Mobile Node.* $\mathrm{ACDPD}_i$ from (8) includes the cost of handling mobility control messages ($\mathrm{CAP}_i$ and $\mathrm{CHP}_{h_i}$) and the cost of data packet delivery

($\mathrm{CDPD}_{k_i}$) from CN to the access layer nodes in the created paths, during the MN's session time. This section describes the parameters in both centralized and decentralized approaches and details these parameters for GTP and PMIP protocols. For the sake of simplicity, we assume that the costs of the processing ($C_p$), routing ($C_r$), and tunneling ($C_t$) in the routers are the same. Furthermore, given the traffic load crossing through the routers at each layer, we assign 1, 1/3, 1/4 unit processing cost per each KB of traffic for the root router (PGW *in centralized architecture*), the distribution routers, and the access routers (SGWs or S/PGW), respectively. This is a rational comparative assignment for the purpose of comparing network architectures.

*4.3.1. The Centralized Architecture*

*The GTP-Based Approach.* In the GTP-based approach, $\mathrm{CAP}_i$ defines the cost of exchanging C.S.Req/Res and M.B.Req/Res tunneled messages between the first attached SGW and PGW. It also includes the cost of constructing/extracting of messages in the first SGW and PGW and the cost of routing the tunneled messages (over GTP-C) in the path between them.

Similarly, $\mathrm{CHP}_{h_i}$ describes the cost of transferring only the M.B.Req/Res tunneled messages (over GTP-C) between the first attached SGW and PGW during eNodeB relocation or the target SGW and PGW for SGW relocation.

*The PMIP-Based Approach.* For the PMIP protocol, $\mathrm{CAP}_i$ comprises (i) the costs of swapping the G.C.S.Req/Res messages, including the SCTP and IPv6 protocol headers, between the first attached SGW and PCRF; (ii) the cost of tunneling (constructing/extracting) the P.B.U/P.B.A messages in the first attached SGW and PGW; and (iii) the cost of routing the tunneled messages over GRE in the path between them.

$\mathrm{CHP}_{h_i}$ for eNodeB relocation only includes the cost of processing G.C.S.Req/Res messages exchanged between the first SGW and PCRF. During SGW relocation, $\mathrm{CHP}_{h_i}$ defines the cost of exchanging aforesaid messages between the target SGW and PCRF. Moreover, it includes the cost of exchanging the tunneled P.B.U/P.B.A messages over GRE between the second SGW and PGW.

$\mathrm{CDPD}_{k_i}$ specifies the cost of routing regular IP packets from CN to PGW and the cost of tunneling (encapsulating) data packets over GTP-U or GRE in PGW. It also counts for the cost of routing the tunneled packet in the path between PGW and SGW and also the cost of decapsulating the tunneled packet in SGW.

The cost of forwarding the tunneled data packets between two SGWs must also be considered in case of a SGW reallocation during the session time.

The detailed derivation of $\mathrm{ACDPD}_i$ for the centralized architecture is given in Appendix B.1.

*4.3.2. The Decentralized Architecture*

*The GTP-Based Approach.* As described in Section 4.2.2, the decentralized architecture does not have a control or data plane tunneling in the core network. Therefore, $\mathrm{CAP}_i$ only includes the cost of processing C.S.Req/Res and M.B.Req/Res messages in the S/PGW.

$\mathrm{CHP}_{h_i}$ includes the cost of processing M.B.Req/Res messages in the first attached S/PGW for eNodeB relocation and in the target S/PGW during S/PGW relocation.

*The PMIP-Based Approach.* In this approach, $\mathrm{CAP}_i$ is related to the cost of processing G.C.S.Req/Res messages, including the SCTP and IPv6 protocol headers, and the P.B.U/P.B.A messages in S/PGW.

$\mathrm{CHP}_{h_i}$ for eNodeB relocation only defines the cost of processing the G.C.S.Req/Res messages, including the SCTP and IPv6 protocol headers, in the first S/PGW. During S/PGW relocation it describes the cost of processing the G.C.S.Req/Res and P.B.U/P.B.A messages in the second S/PGW.

In decentralized architecture, $\mathrm{CDPD}_{k_i}$ specifies the cost of routing IP packets with no tunneling header in the path between CN and S/PGW. If a MN experiences handover during the session time, $\mathrm{CDPD}_{k_i}$ also includes the cost of forwarding the tunneled data packets between the first and target S/GWs.

Note that, in both network architectures, if there is more than one SGW (or S/PGW) between CN and mobile node the data packets are tunneled and forwarded among them. Therefore, the additional delay and cost due to this procedure also must be considered in calculating $\mathrm{LDPD}_{k_i}$ and $\mathrm{CDPD}_{k_i}$.

The detailed derivation of $\mathrm{ACDPD}_i$ for the decentralized architecture is given in Appendix B.2.

## 5. Numerical Results

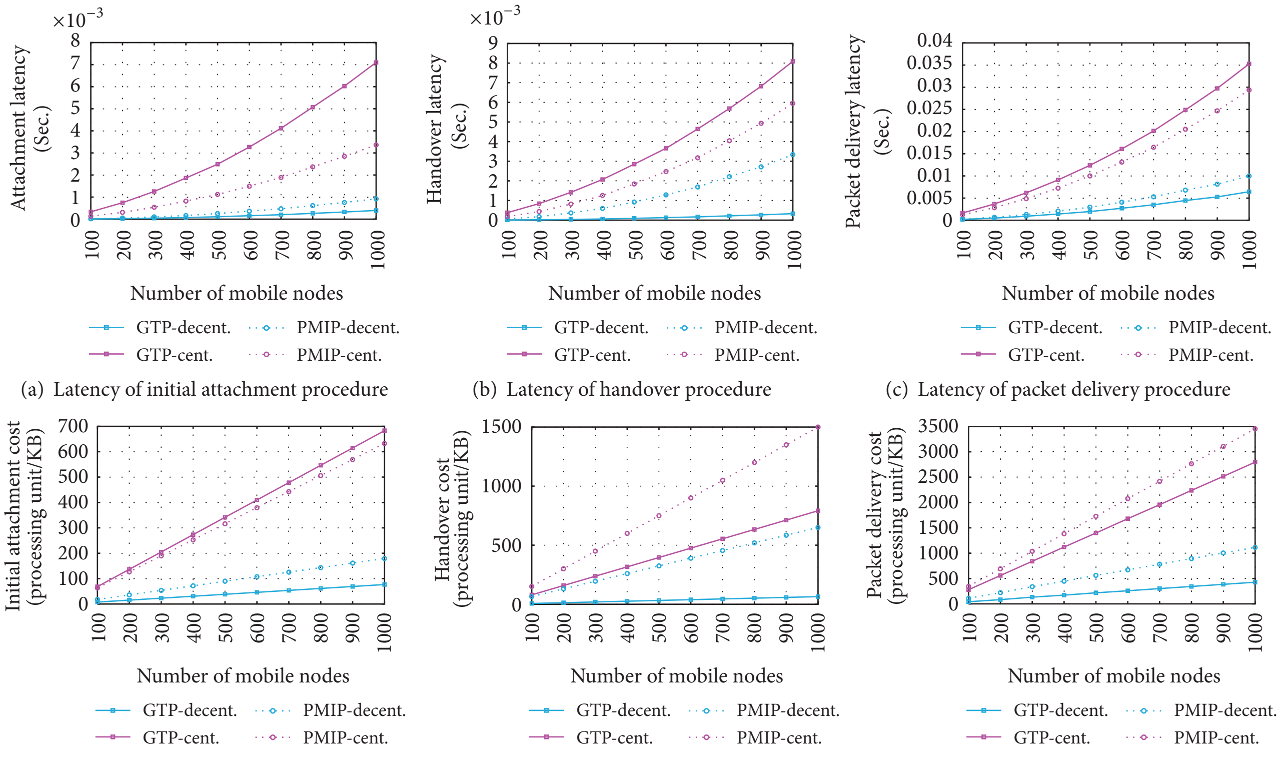
This section presents the numerical results of the performance metrics, defined in Section 4. The obtained results provide scalability indicators for the EPC centralized and decentralized network architectures via a quantitative analogy over the performance of GTP and PMIP protocols. Table 5 lists the input parameters and Table 6 summarizes the default parameters used in Sections 3 and 4.

*5.1. Average Cost and Latency of MN's Initial Attachment, Handover, and Data Packet Delivery Procedures.* The graphs in Figure 8 show the impact of the number of MNs on GTP (solid lines) and PMIP (dash lines) performance in terms of cost and latency, for both the EPC centralized (red lines) and decentralized (blue lines) architectures. It is notable that the decentralized architecture outperforms the centralized one, regardless of the mobility protocol. This is because in the decentralized architecture the control plane messages without tunneling are handled in the S/PGWs of access layer. Furthermore, data traffic with regular IP packets are only transmitted over the paths between CNs and S/PGWs without crossing the root node. Accordingly, latency and cost measures are substantially improved.

*The MN's Initial Attachment Procedure.* Figures 8(a) and 8(d) show that, in centralized architecture during the initial attachment procedure, PMIP achieves better outcomes than GTP particularly in latency. This is due to the fact that four

Table 6: Standard parameters *(Sections 2.2 and 2.3)*.

| Symbol | Definition | Size |
|---|---|---|
| $t_{u(\mathrm{GTP})}$ | *GTP-U* header | 64 B |
| $t_{c(\mathrm{GTP})}$ | *GTP-C* header | 60 B |
| If $m$ = C.S.Req | GTP *Create Session Request* message | 335 B |
| If $m$ = C.S.Res | GTP *Create Session Response* message | 224 B |
| If $m$ = M.B.Req | GTP *Modify Bearer Request* message | 67 B |
| If $m$ = M.B.Res | GTP *Modify Bearer Response* message | 81 B |
| $t_{(\mathrm{PMIPv6})}$ | PMIPv6 tunnel header | 64 B |
| If $m$ = G.C.S.Req | PMIPv6 *Gateway Control Session Request* message | 336 B |
| If $m$ = G.C.S.Res | PMIPv6 *Gateway Control Session Response* message | 972 B |
| If $m$ = P.B.U | PMIPv6 *Proxy Binding Update* message | 104 B |
| If $m$ = P.B.A | PMIPv6 *Proxy Binding Acknowledge* message | 104 B |



(a) Latency of initial attachment procedure

(b) Latency of handover procedure

(c) Latency of packet delivery procedure

(d) Cost of initial attachment procedure

(e) Cost of handover procedure

(f) Cost of packet delivery procedure

Figure 8: The latency and cost of GTP and PMIP mobility protocols in EPC centralized and decentralized architectures.

messages in GTP and two messages in PMIP are exchanged (over GTP-C and GRE, respectively) between the SGWs and PGW. The other two messages in PMIP are exchanged between the SGW and PCRF, through a private network link with no tunneling (Figure 4). In decentralized architecture, the S/PGWs handle these messages. Hence, the related latency and cost are only due to processing of messages on the S/PGWs, having no tunneling header. In this scenario, GTP provides better results as its initial attachment messages are smaller than PMIP messages.

For the centralized architecture, Figure 8(a) shows that the growth rate of the attachment latency in PMIP is 5% less than GTP. However, in the decentralized architecture GTP performs 4% better than PMIP. In addition, using decentralized architecture, the attachment latency is improved ≈44 and ≈8 times in GTP and PMIP, respectively (Table 7).

Figure 8(d) shows that for attachment cost in centralized architecture PMIP provides (>1%) lower increasing slope compared to GTP. However, in the decentralized architecture this metric for GTP is ≈7% lower than for PMIP.

Table 7: Performance of GTP and PMIP in the various EPC network architectures.

| | Centralized | | Decentralized | | Ratio*** | |
|---|---|---|---|---|---|---|
| | GTP | PMIP | GTP | PMIP | GTP | PMIP |
| LAP* | 55% | 50% | 40% | 44% | 44.35 | 7.63 |
| LHP* | 55% | 53% | 40% | 47% | 61.83 | 2.94 |
| LDPD* | 52% | 50% | 41% | 42% | 11.54 | 5.83 |
| CAP** | 155.8% | 154.7% | 144.1% | 151.5% | 8.90 | 3.52 |
| CHP** | 153.8% | 156.4% | 141.6% | 154.3% | 12.31 | 2.30 |
| CDPD** | 156.7% | 157.2% | 147.3% | 156.1% | 6.74 | 3.13 |

*The percentage shows the *growth rate* (GR) of the exponential graphs. **The percentage shows the *increasing slope* (IS) of the liner graphs. The exponential and linear graphs are expressed by $Y_n = Y_{n_0} \times (1 + \mathrm{GR})^n$ and $Y_n = \mathrm{IS} \times n + Y_{n_0}$, respectively. $Y$ and $n$ represent the performance metrics and the number of MNs in $y$- and $x$-axes, respectively. ****Ratio* (R) shows the proportion of the metrics in centralized to decentralized architecture, derived for $Y_{n_0}$ $(n_0 = 100)$.

Furthermore, in decentralized architecture GTP and PMIP decrease the attachment cost by ≈9 and ≈3 times, respectively, compared to the centralized architecture.

*The Handover Procedure.* Figure 8(b) shows that in centralized architecture PMIP offers a lower handover latency than GTP. This is caused by the latency of routing two tunneled messages between the attached SGW and PGW in GTP for every eNodeB relocation. For PMIP the messages are only processed and exchanged between the SGW and PCRP, via a dedicated link. For the cost metric, GTP outperforms PMIP (Figure 8(e)). Because in GTP both the control plane tunnel headers and handover messages are smaller than PMIP. Furthermore, GTP uses fewer number of messages during SGW relocation. In the decentralized architecture, GTP shows better functionality than PMIP both in terms of latency and cost (Figures 8(b) and 8(e)). This is due to the processing of fewer short-sized handover messages in GTP on S/PGWs. Figure 8(b) indicates that for centralized architecture PMIP shows 2% lower growth rate on handover latency than GTP. In decentralized architecture, GTP outperforms PMIP and provides 7% lower growth rate for this metric. Moreover, the handover latency is also improved by ≈62 and ≈3 times in GTP and PMIP, respectively. In terms of handover cost, for both architectures GTP carries out better than PMIP, achieving 2.6% and 12.7% lower increasing slope for this metric in the centralized and decentralized architectures, respectively (Table 7). Furthermore, for decentralized architecture, GTP reduces the handover cost by ≈12 times compared with centralized one, which is ≈2 times when PMIP is used (Figure 8(e)).

*The Data Packet Delivery Procedure.* As shown in Figure 3(a), sizes of the tunnels header in both GTP and PMIP data planes are the same, and hence, one may expect similar latency and cost for both protocols. However, recalling from (4) and (7), these metrics also depend on the latency and cost on the attachment and handover procedures, resulting in the similar outcomes.

Figure 8(c) shows that PMIP performs 2% better than GTP for centralized architecture in terms of growth rate for latency on data packet delivery. However, in decentralized architecture GTP outperforms PMIP and provides 1% lower growth rate for this metric. Furthermore, GTP improves latency for ≈11% times, which is ≈6% far from PMIP.

For the cost metric, GTP provides better results than PMIP with ≈1% and ≈9% lower increasing slope in the centralized and decentralized architectures. In addition, for decentralized architecture GTP reduces the data packet delivery cost by ≈6 times compared to centralized one, while PMIP achieves a ratio of ≈3 times (Figure 8(f)).

Table 7 summarizes the results discussed in this section.

### 5.2. The Data Plane and Control Plane Load on the Network Routers and Links.

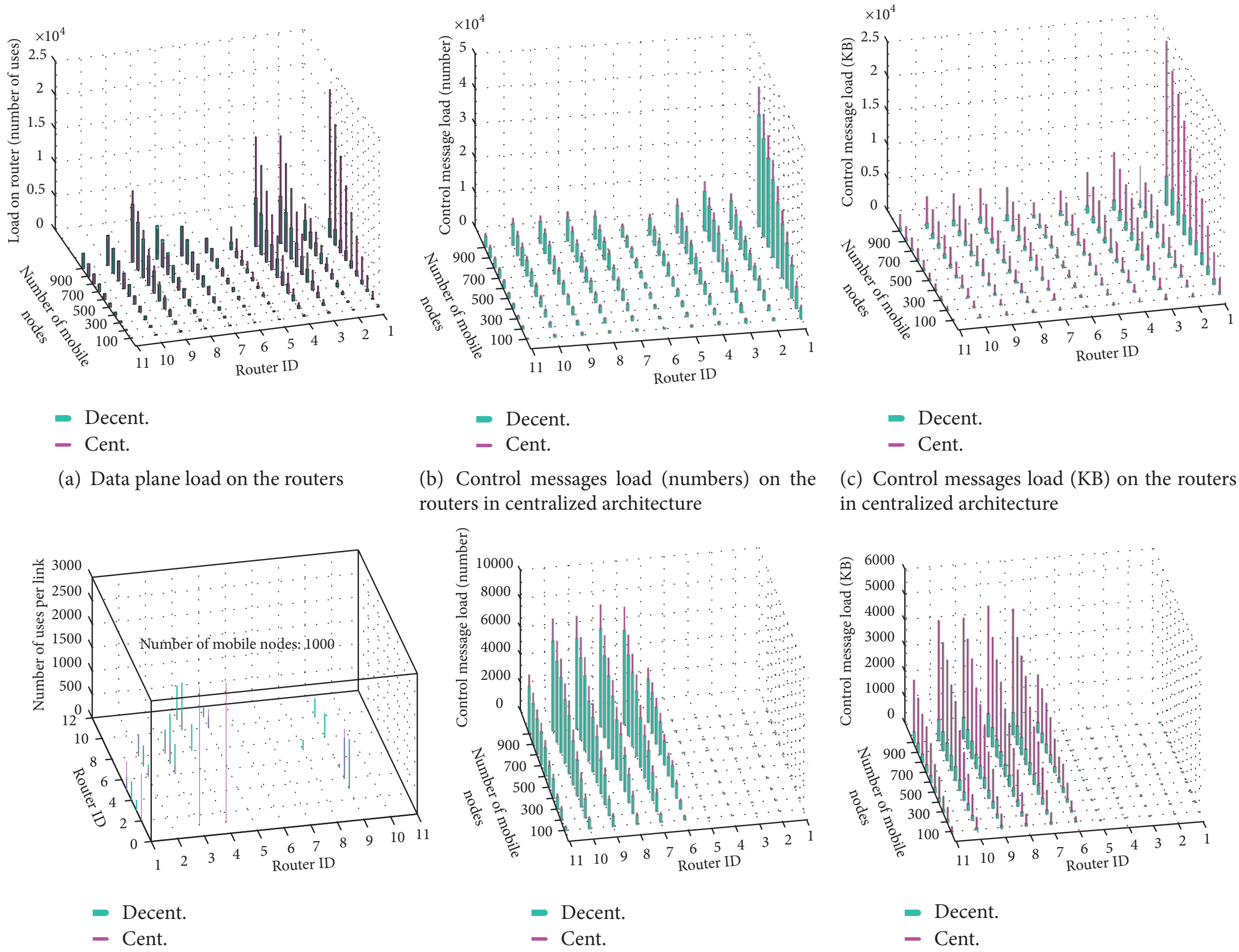
Figure 9 shows the impact of the number of MNs on the network routers and links loads for centralized and decentralized architectures. Although we expected a higher load of both data and control planes in decentralized architecture than in centralized one, we observe that the differences are surprisingly large. That is because in decentralized architecture the control plane messages as well as the data traffic are managed and handled at anchor points placed in the access layer, resulting in reduced load and stress in the upper layers.

*Load of the Data Plane.* Figures 9(a) and 9(d) show the load of the data plane on the network routers and links for different EPC network architectures, respectively. As expected, the root router (R1-PGW, Figure 6) and the routers on the distribution layer (R2 to R5) in the centralized approach are used more often than in the decentralized one. Accordingly, the network links between these two layers, forwarding data traffic between the routers, are also used more in centralized approach.

In decentralized architecture MNs traffic is mainly handled by the access layer routers (R6 to R11-S/PGW) and load is more distributed over the network. This reduces the stress on upper layers and diminishes the load of the core routers and links by growing the number of attached MNs.

Note that, in both architectures, some of the network links are not used. That is because in the proposed model the path construction between the MNs and CNs is only based on the shortest-paths approach, disregarding other functionalities such as load balancing.

Note also that, in decentralized architecture, the load on router R9 is slightly higher than in centralized architecture.



(a) Data plane load on the routers

(b) Control messages load (numbers) on the routers in centralized architecture

(c) Control messages load (KB) on the routers in centralized architecture

(d) Data plane load on the network links

(e) Control messages load (numbers) on the routers in decentralized architecture

(f) Control messages load (KB) on the routers in decentralized architecture

Figure 9: The data plane and control plane loads, over the network routers and links in EPC centralized and decentralized architectures.

This is because in the former R9 is directly connected to CN3 and directly serves the MNs that link to CN3.

In addition in centralized architecture, the obtained results show that the load of the root router (R1) is ≈11 times more than in decentralized architecture. For the routers placed in the distribution and access layers this load ratio on average is ≈2 and ≈1 times, respectively.

*Load of the Control Plane.* Figures 9(b) and 9(e) show the impact of the number of MNs on loads of GTP and PMIP control messages within the routers for different EPC architectures. In centralized architecture all messages related to MN's mobility are handled by the root router (PGW). The routers at the distribution layer (R2 to R5) have also to be involved on crossing the messages within the core network. However, in decentralized architecture control messages are not traversed to the upper layer routers and the access layer routers (S/PGWs in our model) are in charge of managing the mobility messages.

Figures 9(c) and 9(f) show loads on the network routers due to the control plane messages in terms of amount (KBs) of the handling messages. It is observed that PMIP protocol inflicts more load to the network than GTP protocol. This is because size of the mobility messages and the tunneling headers for PMIP are larger than for GTP. Furthermore, GTP generates less number of control messages during the SGW or S/PGW relocation (Figures 3(a) and 4).

Our results indicate that, for both GTP and PMIP protocols, the control plane load on the routers is amplified linearly by the number of attached nodes to the network. It is also inferred that PMIP imposes in average ≈6 times higher load on the routers than GTP.

## 6. Related Work

Using a simulation model, a comparative study has been performed in [23], to analyze performances of the proposed Dynamic Mobility Anchoring (DMA) scheme and Mobile IP (MIP) protocol to handle the MNs' TCP-based traffic. In this study, the handover latency and TCP segment delays have been used as the performance metrics. The work in [16] presents an analytical and experimental evaluation of PMIP-based mobility management in centralized and distributed ways. The metrics of signaling and packet delivery costs, handover latency, and packet loss have been used to present the trade-offs between two architectures. In a similar way, [24] investigated impact of the distribution and dynamic activation of the mobility anchors on performance of PMIP protocol. Here, the evaluation has been accomplished based on the packet delivery cost, anchored/nonanchored packet ratio, and traffic distribution ratio. The authors in [25] proposed a PMIPv6-based distributed mobility management model (D-PMIPv6), which outperformed the conventional PMIPv6 in terms of route optimization, and packet delivery and signaling costs. The discussed approach was based on distribution of access routers with a centralized management model. A comprehensive study of distributed and dynamic mobility management (DDMM) has been done by [26]. The authors discussed an architecture with distributed deployment of mobility anchors and dynamic activation. In this research, it has been shown that DDMM generally achieves higher performance compared to centralized mobility management in terms of packet delivery cost, tunneling overhead, and throughput. The work in [15] proposed a partially (P-DMM) and a fully distributed mobility management (F-DMM). In the former only the data plane was distributed, while in the latter both data and control planes were distributed. In this work, it was shown that the F-DMM outperforms P-DMM strategy in terms of handover latency and packet loss. In [9], the functional differences of the GTP and PMIP, within the EPC, have been discussed and the signaling costs of these protocols, using dynamic QoS and policy control, have been evaluated. An introduction of the different LTE core network architectures and mobility management schemes has been presented in [27]. In this work, an analytical model based on the two-dimensional hexagonal random walk model was proposed for comparing performance of the mobility management on different EPC core architectures. For the performance evaluations the total signaling cost and load on the network nodes have been used as the comparison parameters.

Our work differs from the literature, particularly with respect to comprehensive analysis of four main possible scenarios on LTE system, that is, the current (centralized) and future (decentralized) EPC architectures with GTP or PMIP mobility protocols for 3GPP access. Compared to previous studies, mostly focused on the performance analysis of MIP/PMIP protocols in general centralized and distributed approaches, we paid special attention to the LTE network, the dominant mobile networking technology to accommodate the major part of worldwide mobile data traffic in coming decade.

We conducted a hybrid modeling study, detailing the GTP and PMIP protocols (the only used mobility support mechanisms in the EPC) data and control planes functions for 3GPP access, to quantify the performance and scalability of the current and future LTE network architectures. The proposed model enables detailed analysis on the load of network entities and on the network efficiency parameters for the user's data and control planes, by increasing the number of mobile nodes. To the best of our knowledge, this is the first work to perform a quantitative analysis of such scale on the LTE system in order to compare the performance of different network architectures.

## 7. Conclusions

Although decentralization of the core network architecture is not standardized by the 3GPP yet, it is seen as the vision on emerging future mobile network (e.g., 5G) architectural standards. In this regard there are different ongoing research projects and activities, aiming to come up with the solutions to address the features and demands of current mobile networks in a decentralized architecture. In this article, we have particularly conducted a detailed analysis and comparative study of centralized and decentralized network architectures in the LTE system for 3GPP access. We have carried out a hybrid study comprising simulation and analytical modeling to evaluate the attached node's (device's) data traffic and mobility related messages load, as well as the latency and cost for the initial attachment, handover, and data packet delivery procedures in both network architectures, using GTP and PMIP protocols. Our research aimed, in particular, to quantify the impact of the number of connected devices to the network, on various performance and scalability metrics for both LTE network architectures.

Given the specified scenarios and parameters in our study, the optioned results show that decentralization of the LTE network architecture substantially reduces load of data traffic (≈11 times) on the core of the network. Accordingly, this leads to improvement of the latency for attachment (≈44 times) and handover (≈61 times) procedures during the node mobility and the latency for data delivery procedure (≈11 times) *(using GTP)*, which are the keys in providing higher QoS and QoE for the subscribers. It is also shown that a decentralized architecture *(using GTP)* imposes remarkably lower processing cost on both the data plane (≈7 times) and control plane (≈11 times *during handover*) (see Table 7), which is also an essential concern for network operators. The analysis indicated that, in the centralized architecture, PMIP achieves slightly lower growing rates for the latencies but provides higher increasing slopes for the loads and costs compared to GTP protocol. However, in a decentralized architecture, GTP protocol outperforms PMIP for all the performance metrics.

The presented approach of analysis helps to assess the network efficiency (in terms of the data and control planes latency and processing cost) and the network scalability (in terms of handling the data traffic load and signaling overhead) in the LTE network with different core network architectures. The outcomes of this research provide clear intuition on the impact of the growing number of users on the current and future LTE systems. In a future study, our analysis can be further extended by taking into account other parameters such as the additional control plane overhead, demanded to address the core network features (e.g., mobility management, policy controlling, and accounting), and the required investment and level of complexity for modification and maintenance of the network architectures. This can provide the mobile network operators a trustworthy insight during decision making and policy development procedures in order to accommodate the network infrastructure for coping with the future demands on mobile data traffic.

## 8. Discussion

This section briefly discusses the major modifications that would be required on the current EPS system as well as some of technical challenges to realize a decentralized LTE network architecture and to support MN's mobility in this architecture.

*8.1. Anchor Point Relocation.* In the current 3GPP LTE specification, IP-based traffic continuity is not supported when a MN changes its EPS traffic anchor point (PGW), for example, during interoperator roaming procedure. In the existing LTE system, MN's traffic remains anchored to a single PGW until it moves out of the access network, and upon a handover to a different anchor point, flows initiated at the previous PGW will be stopped. This is due to the fact that, in the current LTE technical specifications, there is not a standard mechanism to support a handover procedure with PGW relocation and to provide the continuity of bearers after PGW relocation. Following a decentralized LTE architecture, the anchor points (S/PGW) are placed closer to the edge network to locally handle the MNs' traffic and mobility, and the MNs are expected to change their anchor point far more often. In this case, two layers of mobility management are needed in order to handle the MNs' IP traffic continuity during a handover with a S/PGW relocation: (i) within the EPC network (between the S/PGWs and eNodeBs) and (ii) above the EPC network (between the S/PGWs and CNs). The main obstacle in implementation of IP address continuity within the EPC network comes by the fact that, in the current EPS architecture, there is neither signaling nor data forwarding scheme available between two different PGW entities. However, by combining the PGW and SGW functionalities into a single entity (S/PGW), current standard solution and messages used for a handover procedure with SGW relocation can be revised and modified to support IP traffic continuity between different S/PGW domains. We have addressed this issue in our previous works and detailed information about the related modifications can be found in [19, 20].

The mobility above the anchor points (S/PGWs) is discussed in the following section.

*8.2. Traffic Steering on Top of the Anchor Points.* As described previously, in the current EPC architecture, the PGW as a central anchor point handles all the MNs data traffic and mobility related functions. However in a LTE with decentralized architecture, the S/PGWs are distributed closer to the edge of the network, anchoring the MNs attached locally and handling mobility for those users moving between eNodeBs. In this case, an additional mobility support mechanism also needs to be implemented to keep ongoing sessions active (above the EPC) for the MNs performing handover with mobility anchor (S/PGW) relocation. To address this issue, we have developed two *network layer* [19, 20] and one *transport layer* [21] solutions. Generally, the mobility support approaches running to the network layer handle MN's mobility requirements in a transparent manner and hide any changes from upper layers. However, they require some infrastructural modifications and impose extra overhead. Transport layer mobility management schemes keep the network infrastructure intact and implement the whole functionality for supporting MN's mobility in the transport layer of the end host entities. Different approaches may impose further signaling efforts in the network, which must also be taken into account on calculation of the performance metrics.

*8.3. Unifying Anchor Points Data.* Besides the traffic forwarding and mobility anchoring functions, PGW is also centrally in charge of other tasks such as the policy enforcement, packet filtering, packet screening, lawful interception, and charging for each MN. Moving towards a decentralized EPC architecture leads to the distribution of several S/PGWs on the edge of the access network. This demands common templates of the aforementioned functions for S/PGWs to carry out the unique regulations for the MNs performing handover among them. To do so, additional network connections and synchronization mechanisms (or, e.g., resource pooling and memory sharing) need to be applied between the S/PGWs and also with the other EPC components (e.g., MME and PCRF).

## Appendix

## A. Average Latency of Data Packet Delivery for Mobile Node *i*

$\mathrm{ALDPD}_i$ is derived using (A.1).

$$\mathrm{ALDPD}_i = \mathrm{LAP}_i + \frac{\sum_{h_i=1}^{N_{h_i}} \mathrm{LHP}_{h_i}}{N_{h_i}} + \frac{\sum_{k_i=1}^{N_{k_i}} \mathrm{LDPD}_{k_i}}{N_{k_i}}. \tag{A.1}$$

In the following, we calculate $\mathrm{ALDPD}_i$ items for EPC centralized and decentralized architectures with GTP and PMIP protocols.

*A.1. The Centralized Architecture*

(▶)

$$\begin{aligned}
\mathrm{LAP}_{i_{(\mathrm{GTP})}} = {} & \Big\{\Big[\tau_{t_{(\mathrm{C.S.Req}+t_c)}} + \tau_{t_{(\mathrm{C.S.Res}+t_c)}} + \tau_{t_{(\mathrm{M.B.Req}+t_c)}} \\
& + \tau_{t_{(\mathrm{M.B.Res}+t_c)}}\Big] \text{ in SGW}\Big\} \\
& + \Bigg\{\sum_{l=1,j=1}^{H_{\mathrm{SGW}\to\mathrm{PGW}}} \Big[(\delta + d)_{(\mathrm{C.S.Req}+t_c)_l} \\
& + (\delta + d)_{(\mathrm{M.B.Req}+t_c)_l} + \tau_{r_{(\mathrm{C.S.Req}+t_c)_j}} + \tau_{r_{(\mathrm{M.B.Req}+t_c)_j}}\Big]\Bigg\} \\
& + \Bigg\{\sum_{l=1,j=1}^{H_{\mathrm{PGW}\to\mathrm{SGW}}} \Big[(\delta + d)_{(\mathrm{C.S.Res}+t_c)_l}
\end{aligned}$$

$$\begin{aligned}
&+ (\delta + d)_{(\text{M.B.Res}+t_c)_l} + \tau_{r_{(\text{C.S.Res}+t_c)_j}} + \tau_{r_{(\text{M.B.Res}+t_c)_j}} \Big]\Bigg\} \\
&+ \Big\{\Big[\tau_{t_{(\text{C.S.Req}+t_c)}} + \tau_{t_{(\text{C.S.Res}+t_c)}} + \tau_{t_{(\text{M.B.Req}+t_c)}} \\
&+ \tau_{t_{(\text{M.B.Res}+t_c)}}\Big] \text{ in PGW}\Big\}.
\end{aligned} \tag{A.2}$$

(▷)

$$\begin{aligned}
&\text{LAP}_{i_{(\text{PMIP})}} \\
&= \Big\{\Big[\tau_{(\text{G.C.S.Req+SCTP+IPv6})} + \tau_{(\text{G.C.S.Res+SCTP+IPv6})}\Big] \\
&\text{in SGW}\Big\} + \Big\{\Big[\tau_{t_{(\text{P.B.U+GRE})}} + \tau_{t_{(\text{P.B.A+GRE})}}\Big] \text{ in SGW}\Big\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{SGW}\rightarrow\text{PGW}}} \Big[(\delta + d)_{(\text{P.B.U+GRE})_l} + \tau_{r_{(\text{P.B.U+GRE})_j}}\Big]\right\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{PGW}\rightarrow\text{SGW}}} \Big[(\delta + d)_{(\text{P.B.A+GRE})_l} + \tau_{r_{(\text{P.B.A+GRE})_j}}\Big]\right\} \\
&+ \Big\{\Big[\tau_{t_{(\text{P.B.U+GRE})}} + \tau_{t_{(\text{P.B.A+GRE})}}\Big] \text{ in PGW}\Big\}.
\end{aligned} \tag{A.3}$$

(▶)

$$\begin{aligned}
&\text{LHP}^{\dagger}_{h_{i(\text{GTP})}} \\
&= \Big\{\Big[\tau_{t_{(\text{M.B.Req}+t_c)}} + \tau_{t_{(\text{M.B.Res}+t_c)}}\Big] \text{ in SGW}\Big\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{SGW}\rightarrow\text{PGW}}} \Big[(\delta + d)_{(\text{M.B.Req}+t_c)_l} + \tau_{r_{(\text{M.B.Req}+t_c)_j}}\Big]\right\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{PGW}\rightarrow\text{SGW}}} \Big[(\delta + d)_{(\text{M.B.Res}+t_c)_l} + \tau_{r_{(\text{M.B.Res}+t_c)_j}}\Big]\right\} \\
&+ \Big\{\Big[\tau_{t_{(\text{M.B.Req}+t_c)}} + \tau_{t_{(\text{M.B.Res}+t_c)}}\Big] \text{ in PGW}\Big\}.
\end{aligned} \tag{A.4}$$

(†) During each SGW handover, the $\text{LHP}_{h_{i(\text{GTP})}}$ is repeated three times for eNodeB relocation and one time for SGW relocation by replacing SGW → SGW′.

(▷)

$$\begin{aligned}
&\text{LHP}_{h_{i(\text{PMIP})}} = 3 \\
&\times \Big\{\Big[\tau_{(\text{G.C.S.Req+SCTP+IPv6})} + \tau_{(\text{G.C.S.Res+SCTP+IPv6})}\Big] \\
&\text{in SGW}\Big\} \\
&+ \Big\{\Big[\tau_{(\text{G.C.S.Req+SCTP+IPv6})} + \tau_{(\text{G.C.S.Res+SCTP+IPv6})}\Big] \\
&\text{in SGW}'\Big\} + \Big\{\Big[\tau_{t_{(\text{P.B.U+GRE})}} + \tau_{t_{(\text{P.B.A+GRE})}}\Big] \text{ in SGW}'\Big\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{SGW}'\rightarrow\text{PGW}}} \Big[(\delta + d)_{(\text{P.B.U+GRE})_l} + \tau_{r_{(\text{P.B.U+GRE})_j}}\Big]\right\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{PGW}\rightarrow\text{SGW}'}} \Big[(\delta + d)_{(\text{P.B.A+GRE})_l} + \tau_{r_{(\text{P.B.A+GRE})_j}}\Big]\right\} \\
&+ \Big\{\Big[\tau_{t_{(\text{P.B.U+GRE})}} + \tau_{t_{(\text{P.B.A+GRE})}}\Big] \text{ in PGW}\Big\}.
\end{aligned} \tag{A.5}$$

(▶)

$$\begin{aligned}
\text{LDPD}_{k_{i(\text{GTP})}} &= \left\{\sum_{l=1,j=1}^{H_{\text{CN}\rightarrow\text{PGW}}} \Big[(\delta + d)_{(p)_l} + \tau_{r_{(p)_j}}\Big]\right\} \\
&+ \Big\{\tau_{t_{(p+t_u)_{\text{PGW}}}}\Big\} + \Big\{\tau_{t_{(p+t_u)_{\text{SGW}}}}\Big\} \\
&+ \left\{\sum_{l=1,j=1}^{H_{\text{PGW}\rightarrow\text{SGW}}} \Big[(\delta + d)_{(p+t_u)_l} + \tau_{r_{(p+t_u)_j}}\Big]\right\} \\
&+ \Big\{\tau_{t_{(p+t_u)_{\text{SGW}}}} + (\delta + d)_{(p+t_u)_l} + \tau_{r_{(p+t_u)_j}} \\
&+ \tau_{t_{(p+t_u)_{\text{SGW}'}}}\Big\}^{\dagger\dagger}.
\end{aligned} \tag{A.6}$$

(††) This expression equals zero if there is no handover during session time.

(▷) $\text{LDPD}_{k_{i(\text{PMIP})}}$ is calculated the same as $\text{LDPD}_{k_{i(\text{GTP})}}$ by replacing $t_u$ → GRE.

### *A.2. The Decentralized Architecture*

(▶)

$$\begin{aligned}
&\text{LAP}_{i_{(\text{GTP})}} \\
&= \Big\{\Big[\tau_{p_{(\text{C.S.Req})}} + \tau_{p_{(\text{C.S.Res})}} + \tau_{p_{(\text{M.B.Req})}} + \tau_{p_{(\text{M.B.Res})}}\Big] \\
&\text{in S/PGW}\Big\}.
\end{aligned} \tag{A.7}$$

(▷)

$$\begin{aligned}
\text{LAP}_{i_{(\text{PMIP})}} &= \Big\{\Big[\tau_{p_{(\text{G.C.S.Req+SCTP+IPv6})}} + \tau_{p_{(\text{G.C.S.Res+SCTP+IPv6})}} \\
&+ \tau_{p_{(\text{P.B.U})}} + \tau_{p_{(\text{P.B.A})}}\Big] \text{ in S/PGW}\Big\}.
\end{aligned} \tag{A.8}$$

(▶)

$$\begin{aligned}
\text{LHP}_{h_{i(\text{GTP})}} &= 3 \times \Big\{\Big[\tau_{p_{(\text{M.B.Req})}} + \tau_{p_{(\text{M.B.Res})}}\Big] \text{ in S/PGW}\Big\} \\
&+ \Big\{\Big[\tau_{p_{(\text{M.B.Req})}} + \tau_{p_{(\text{M.B.Res})}}\Big] \text{ in S/PGW}'\Big\}.
\end{aligned} \tag{A.9}$$

(▷)

$$\begin{aligned}\mathrm{LHP}_{h_{i(\text{PMIP})}} &= 3 \times \left\{\left[\tau_{p_{(\text{G.C.S.Req+SCTP+IPv6})}} + \tau_{p_{(\text{G.C.S.Res+SCTP+IPv6})}}\right] \text{ in S/PGW}\right\} \\ &+ \left\{\left[\tau_{p_{(\text{G.C.S.Req+SCTP+IPv6})}} + \tau_{p_{(\text{G.C.S.Res+SCTP+IPv6})}} + \tau_{p_{(\text{P.B.U})}} + \tau_{p_{(\text{P.B.A})}}\right] \text{ in S/PGW}'\right\}.\end{aligned} \tag{A.10}$$

(▶)

$$\begin{aligned}\mathrm{LDPD}_{k_{i(\text{GTP})}} &= \left\{\sum_{l=1,j=1}^{H_{\text{CN}\rightarrow\text{S/PGW}}} \left[(\delta + d)_{(p)_l} + \tau_{r_{(p)_j}}\right]\right\} \\ &+ \left\{\tau_{t_{(p+t_u)_{\text{S/PGW}}}} + (\delta + d)_{(p+t_u)_l} + \tau_{r_{(p+t_u)_j}} + \tau_{t_{(p+t_u)_{\text{S/PGW}'}}}\right\}^{\dagger\dagger\dagger}.\end{aligned} \tag{A.11}$$

(†††) This expression equals zero if there is no handover during session time.

(▷) $\mathrm{LDPD}_{k_{i(\text{PMIP})}}$ is calculated the same as $\mathrm{LDPD}_{k_{i(\text{GTP})}}$ by replacing $t_u \rightarrow$ GRE.

## B. Average Processing Cost of Data Packet Delivery for Mobile Node $i$

By (B.1), $\mathrm{ACDPD}_i$ is calculated as follows:

$$\mathrm{ACDPD}_i = \mathrm{CAP}_i + \frac{\sum_{h_i=1}^{N_{h_i}} \mathrm{CHP}_{h_i}}{N_{h_i}} + \frac{\sum_{k_i=1}^{N_{k_i}} \mathrm{CDPD}_{k_i}}{N_{k_i}}. \tag{B.1}$$

This section presents calculation of $\mathrm{ACDPD}_i$, in the centralized and decentralized architectures for both GTP and PMIP protocols.

### *B.1. The Centralized Architecture*

(▶)

$$\begin{aligned}\mathrm{CAP}_{i(\text{GTP})} &= \left\{\left[C_{t_{(\text{C.S.Req}+t_c)}} + C_{t_{(\text{C.S.Res}+t_c)}} + C_{t_{(\text{M.B.Req}+t_c)}} + C_{t_{(\text{M.B.Res}+t_c)}}\right] \text{ in SGW}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{SGW}\rightarrow\text{PGW}}} \left[C_{r_{(\text{C.S.Req}+t_c)_j}} + C_{r_{(\text{M.B.Req}+t_c)_j}}\right]\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{PGW}\rightarrow\text{SGW}}} \left[C_{r_{(\text{C.S.Res}+t_c)_j}} + C_{r_{(\text{M.B.Res}+t_c)_j}}\right]\right\} \\ &+ \left\{\left[C_{t_{(\text{C.S.Req}+t_c)}} + C_{t_{(\text{C.S.Res}+t_c)}} + C_{t_{(\text{M.B.Req}+t_c)}} + C_{t_{(\text{M.B.Res}+t_c)}}\right] \text{ in PGW}\right\}.\end{aligned} \tag{B.2}$$

(▷)

$$\begin{aligned}\mathrm{CAP}_{i(\text{PMIP})} &= \left\{\left[C_{p_{(\text{G.C.S.Req+SCTP+IPv6})}} + C_{p_{(\text{G.C.S.Res+SCTP+IPv6})}}\right] \text{ in SGW}\right\} \\ &+ \left\{\left[C_{t_{(\text{P.B.U+GRE})}} + C_{t_{(\text{P.B.A+GRE})}}\right] \text{ in SGW}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{SGW}\rightarrow\text{PGW}}} C_{r_{(\text{P.B.U+GRE})_j}}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{PGW}\rightarrow\text{SGW}}} C_{r_{(\text{P.B.A+GRE})_j}}\right\} \\ &+ \left\{\left[C_{t_{(\text{P.B.U+GRE})}} + C_{t_{(\text{P.B.A+GRE})}}\right] \text{ in PGW}\right\}.\end{aligned} \tag{B.3}$$

(▶)

$$\begin{aligned}\mathrm{CHP}^{\ddagger}_{h_{i(\text{GTP})}} &= \left\{\left[C_{t_{(\text{M.B.Req}+t_c)}} + C_{t_{(\text{M.B.Res}+t_c)}}\right] \text{ in SGW}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{SGW}\rightarrow\text{PGW}}} C_{r_{(\text{M.B.Req}+t_c)_j}}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{PGW}\rightarrow\text{SGW}}} C_{r_{(\text{M.B.Res}+t_c)_j}}\right\} \\ &+ \left\{\left[C_{t_{(\text{M.B.Req}+t_c)}} + C_{t_{(\text{M.B.Res}+t_c)}}\right] \text{ in PGW}\right\}.\end{aligned} \tag{B.4}$$

(‡) During each SGW handover, the $\mathrm{CHP}_{h_{i(\text{GTP})}}$ is repeated three times for eNodeB relocation and one time for SGW relocation by replacing SGW → SGW′.

(▷)

$$\begin{aligned}\mathrm{CHP}_{h_{i(\text{PMIP})}} &= 3 \times \left\{\left[C_{p_{(\text{G.C.S.Req+SCTP+IPv6})}} + C_{p_{(\text{G.C.S.Res+SCTP+IPv6})}}\right] \text{ in SGW}\right\} \\ &+ \left\{\left[C_{p_{(\text{G.C.S.Req+SCTP+IPv6})}} + C_{p_{(\text{G.C.S.Res+SCTP+IPv6})}}\right] \text{ in SGW}'\right\} + \left\{\left[C_{t_{(\text{P.B.U+GRE})}} + C_{t_{(\text{P.B.A+GRE})}}\right] \text{ in SGW}'\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{SGW}'\rightarrow\text{PGW}}} C_{r_{(\text{P.B.U+GRE})_j}}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\text{PGW}\rightarrow\text{SGW}'}} C_{r_{(\text{P.B.A+GRE})_j}}\right\} \\ &+ \left\{\left[C_{t_{(\text{P.B.U+GRE})}} + C_{t_{(\text{P.B.A+GRE})}}\right] \text{ in PGW}\right\}.\end{aligned} \tag{B.5}$$

(▶)

$$\begin{aligned}\mathrm{CDPD}_{k_{i(\mathrm{GTP})}} &= \left\{\sum_{j=1}^{H_{\mathrm{CN}\to\mathrm{PGW}}} C_{r_{(p)_j}}\right\} + \left\{C_{t_{(p+t_u)_{\mathrm{PGW}}}}\right\} \\ &+ \left\{\sum_{j=1}^{H_{\mathrm{PGW}\to\mathrm{SGW}}} C_{r_{(p+t_u)_j}}\right\} + \left\{C_{t_{(p+t_u)_{\mathrm{SGW}}}}\right\} \\ &+ \left\{C_{t_{(p+t_u)_{\mathrm{SGW}}}} + C_{r_{(p+t_u)_j}} + C_{t_{(p+t_u)_{\mathrm{SGW}'}}}\right\}^{\ddagger\ddagger}.\end{aligned} \tag{B.6}$$

(‡‡) This expression equals zero if there is no handover during session time.

(▷) $\mathrm{CDPD}_{k_{i(\mathrm{PMIP})}}$ is calculated the same as the $\mathrm{CDPD}_{k_{i(\mathrm{GTP})}}$ by replacing the $t_u \to$ GRE.

*B.2. The Decentralized Architecture*

(▶)

$$\begin{aligned}\mathrm{CAP}_{i_{(\mathrm{GTP})}} &= \left\{\left[C_{p_{(\mathrm{C.S.Req})}} + C_{p_{(\mathrm{C.S.Res})}} + C_{p_{(\mathrm{M.B.Req})}} + C_{p_{(\mathrm{M.B.Res})}}\right]\right. \\ &\left.\text{in S/PGW}\right\}.\end{aligned} \tag{B.7}$$

(▷)

$$\begin{aligned}\mathrm{CAP}_{i_{(\mathrm{PMIP})}} &= \left\{\left[C_{p_{(\mathrm{G.C.S.Req+SCTP+IPv6})}} + C_{p_{(\mathrm{G.C.S.Res+SCTP+IPv6})}}\right.\right. \\ &\left.\left.+ C_{p_{(\mathrm{P.B.U})}} + C_{p_{(\mathrm{P.B.A})}}\right] \text{in S/PGW}\right\}.\end{aligned} \tag{B.8}$$

(▶)

$$\begin{aligned}\mathrm{CHP}_{h_{i(\mathrm{GTP})}} &= 3 \times \left\{\left[C_{p_{(\mathrm{M.B.Req})}} + C_{p_{(\mathrm{M.B.Res})}}\right] \text{in S/PGW}\right\} \\ &+ \left\{\left[C_{p_{(\mathrm{M.B.Req})}} + C_{p_{(\mathrm{M.B.Res})}}\right] \text{in S/PGW}'\right\}.\end{aligned} \tag{B.9}$$

(▷)

$$\begin{aligned}\mathrm{CHP}_{h_{i(\mathrm{PMIP})}} &= 3 \times \left\{\left[C_{p_{(\mathrm{G.C.S.Req+SCTP+IPv6})}}\right.\right. \\ &\left.\left.+ C_{p_{(\mathrm{G.C.S.Res+SCTP+IPv6})}}\right] \text{in S/PGW}\right\} \\ &+ \left\{\left[C_{p_{(\mathrm{G.C.S.Req+SCTP+IPv6})}} + C_{p_{(\mathrm{G.C.S.Res+SCTP+IPv6})}}\right.\right. \\ &\left.\left.+ C_{p_{(\mathrm{P.B.U})}} + C_{p_{(\mathrm{P.B.A})}}\right] \text{in S/PGW}'\right\}.\end{aligned} \tag{B.10}$$

(▶)

$$\begin{aligned}\mathrm{CDPD}_{k_{i(\mathrm{GTP})}} &= \left\{\sum_{j=1}^{H_{\mathrm{CN}\to\mathrm{S/PGW}}} C_{r_{(p)_j}}\right\} \\ &+ \left\{C_{t_{(p+t_u)_{\mathrm{S/PGW}}}} + C_{r_{(p+t_u)_j}} + C_{t_{(p+t_u)_{\mathrm{S/PGW}'}}}\right\}^{\ddagger\ddagger\ddagger}.\end{aligned} \tag{B.11}$$

(‡‡‡) This expression equals zero if there is no handover during session time.

(▷) $\mathrm{CDPD}_{k_{i(\mathrm{PMIP})}}$ is calculated the same as the $\mathrm{CDPD}_{k_{i(\mathrm{GTP})}}$ by replacing the $t_u \to$ GRE.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update," 2016-2021 White Paper, 2017.

[2] "Ericsson Mobility Report," November 2016.

[3] "Deployment Strategies for Heterogeneous Networks," Nokia Whitepaper, 2013.

[4] "Dealing with Density: The Move to Small-Cell Architectures," RUCKUS Wireless White paper, 2014.

[5] H. A. Chan, H. Yokota, J. Xie, P. Seite, and D. Liu, "Distributed and dynamic mobility management in mobile internet: Current approaches and issues," *Journal of Communications*, vol. 6, no. 1, pp. 4–15, 2011.

[6] L. Bokor, Z. Faigl, and S. Imre, "Flat architectures: Towards scalable future internet mobility," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 6656, pp. 35–50, 2011.

[7] T. Taleb, K. Samdanis, and F. Filali, "Towards supporting highly mobile nodes in decentralized mobile operator networks," in *Proceedings of the 2012 IEEE International Conference on Communications, ICC 2012*, pp. 5398–5402, June 2012.

[8] C. Singhal and S. De, *Resource allocation in next-generation broadband wireless access networks*, IGI Global, 2017.

[9] S. Frei, W. Fuhrmann, A. Rinkel, and B. Ghita, "Signalling effort evaluation of mobility protocols within Evolved Packet Core network," in *Proceedings of the 8th International Network Conference, INC 2010*, pp. 99–108, July 2010.

[10] 3GPP.TS.23.402, "Universal Mobile Telecommunications System (UMTS); LTE; Architecture enhancements for non-3GPP accesses," 2015.

[11] 3GPP.TS.23.401, "General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access," 2015.

[12] M. Olsson and C. Mulligan, *EPC and 4G Packet Networks: Driving the Mobile Broadband Revolution*, Academic Press, 2nd edition, 2012.

[13] C. Schindelhauer, “Mobility in Wireless Networks,” in *SOFSEM 2006: Theory and Practice of Computer Science*, vol. 3831 of *Lecture Notes in Computer Science*, pp. 100–116, Springer, Berlin, Heidelberg, 2006.

[14] R. R. Roy, *Handbook of mobile ad hoc networks for mobility models*, Springer Science Business Media, 2010.

[15] M. K. Murtadha, N. K. Noordin, B. M. Ali, and F. Hashim, “Design and evaluation of distributed and dynamic mobility management approach based on PMIPv6 and MIH protocols,” *Wireless Networks*, vol. 21, no. 8, pp. 2747–2763, 2015.

[16] F. Giust, C. J. Bernardos, and A. De La Oliva, “Analytic evaluation and experimental validation of a network-based IPv6 distributed mobility management solution,” *IEEE Transactions on Mobile Computing*, vol. 13, no. 11, pp. 2484–2497, 2014.

[17] 3GPP.TS.23.060, “Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); General Packet Radio Service (GPRS); Service description; Stage 2,” 2011.

[18] H. Chan, K. Pentikousis, P. Seite, and A. Dutta, “Distributed Mobility Management Framework,” *Internet draft*, 2013.

[19] M. Karimzadeh, L. Valtulina, A. Pras et al., “Double-NAT Based Mobility Management for Future LTE Networks,” in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, San Francisco, Calif, USA, March 2017.

[20] M. Karimzadeh, L. Valtulina, H. v. Berg, A. Pras, M. Liebsch, and T. Taleb, “Software Defined Networking to support IP address mobility in future LTE network,” in *Proceedings of the 2017 Wireless Days (WD)*, pp. 46–53, Porto, Portugal, March 2017.

[21] M. Karimzadeh, L. Valtulina, H. v. Berg, A. Pras, P. G. Ortiz, and R. Sadre, “MultiPath TCP to Support User’s Mobility in Future LTE Network,” in *Proceedings of IFIP/IEEE Wireless and Mobile Networking Conference (WMNC)*, March 2017.

[22] F. Giust, L. Cominardi, and C. Bernardos, “Distributed mobility management for future 5G networks: Overview and analysis of existing approaches,” *IEEE Communications Magazine*, vol. 53, no. 1, pp. 142–149, 2015.

[23] P. Bertin, S. Bonjour, and J.-M. Bonnin, “Distributed or centralized mobility?” in *Proceedings of the 2009 IEEE Global Telecommunications Conference, GLOBECOM 2009*, December 2009.

[24] S. Jeon, S. Figueiredo, and R. L. Aguiar, “On the impacts of distributed and Dynamic Mobility Management strategy: A simulation study,” in *Proceedings of the 6th IFIP/IEEE Wireless Days Conference, WD 2013*, November 2013.

[25] L. Yi, H. Zhou, F. Ren, and H. Zhang, “Analysis of route optimization mechanism for distributed mobility management,” *Journal of Networks*, vol. 7, no. 10, pp. 1662–1669, 2012.

[26] S. Jeon, N. Kang, D. Corujo, and R. L. Aguiar, “Comprehensive performance evaluation of distributed and dynamic mobility routing strategy,” *Computer Networks*, vol. 79, pp. 53–67, 2015.

[27] W. Meng, M. Georgiades, and R. Tafazolli, “Signalling cost evaluation of mobility management schemes for different core network architectural arrangements in 3GPP LTE/SAE,” in *Proceedings of the 2008 IEEE 67th Vehicular Technology Conference-Spring, VTC*, pp. 2253–2258, May 2008.

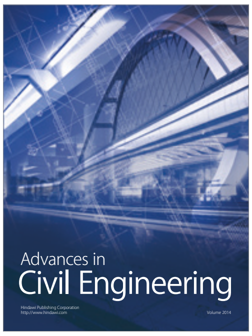
Journal of
Engineering

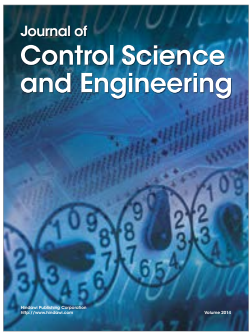
The Scientific World Journal

International Journal of
Rotating Machinery

Journal of
Sensors

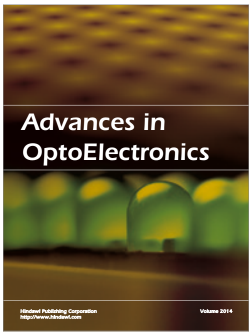
International Journal of
Distributed Sensor Networks


Advances in
Civil Engineering


Journal of
Control Science and Engineering

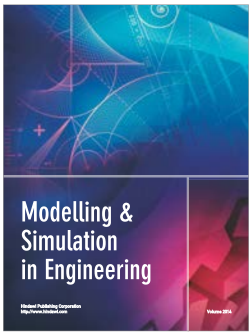
Journal of
Robotics

Journal of
Electrical and Computer Engineering

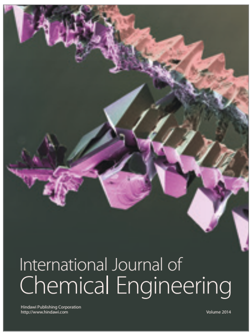

Advances in
OptoElectronics

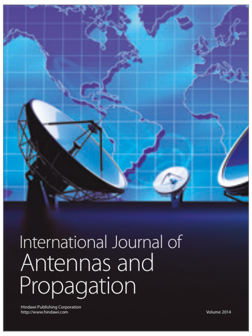
VLSI Design

International Journal of
Navigation and Observation

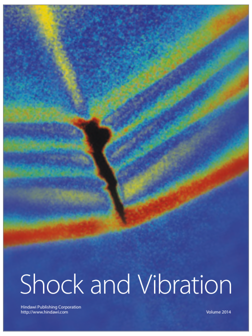

Modelling & Simulation in Engineering

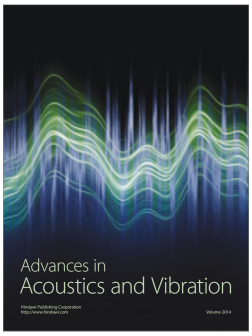
International Journal of
Aerospace Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and Propagation

Active and Passive Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration