Master's Thesis

Title

Bearer Aggregation Methods in Mobile Core Networks with C/U Plane Separation

Supervisor

Professor Morito Matsuoka

Author

Shuya Abe

February 9th, 2018

Graduate School of Information Science and Technology

Osaka University

Master’s Thesis

Bearer Aggregation Methods in Mobile Core Networks
with C/U Plane Separation

Shuya Abe

Abstract

In response to the growing demand for cellular networks, it is essential to improve the capacity of mobile core networks. Especially, in terms of accommodating machine-to-machine/Internet-of-Things (M2M/IoT) terminals into cellular networks, the load on the control and the user planes of the mobile core network increases massively. To deal with this problem, it is possible to apply virtualization technologies such as software-defined network and network function virtualization. However, few existing studies evaluate such solutions for mobile core networks numerically and in detail.

In this thesis, we first evaluate mobile core network architectures with virtualization technologies and control/user (C/U) plane separation using the mathematical analysis. We analyze and evaluate the performance of mobile core networks in terms of accommodating massive M2M/IoT terminals based on queueing theory and the actual source codes of mobile core network implementation. We then propose a novel bearer aggregation method that one bearer between gateway nodes is shared by multiple UEs, whereas one bearer corresponds to one UE in the traditional mobile core networks. This method reduces the control plane load especially when accommodating massive M2M/IoT terminals.

The result of numerical evaluation shows that the capacity of the mobile core network can be increased by up to 32.8% with node virtualization and C/U plane separation, and further by 201.4% by using bearer aggregation. Moreover, to maintain the performance of the mobile core network, we should carefully determine where the bearer aggregation is applied and when the shared bearer for each terminal is determined based on application characteristics and the number of accommodated M2M/IoT terminals.

Keywords

Mobile Core Network

M2M/IoT Communication

Software Defined Networks

C/U Plane Separation

Bearer Aggregation

Contents

List of Figures

List of Tables

1 Introduction

With increasing demand for cellular networks owing to the proliferation of rich user terminals such as smartphones and massive machine-to-machine/Internet-of-Things (M2M/IoT) terminals [1], increasing the capacity of mobile core networks is important [2]. Some M2M/IoT communications have characteristics different from those of rich user terminals—communication may occur periodically or intermittently with small amounts of data, while the number of terminals may be enormous. In addition, many M2M/IoT terminals have almost no mobility, and most of them only transmit data (i.e., no data are received).

One possible way to accommodate such terminals is to exploit a non-cellular wireless network called "Low Power, Wide Area Network (LPWAN) [3]". However, the realization of LPWANs which are not based on cellular network is costly because it requires the construction of a new network infrastructure. On the other hand, when accommodating massive M2M/IoT terminals to existing cellular networks, such as enhanced Machine Type Communication (eMTC), Narrowband Internet of Things (NB-IoT), and Extended Coverage GSM IoT (EC-GSM-IoT) [4–6], existing network infrastructure can be utilized. However, as more and more M2M/IoT terminals are accommodated to the cellular networks, the load on the core networks increases, especially on the control plane nodes. In this thesis, we focus on the signaling procedures based on the existing cellular network and propose a method to improve the network capacity.

In existing studies, the effects of virtualization technologies such as software-defined network (SDN) and network function virtualization (NFV) on long-term evolution/evolved packet core (LTE/EPC) networks have been discussed to address such problems [7–18]. In [19], our research group considered the mobile core network architecture for accommodating massive M2M/IoT terminals and showed the conceptual idea of the bearer aggregation method. However, these evaluations were based on a simple analysis model with severe assumptions.

Most of existing studies evaluate the performance of mobile core network based on the number and size of signaling messages. In this thesis, we introduce an analysis method for evaluating the performance of mobile core networks in terms of accommodating massive M2M/IoT terminals based on queueing theory and the actual source codes of mobile core network implementation. Moreover, we evaluate the performance of mobile core networks with node virtualization and C/U plane separation with SDN numericaly. We apply the node virtualization and C/U plane separation to a SGW/PGW and eNodeBs. We then propose and evaluate the bearer aggregation method that one bearer is shared by multiple UEs to concretize the primal idea proposed in [19]. In [19], we proposed the bearer aggregation at only SGW and did not discussed when the group of a UE for the bearer aggregation is determined. In this thesis, we defined two design parameters of the bearer aggregation. One is the node on which aggregation is applied, and the other is the timing when a group of UEs for a shared bearer is determined. In conclusion, we present of parameter design according to the characteristics of M2M/IoT terminals such as mobility, latency and quantity.

In Section 2, extant research related to our study is summarized. Section 3 presents the architectures of the mobile core network used herein. Section 4 introduces the bearer aggregation method. Section 5 describes the mathematical analysis of the performance of the mobile core network. In Section 6, we provide extensive numerical evaluation results and discussions. In Section 7, based on the evaluation results, we show the design of parameter settings for M2M/IoT terminals with various characteristics. Finally, in Section 8, we conclude this thesis with a brief summary and an outline of future work.

2 Related Work

Various methods have been proposed for improving the capacity of M2M/IoT communications in mobile cellular networks [20]. These existing works, as well as the method proposed in this thesis, are listed in Table 1.

In [7], the authors showed that efficient resource utilization can be achieved by implementing EPC nodes as software. In addition, [8–10] showed that low-cost mobility support can be realized by virtualizing EPC nodes as functions of NFV and distributing them over the network. The authors of [11] and [12] showed that applying NFV to EPC nodes decreases the amount of signaling traffic and the cost of devices, infrastructure, and energy consumption. In [12, 13, 16, 17], application of SDN to mobile core networks and virtualization of nodes in a cloud environment were studied.

The authors of [14] presented the design, implementation, and evaluation of two LTE/EPC architectures, one of which is based on SDN, and the other is based on NFV. From the evaluation results, an SDN-based EPC is better when handling large amounts of data traffic because it decreases the overhead of forwarding data packets. On the other hand, an NFV-based EPC is better at handling large signaling load, because every signaling message is handled with the SDN controller in an SDN architecture.

In [15], the authors proposed a hybrid SDN/NFV architecture which applies both the SDN decomposition and NFV concept for LTE/EPC networks. In the proposed architecture, the data plane functions of SGW and PGW can be located dynamically in either of a data center (cloud environment) in case of NFV deployment or a transport network in case of SDN deployment depending on QoS requirement. The authors evaluated the performance of these solutions and showed that the SDN decomposition decreases the network delay while increases the total network load, and that the NFV deployment increases the traffic delay while it does not increase the network load.

In these studies, the authors argued that the utilization of server resources can be improved and cost can be decreased by virtualizing nodes and applying SDN and NFV. However, the number of signaling procedures for virtualized functional modules and additional SDN control messages in mobile core networks may increase. Accommodating massive M2M/IoT terminals, the overhead on the control plane nodes cannot be ignored, especially when such terminals may be synchronized when in sending data. However, there has been almost no detailed evaluation of the relationship between SDN and increased signaling overhead, with the exception of [16] where the authors evaluated additional network traffic due to the application of SDN to the mobile core network. In [12, 13, 16], the signal processing load was evaluated on the basis of the number and size of messages sent and received by EPC nodes. However, the processing load of signaling messages is determined by many other factors such as the number of instructions executed in the node to process messages and node resources.

In [18], the authors proposed a modified packet core architecture and tunnel management methods, including bearer aggregation, specific to M2M traffic, and evaluated the performance of a mobile core network by using OpenAirInterface (OAI) [21], a software application for LTE/EPC networks written in C. However, this method requires the introduction of a new node in the mobile core network and extensive modification to the signaling message flows. On the other hand, our proposed method focuses on bearer aggregation with minimal modification of the current mobile core network.

Table 1: Comparison of existing methods and proposed method

	[7–11]	[12–16]	[17]	[18]	Proposed
Applying softwarelization technologies to EPC nodes	✓	✓	✓	✓	✓
Qualitative discussions on applying SDN		✓	✓	✓	✓
Overhead evaluation of virtualization and SDN		✓		✓	✓
Evaluation considering signal processing load					✓
Aggregating multiple bearers				✓	✓

3 Network Architecture

To evaluate the effects of node virtualization and C/U plane separation in mobile core networks, we consider the following three network architectures.

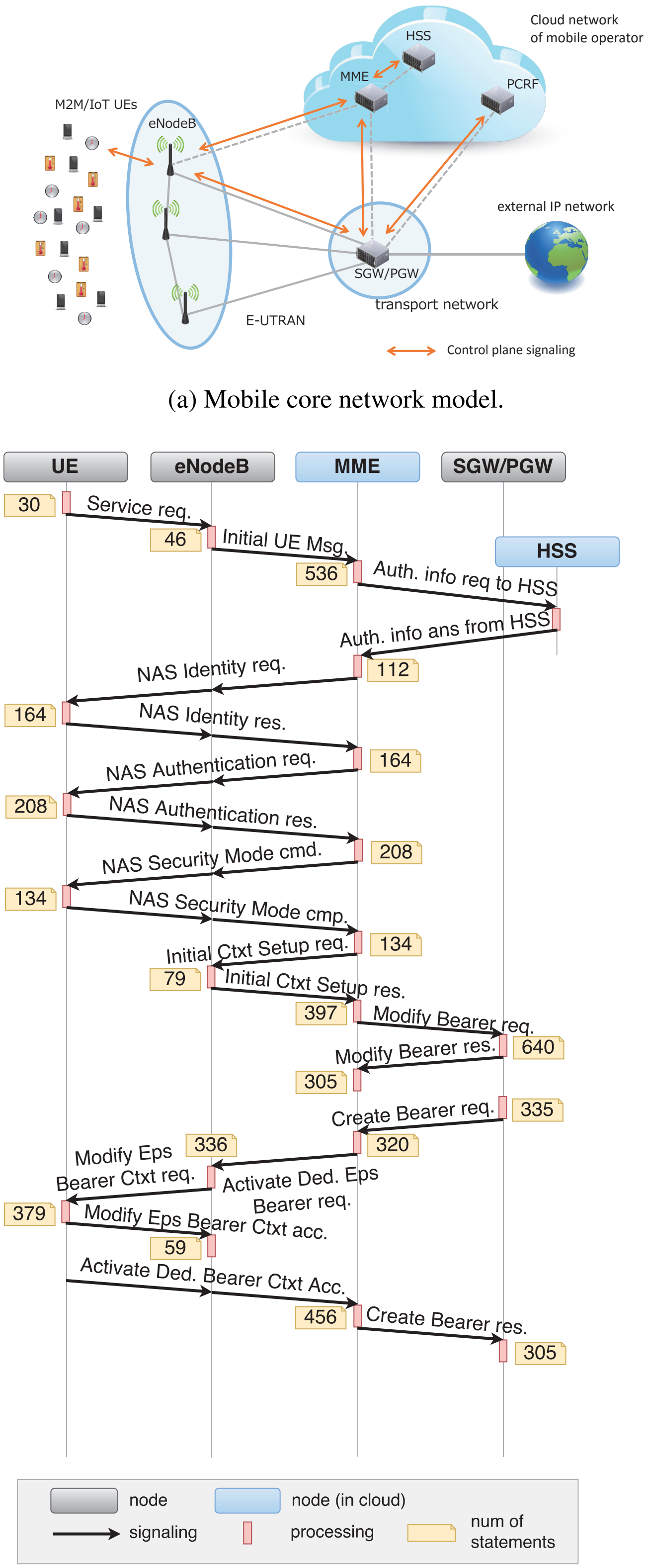
3.1 Conventional Core Network (CN)

The conventional mobile core network architecture is shown in Fig. 1(a). It consists of user equipments (UEs), eNodeBs, a serving gateway/packet data network gateway (SGW/PGW), a mobile management entity (MME), a home subscriber server (HSS), and a policy and charging rules function (PCRF). Note that SGW and PGW are integrated into a single node, as in the state-of-the-art implementation design of EPC. We assume that MME, HSS, and PCRF are virtualized and located in the cloud environment owned by the mobile network operator, and eNodeB and SGW/PGW are located in the transport network without virtualization. Fig. 1(b) shows the signaling flow when a UE changes its state from idle to active and requests start of communication. In this figure, req. and res. mean the signaling message is a request and a response message, respectively, and Msg. stands for "message". Ctxt, Ded., Acc mean "Context", "Dedicated", and "Accept", respectively. The figure includes the number of statements in programs executed by each node for processing each signaling message. The number of statements was obtained by analyzing the source code of OAI. Note that each processing of signaling messages has a different number of statements, meaning that each message imposes a different load on the corresponding mobile core node.

The networks where data is carried by RRC or NAS messages such as small data transmission in NB-IoT [5] can also be evaluated by assuming that the signaling procedure shown in Fig. 1(b) is terminated at NAS Security Mode cmp and a pair of a MME and an SGW can perform as a Cellular IoT Serving Gateway Node [5].

3.2 C/U Plane-separated SGW/PGW (PS)

Fig. 2(a) shows the architecture of the mobile core network with node virtualization and C/U plane separation in SGW/PGW. SGW/PGW is separated into SGW/PGW$_c$ for control plane functions and SGW/PGW$_d$ for data plane functions. SGW/PGW$_c$ is virtualized and located in the cloud environment, while SGW/PGW$_d$ is in the transport network without virtualization, as in CN. By applying C/U plane separation, the propagation delay between SGW/PGW$_c$ and MME becomes



(a) Mobile core network model.

(b) Signaling flow for bearer establishment.

Figure 1: Conventional Core Network (CN)

smaller than that in CN.

In this architecture, a GTP module [17], that is, a function matching the general packet radio service tunneling protocol (GTP) bearers established at SGW, is installed in the cloud network. This is because all control plane functions are located in the cloud environment in PS. Therefore, it is necessary for all data packets to pass through the cloud environment when UEs perform data communication after establishing the bearer, resulting in large propagation delays in the mobile core network. For the reason, this architecture is unsuitable for UEs with large amount of communication data such as smart phones, while it is acceptable for M2M/IoT terminals with small amount of transmitting and receiving data.

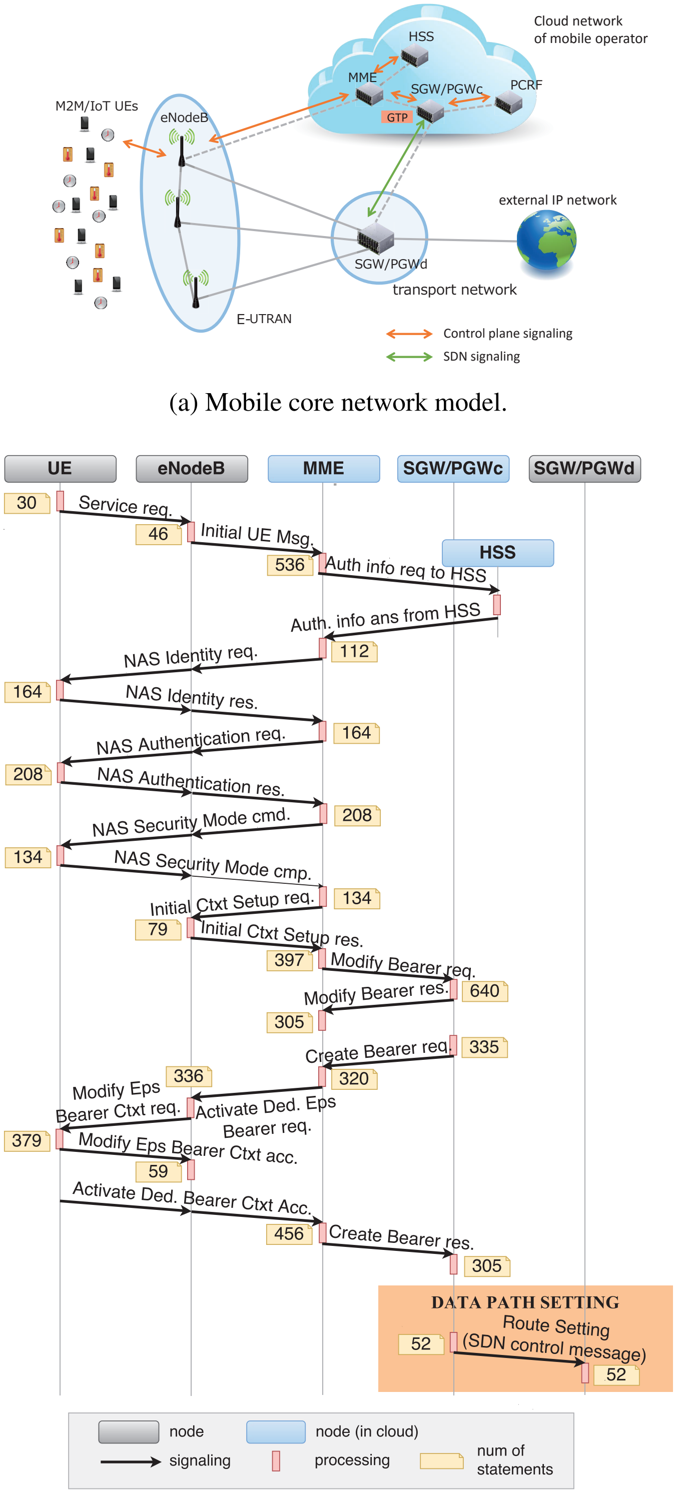
Fig. 2(b) shows the signaling flow when a UE changes its state from idle to active and requests the start of communication. Compared with Fig. 1(b), a control message related to route setting by SDN is required to be sent from SGW/PGW$_c$ to SGW/PGW$_d$ after signaling bearer establishment is complete. The number of statements necessary for processing this message is determined based on the source code of similar functions in OAI. The total propagation delay in signaling messages related to bearer establishment in this architecture is smaller than that in CN since messages between MME and SGW/PGWC are exchanged within the cloud environment.

3.3 C/U Plane-separated SGW/PGW with GTP Module Located in Data Plane (PS$_g$)

Fig. 3(a) shows the mobile core network architecture where the GTP module is implemented in the transport network. The GTP module can be implemented in the form of special hardware or software in SGW/PGW$_d$. This architecture can prevent the increase in propagation delay in the data plane found in PS, while an additional signaling message is required to configure the GTP module for bearer establishment, as shown in Fig. 3(b).

3.4 C/U Plane-separated SGW/PGW and eNodeB with GTP Modules Located in Data Plane (PSE$_g$)

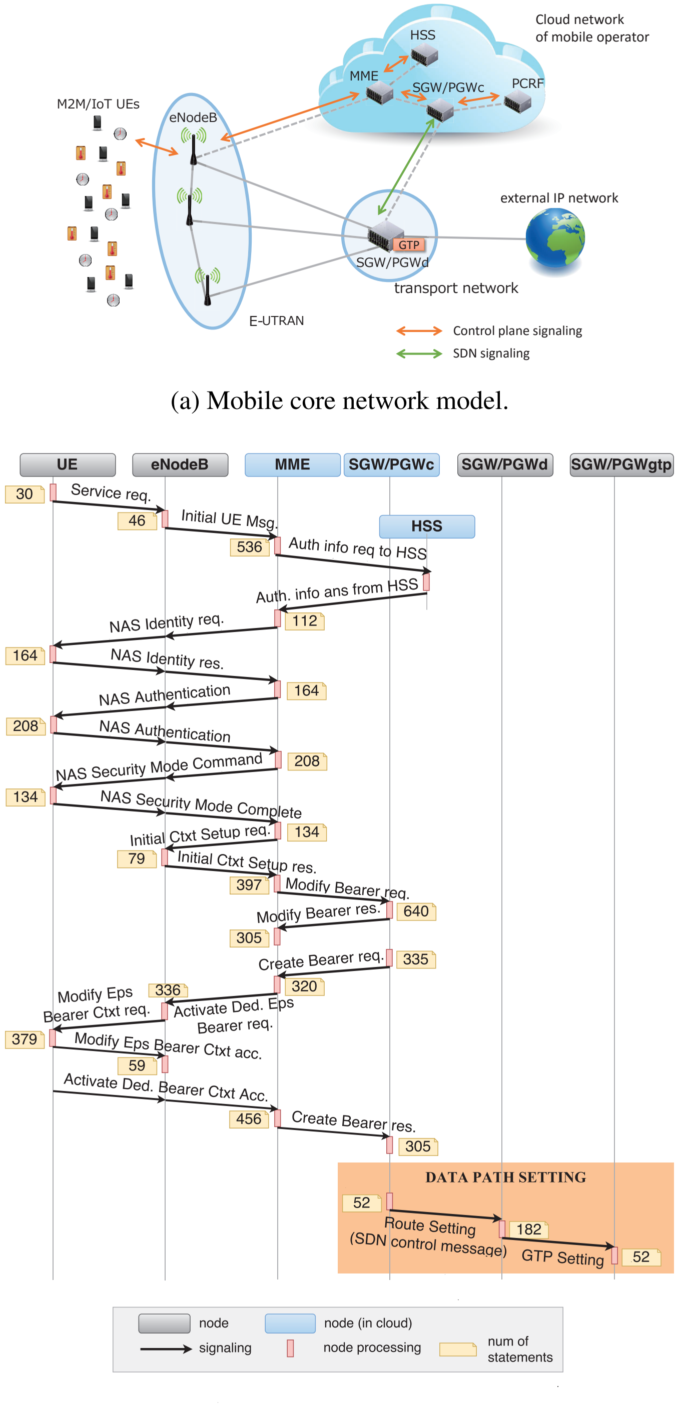
Fig. 4(a) shows the mobile core network architecture where the GTP modules of SGW and eNodeB are implemented in the transport network and E-UTRAN. Compared to PS and PS$_g$, in this architecture, C/U plane separation also applies to eNodeBs. Like the SGW/PGW in PS$_g$, the GTP



(a) Mobile core network model.

(b) Signaling flow for bearer establishment.

Figure 2: C/U Plane-separated SGW (PS)

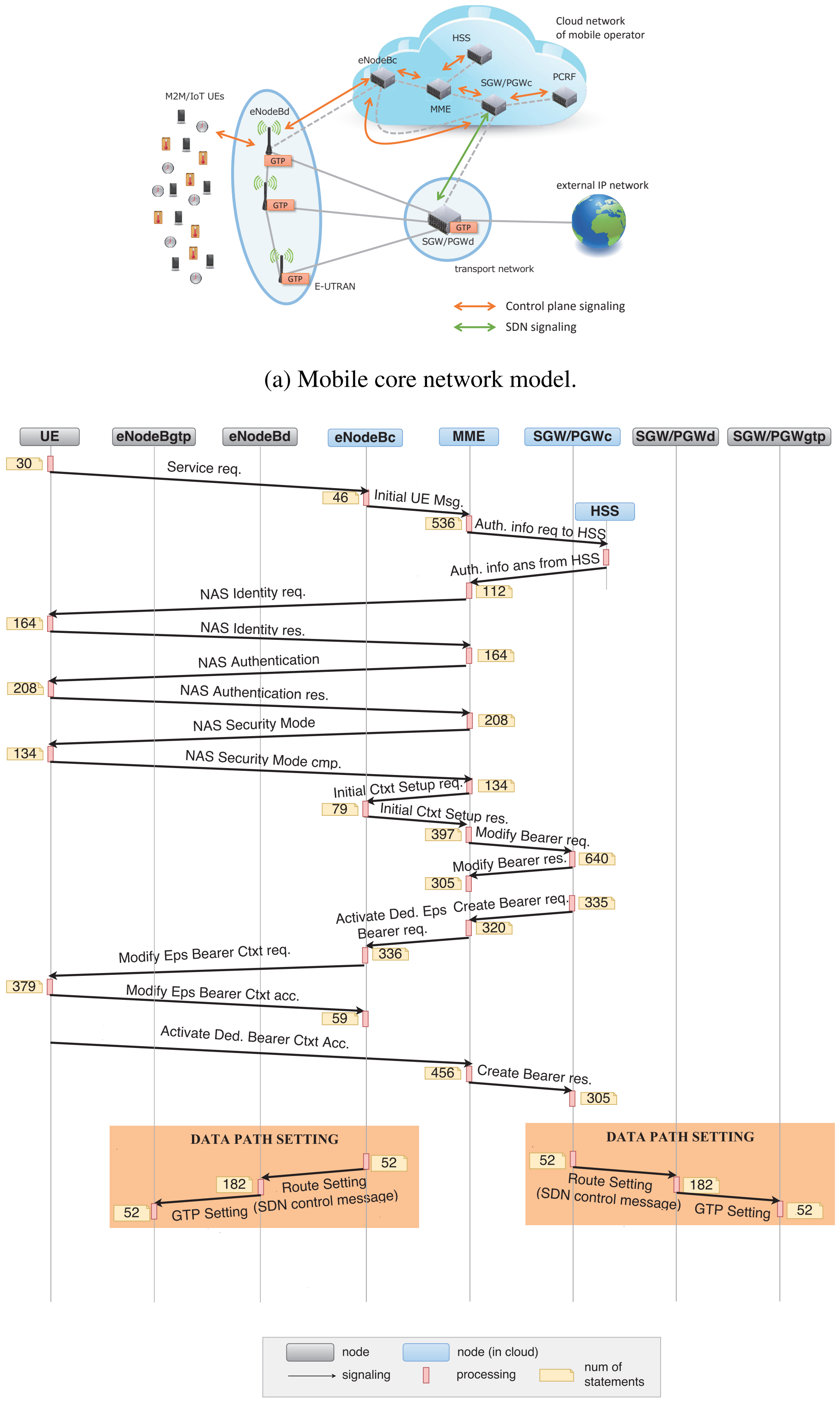


(a) Mobile core network model.

(b) Signaling flow for bearer establishment.

Figure 3: C/U Plane-separated SGW with GTP module located at Data Plane (PS$_g$)

modules of eNodeBs are implemented in eNodeB$_d$s in order to prevent the increase in propagation delay in the data plane. The similar architecture has been proposed to cooperatively control among neighboring base stations [22] and to integrate C-plane functions of multiple base stations [23]. We generalize such architectures to obtain the architecture in this subsection.



(a) Mobile core network model.

(b) Signaling flow for bearer establishment.

Figure 4: C/U Plane-separated SGW and eNodeB with GTP modules located at Data Plane (PSE$_g$)

4 Bearer Aggregation Method

One of the problems associated with accommodating massive M2M/IoT terminals in cellular networks is the increase in the number of bearers to be handled concurrently in the mobile core network. In this section, a bearer aggregation method is proposed to decrease the number of concurrent bearers.

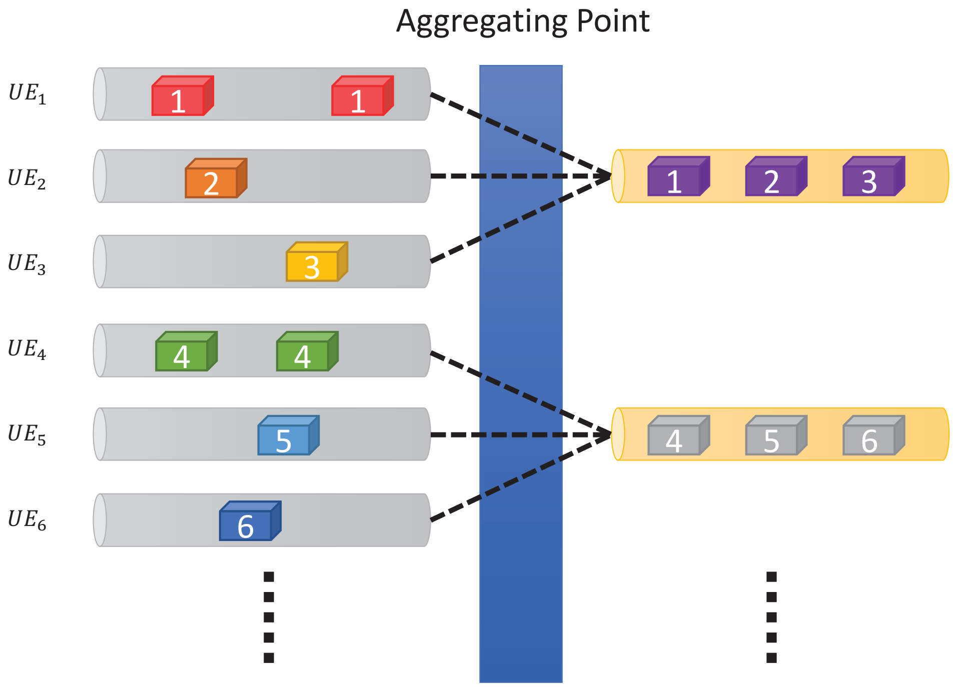
4.1 Overview

The bearer aggregation method reduces the load on EPC nodes by ensuring that one bearer is shared by multiple UEs; by contrast, in current mobile core networks, a single bearer corresponds to a single UE.

An illustration of the bearer aggregation method is given in Fig. 5(a). At the node where the aggregation method is applied, called an aggregation point, bearers from a *group* of UEs are aggregated into a single shared bearer. For example, in Fig. 5(a), when the aggregation point is an SGW and it aggregates multiple S1–u bearers between an eNodeB and the SGW into a single S5/S8 bearer between the SGW and a PGW, packets from UE_1, UE_2, and UE_3 passing through their S1–u bearers are injected into a shared S5/S8 bearer to be transmitted to the PGW. By this mechanism, the processing overhead of the SGW for handling signaling messages is reduced by decreasing the number of concurrent bearers at the node, while additional processing is required for maintaining the shared bearers. Note that the number of bearers aggregated by a single shared bearer, that is defined as *aggregation level* in this thesis, can vary. Furthermore, the aggregation level does not affect the U-plane performance since it only omit a part of signaling procedure in C-plane when establishing bearers.

One possible shortcoming of bearer aggregation is that the Quality of Service (QoS) of the data transmission can be considered within a unit of an aggregated bearer, while per-UE guarantee can be achieved in the conventional mobile core network. However, especially when considering M2M/IoT communication, it is likely that massive but homogeneous terminals from a single user are accommodated. In such a case, the degraded resolution of QoS guarantee would be acceptable. For example, 3GPP categorizes M2M/IoT UEs accomodated to the cellular network into some types [24] and UEs in the same categories can be aggregated into the same group.

We expect that this method can be applied to the LPWANs which have the similar signal-



(a) Bearer aggregation method.

	IMSI	vIMSI
UE_1	000 00 0000000001	000 00 1000000001
UE_2	000 00 0000000002	000 00 1000000001
UE_3	000 00 0000000003	000 00 1000000001
UE_4	000 00 0000000004	000 00 1000000002
UE_5	000 00 0000000005	000 00 1000000002
UE_6	000 00 0000000006	000 00 1000000002

(b) IMSI table.

Figure 5: Bearer Aggregation

ing procedures such as eMTC. Moreover, the aggregation method can be applied to large data transmission in NB-IoT that utilize bearers.

4.2 Virtual IMSI

To realize the bearer aggregation method with minimal modification to the behavior of the conventional mobile core network, we introduce the concept of *virtual International Mobile Subscriber Identity (vIMSI)*, which is associated with a shared bearer, in contrast to a normal IMSI, which is assigned uniquely to each UE and a corresponding bearer. MME handles matching between IMSIs and vIMSIs by maintaining an *IMSI table* (Fig. 5(b)) which represents the current status of bearer aggregation.

In the signaling flows shown in Fig. 1(b), Fig. 2(b) and Fig. 3(b), from the time when the flows begin to the time when a "NAS Security Mode" message is sent from a UE to an MME, the signaling messages include only a normal IMSI that corresponds to the UE. When an "NAS Security Mode" response message arrives at the MME, the MME searches the IMSI table in Fig. 5(b) to locate a vIMSI for the UE. Then, in the following signaling flow, signaling messages include both the normal IMSI for the UE and the vIMSI for the shared bearer. In addition, the MME notifies the PCRF of the correspondence between the IMSIs and vIMSIs when it updates the IMSI table.

4.3 Design Options

The bearer aggregation method has two design parameters. One is the node on which aggregation is applied, and the other is the timing when a group of UEs for a shared bearer is determined.

4.3.1 Aggregation Point

The bearers are aggregated at SGW or eNodeB. Table 2 shows the positive and negative effects with SGW or eNodeB aggregation.

Aggregation at SGW Multiple S1–u bearersbetween an eNodeB and an SGW are aggregated into a single S5/S8 bearer between the SGW and a PGW. This leads to a decrease in the number of Modify Bearer req/res. messages and Create Bearer req/res. messages for creating S5/S8 bearers.

Since an S5/S8 bearer is maintained while a UE is attached to the network, bearer aggregation at the SGW does not significantly influence the protocol for the establishment and release of S5/S8 bearers.

Aggregation at eNodeB Multiple radio bearers between UEs and an eNodeB are aggregated into a single S1–u bearer between the eNodeB and an SGW. The number of S5/S8 bearers is also reduced because an S1–u bearer and an S5/S8 bearer have a one-to-one relationship. Consequently, the decrease in the signaling overhead owing to aggregation at the SGW can be realized. Additionally, the number of Initial Context Setup req/res. messages for establishing S1–u bearers decreases.

However, we believe that this aggregation significantly affects the protocol. In the current mobile core networks, an S1–u bearer and a corresponding radio bearer are released simultaneously when a UE becomes idle. In contrast, when aggregation is applied at eNodeB, a shared S1–u bearer should be maintained until all UEs in the group for the shared bearer become idle.

Aggregation at both SGW and eNodeB The combination of aggregations at SGW and eNodeB can be considered to further decrease the number of bearers, in which some of the aggregated S1–u bearers are again aggregated into a single aggregated S5/S8 bearer.

In this aggregation, the IMSI table in Fig. 5(b) can also be used. In the second stage of the aggregation at the SGW, the vIMSI of the aggregated S1–u bearer is added to the IMSI column and the vIMSI of the S5/S8 bearer is added to the vIMSI column.

This aggregation inherits the advantages and disadvantages of both aggregations at SGW and at eNodeB. Therefore, it significantly affects the protocol in the sense that a shared S1–u bearer should be maintained until all UEs in the group for the shared bearer become idle, as in the aggregation at eNodeB.

4.3.2 Aggregation Timing

The group of a UE for bearer aggregation is determined when the UE attaches to the network or when the UE becomes active and the communication request is issued. Table 3 describes the positive and negative effects with pre-determined aggregation or on-demand aggregation.

Pre-determined Aggregation The group of a UE for bearer aggregation is determined when the UE attaches to the network. The assignment of a vIMSI by MME and the notification to the PCRF are conducted after that.

Fig. 6(a) shows a timeline of the signal processing at the MME with a pre-determined aggregation method. The vertical dashed lines represent arrivals of the communication requests from a group of UEs. When the first UE (UE_1 in the figure) arrives, the corresponding shared bearer is established (Bearer Establishment depicted as a blue box in the figure). Therefore, the following UEs ($UE_2...UE_K$) do not require the establishment procedure of the shared bearer. However, the data path setting to data plane nodes (green box) is necessary for each UE.

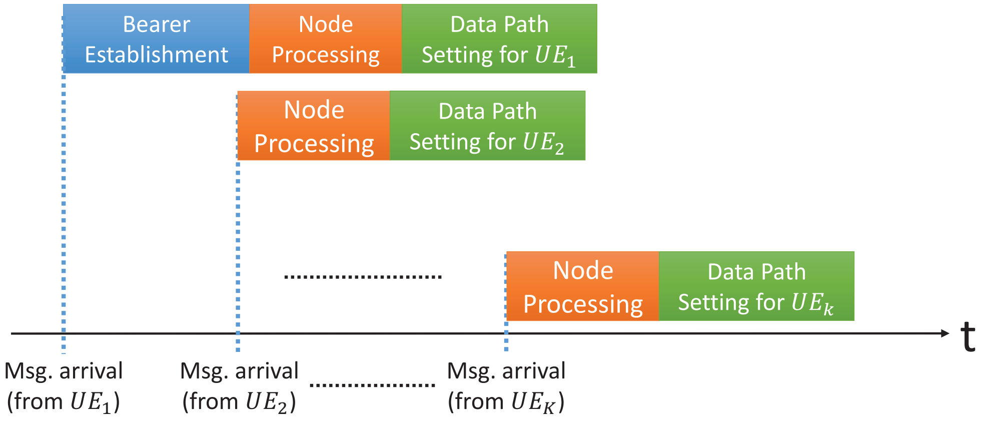
On-demand Aggregation The group of a UE for bearer aggregation is determined when the UE becomes active and the communication request is issued, not when the UE attaches to the network. Therefore, a notification is sent to the PCRF every time the UE initiates communication. Fig. 6(b) shows an example of the on-demand aggregation method. Each UE waits until the number of UEs reaches the required aggregation level, which is the number of UEs in each group. Then, assignment of the vIMSI, sending of a notification to the PCRF, and establishment of the shared bearer occur. Note that this method requires only one data path setting procedure for all UEs in the group. However, a UE experiences a waiting time between the time it sends a communication request and the time at which the shared bearer is established.

Table 2: Positive and Negative Effect with SGW or eNodeB Aggregation

Aggregation Point	Positive Effect	Negative Effect
SGW	required small modification	low network capacity, large bearer establishment time
eNodeB	high network capacity, small bearer establishment time	required large modification

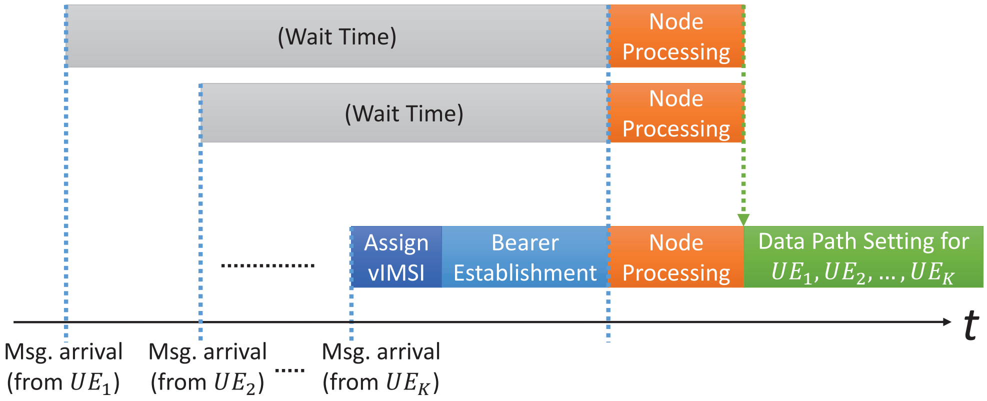
Table 3: Positive and Negative Effect with Pre-determined or On-demand Aggregation

Aggregation Timing	Positive Effect	Negative Effect
Pre-determined	not required wait time	low network capacity,
On-demand	high network capacity	required wait time



Bearer Establishment
Node Processing
Data Path Setting for UE_1
Node Processing
Data Path Setting for UE_2
Node Processing
Data Path Setting for UE_k
t
Msg. arrival (from UE_1)
Msg. arrival (from UE_2)
Msg. arrival (from UE_K)

(a) Pre-determined aggregation method.



(Wait Time)
Node Processing
(Wait Time)
Node Processing
Assign vIMSI
Bearer Establishment
Node Processing
Data Path Setting for $UE_1, UE_2, \ldots, UE_K$
t
Msg. arrival (from UE_1)
Msg. arrival (from UE_2)
Msg. arrival (from UE_K)

(b) On-demand aggregation method.

Figure 6: Aggregation Timing

5 Performance Analysis

We calculate *the network capacity* and *the bearer establishment time* to evaluate the performance of mobile core networks. We define the network capacity as the maximum number of UEs that can be accommodated with saturated utilization of nodes in a mobile core network. The bearer establishment time is defined as the time from when the signaling flow starts to the time when it ends, assuming the network models and signaling flows explained in Section 3.

We first calculate the processing time for a signaling message at a node. We then derive the network capacity and the bearer establishment time based on the processing time.

5.1 Notations

m is one of CN, PS and PS$_g$ defined in Section 3. For the individual node, we abbreviate UE, eNodeB, MME, and SGW/PGW as U, B, M, and G, respectively. B_c, B_d, and B_g respectively represent a control plane node, a data plane node, and a GTP module for eNodeB and G_c, G_d, and G_g respectively represent a control plane node, a data plane node, and a GTP module for SGW/PGW. The propagation delay of signaling messages between nodes N_1 and N_2 is denoted by τ_{N_1,N_2}. The average processing time for a signaling message at node N is denoted by t_N. C_{N_1,N_2} means the number of signaling messages transmitted from N_1 to N_2. The number of messages processed at node N in the signaling flow is denoted by P_N. n_N represents the number of nodes N in the network. A_N is the server resources of node N in terms of the number of statements that can be processed per second. L_{N_i} is the number of statements required for processing the ith signaling message at node N. Note that N, N_1, and N_2 mean one of U, B, B_c, B_d, B_g, M, G, G_c, G_d, and G_g.

We assume that each UE starts the communication at regular intervals of D, which is called a communication period of a UE. In addition, there is a randomness in the start timing within the period. Specifically, all UEs initiate communication requests randomly within a certain time interval $D'(< D)$. K represents the aggregation level in bearer aggregation ($K = 3$ in Fig. 5). We set K to a constant value in this evaluation for simplicity.

5.2 Processing Time

To derive the average processing time at a node N, we employ the M/G/1/PS queuing model. We assume that the arrival of the signaling messages at a node follows the Poisson distribution. In the M/G/1/PS model, the mean sojourn time $E[R]$ can be derived as

$$E[R] = \frac{\rho^r}{1-\rho} \cdot \frac{E[S^2]}{2E[S]} + \frac{1-\rho^r}{1-\rho} \cdot E[S], \tag{1}$$

where λ is the job arrival rate, $S(x)$ is the workload distribution, $E[S]$ is the mean workload, r is the maximum number of parallel processing runs, and the system utilization is determined as follows:

$$\rho = \lambda \cdot E[S]. \tag{2}$$

In the analysis, the number of signaling messages to be processed per unit time at node N, also called *the signaling processing frequency*, is regarded as the job arrival rate. The time distribution of the processing of signaling messages at node N is used for the workload distribution, S_N. Then, the mean workload $E[S_N]$ can be calculated on the basis of the average number of statements for processing signaling messages and the server resources of node N. Therefore, λ_N, $E[S_N]$, and $E[S_N^2]$ are derived as

$$\begin{aligned} \lambda_N &= \frac{P_N \cdot n_U}{D \cdot n_N}, & (3)\\ E[S_N] &= \sum_{i=1}^{P_N} \frac{L_{N_i}}{A_N \cdot P_N}, & (4)\\ E[S_N^2] &= \sum_{i=1}^{P_N} \frac{L_{N_i}^2}{A_N^2 \cdot P_N}. & (5) \end{aligned}$$

5.3 Network Capacity

Solving (2) for n_U with (3) and (4), we obtain

$$n_U = D \cdot \rho \frac{n_N \cdot A_N}{\sum_{i=1}^{P_N} L_i}. \tag{6}$$

By substituting $\rho = 1$ for (6), we can obtain the number of UEs that can be accommodated at the node N. We denote it by n_{UNmax}. The network capacity, denoted as $n_{U_{max}}$ is the minimum value of n_{UNmax} for all nodes in the network.

$$n_{U_{max}} = \min_{N \in \mathbb{V}_m} \left(D \frac{n_N \cdot A_N}{\sum_{i=1}^{P_N} L_i} \right). \tag{7}$$

We assume that some nodes in the network are located in the cloud environment and their server resources can be configured while ensuring that the total amount of server resources remains fixed. When we optimize the server resources of the nodes in the network to obtain the maximum value of n_{UNmax} for all nodes becomes identical, that also equals to the network capacity in (7). We can calculate the server resources of each N in such a situation, called as *optimized server resources* of node N and denoted as A'_N, as follows. Note that $\mathbb{W}$ represents a set of the nodes whose server resources are optimized.

$$A'_N = \left(\sum_{I \in \mathbb{W}} A_I \right) \cdot \frac{\sum_{i=1}^{P_N} L_i}{\sum_{I \in \mathbb{W}} \sum_{i=1}^{P_I} L_i} \cdot \frac{1}{n_N}. \tag{8}$$

When applying a bearer aggregation method, both of the signaling processing frequency λ and the average signaling processing time $E[S]$ change. λ decreases because P_N in (3) gets smaller. On the other hand, $E[S]$ decreases because L_i in (4) decreases. Specifically, when applying a bearer aggregation method with the aggregation level K, the number of signaling messages related to the bearer establishment decreases to $1/K$. In detail, when i is the signaling processing to be aggregated, the number of statements L_i becomes L_i/K with the bearer aggregation.

From (6), we can see that the network capacity is proportional to the number of bottleneck nodes. Also, the network capacity is proportional to the amount of server resources. Therefore, when the server resources are optimized, because a more proportion of server resources is allocated to the bottleneck nodes, the network capacity increases. In addition, applying bearer aggregation to the mobile core network reduces the number of statements and signaling processings, resulting in increased network capacity.

5.4 Bearer Establishment Time

The bearer establishment time T is the sum of propagation delay of all signaling messages, T_τ; processing times of all messages, T_t; and the waiting time required when using on-demand aggregation, T_w. We derive the bearer establishment time by (9).

$$\begin{aligned} T &= T_\tau + T_t + T_w \\ &= \sum_{N_1,N_2 \in \mathbb{V}_m} (C_{N_1,N_2} \cdot \tau_{N_1,N_2}) + \sum_{N \in \mathbb{V}_m} (P_N \cdot t_N) + T_w, \end{aligned} \tag{9}$$

where T_w is calculated by (10) on the basis of the communication period of a UE, the number of UEs attached to the network, and the aggregation level.

$$T_w = \begin{cases} \frac{K \cdot D}{2n_U} & \text{(Aggregation at SGW)} \\ \frac{K \cdot D \cdot n_B}{2n_U} & \text{(Aggregation at eNodeB)} \end{cases} \tag{10}$$

In what follows, we assess the effect of n_U on T. Since T_τ does not depend on n_U, we can obtain the following equation for PS$_g$ with on-demand aggregation at SGW by differentiating T with respect to n_U. Note that in case of aggregation at eNodeB, the second term of the following equation is multiplied by n_B.

$$\begin{aligned} \frac{dT}{dn_U} &= \frac{dT_t}{dn_U} + \frac{dT_w}{dn_U} \\ &= \sum_{N \in \mathbb{V}_m} \left(P_N \frac{dt_N}{dn_U} \right) + \left(-\frac{D \cdot K}{2n_U^2} \right) \end{aligned} \tag{11}$$

$\frac{dt_N}{dn_U}$ is expressed as (12). This represents the increase in the average signaling processing time t_N at the node N when the number of UEs increases.

$$\begin{aligned}\frac{dt_N}{dn_U} &= \frac{dt_N}{d\rho} \cdot \frac{d\rho}{dn_U} \\ &= \big((2E[S]^2 - E[S^2])(r-1)\rho^r \\ &- (2E[S]^2 - E[S^2])r \cdot \rho^{r-1} + 2E[S]^2\big) \\ &/2E[S](1-\rho)^2 \frac{d\rho}{dn_U}\end{aligned} \tag{12}$$

Assuming that the maximum number r of parallel processing runs corresponding to each node is one, (12) can be converted as

$$\begin{aligned}\frac{dt_N}{dn_U} &= \frac{P_N \cdot D \cdot n_N \cdot E[S^2]}{2(D \cdot n_N - P_N \cdot E[S] \cdot n_U)^2} \\ &= \frac{D \cdot n_N \cdot \sum_{i=1}^{P_N} L_i^2}{2\left(D \cdot n_N \cdot A_N - \sum_{i=1}^{P_N} L_i \cdot n_U\right)^2}.\end{aligned} \tag{13}$$

Substituting (13) for (11), we obtain

$$\frac{dT}{dn_U} = \sum_{N \in \mathbb{V}_m} \frac{P_N \cdot D \cdot n_N \cdot \sum_{i=1}^{P_N} L_i^2}{2\left(D \cdot n_N \cdot A_N - \sum_{i=1}^{P_N} L_i \cdot n_U\right)^2} - \frac{D \cdot K}{2n_U^2}. \tag{14}$$

From (13), we can see that the increase in the average signaling processing runs corresponding to time is inversely proportional to the square of the number of accommodated terminals. $\frac{dt_N}{dn_U}$ diverges to positive infinity when n_U approaches $\frac{D \cdot n_N \cdot A_N}{\sum_{i=1}^{P_N} L_i}$, which is identical to $n_{U N max}$ in (6) with $\rho = 1$. Therefore, with (14), by appropriately allocating server resources to EPC nodes as in (8), we can maximize the number of accommodated UEs with finite value of the bearer establishment time.

6 Numerical Evaluation

In this section, we show the numerical results of the analysis in Section 5 for evaluating the effects of the node virtualization and C/U plane separation with SDN discussed in Section 3 and the bearer aggregation method proposed in Section 4.

6.1 Evaluation Candidates and Parameter Settings

We evaluate the performance of six different bearer aggregation methods, each of which combines aggregation point and aggregation timing. For comparison, we evaluate the performance of a model without bearer aggregation. The notations for these methods are as follows.

- NA: no aggregation
- PA–SGW: pre-determined aggregation at SGW
- OA–SGW: on-demand aggregation at SGW
- PA–eNB: pre-determined aggregation at eNodeB
- OA–eNB: on-demand aggregation at eNodeB
- PA–SGWeNB: pre-determined aggregation at both of SGW and eNodeB
- OA–SGWeNB: on-demand aggregation at both of SGW and eNodeB

The communication period of a UE is set to 600 seconds. The network model has 2,000 eNodeBs, one MME and one SGW/PGW for CN. In PS and PS$_g$, SGW/PGW is divided into one SGW/PGW$_c$ and SGW/PGW$_d$. One GTP module of SGW/PGW exists in the network for PS and PS$_g$. In PSE$_g$, eNodeBs are divided into one eNodeB$_c$ and multiple eNodeB$_d$s and GTP modules of eNodeBs exist in the network. We change the number of UEs to be attached to the network, while each eNodeB accommodates an identical number of UEs. The propagation delays of signaling messages between nodes are configured as follows. Note that the propagation delays do not include the processing time for signaling messages.

- UE–eNodeB, eNodeB$_d$: 20 msec
- UE–eNodeB$_c$: 30 msec

- eNodeB, eNodeB$_d$–SGW/PGW, SGW/PGW$_d$: 7.5 msec

- eNodeB, eNodeB$_d$–MME, eNodeB$_c$: 10 msec

- SGW/PGW, SGW/PGW$_d$–MME, SGW/PGW$_c$: 10 msec

- MME–eNodeB$_c$, SGW/PGW$_c$: 1 msec

- GTP module–SGW/PGW$_d$, eNodeB$_d$: 1 msec

The default values of the server resources of nodes as follows.

- UE: 3,000 statements/sec

- eNodeB: 300,000 statements/sec

- eNodeB$_c$: 600,000,000 statements/sec

- eNodeB$_d$: 300,000 statements/sec

- MME: 3,000,000 statements/sec

- SGW/PGW: 3,000,000 statements/sec

- SGW/PGW$_c$: 3,000,000 statements/sec

- SGW/PGW$_d$: 3,000,000 statements/sec

- GTP module: 600,000 statements/sec

These values were determined on the basis of discussions with researchers from a mobile network operator in Japan, assuming a nation-wide mobile core network and we set the average values of propagation delays assuming EPC deployed in Japan as follows. The cloud environment which has MME, eNodeB$_c$ and SGW/PGW$_c$ is located in major metropolitan areas in Japan, namely, Tokyo, Osaka and Nagoya. SGW/PGWs or SGW/PGW$_d$s locate in these major metropolitan areas and regional hub cities, namely, Sapporo, Sendai, Hiroshima and Fukuoka. ENodeBs or eNodeB$_d$s are deployed to the whole part of Japan and we configured the number of eNodeB from the white paper from Ministry of Internal Affairs and Communications [25].

We assume that in PS and PS$_g$, the server resources located in the cloud environment (MME, eNodeB$_c$ and SGW/PGW$_c$) can be optimized so that the loads on the servers are identical, while the sum of the server resources remains unchanged from the above mentioned default values.

The number of statements for processing each signaling message in the signaling flow in Section 4 is determined on the basis of the source code of OAI. Note that we ignore the number of statements for maintaining and searching the IMSI table in the bearer aggregation because we assume it is sufficiently smaller than that for other signaling messages.

6.2 Evaluation Results

6.2.1 Network Architecture Comparisons

Fig. 7 shows the relationship between the number of accommodated UEs and the bearer establishment time when we utilize the mobile core networks based on the four architecture discussed in Section 3. In the figure, (unoptimized) means that the server resources are set to the default values and are not optimized, while (optimized) represents the results with the optimization of server resources. As shown in the figure, when the number of UEs reaches a certain value, the bearer establishment time increases sharply. This is because the load on one of the nodes in the network increases to 100%. In what follows, we use that number of UEs as the network capacity.

Without server resource optimization, there is almost no difference among the four networks in terms of network capacity. In these networks, the bottleneck node of the network capacity is MME, and even when the load on SGW/PGW$_c$ and eNodeB$_c$ decreases owing to C/U plane separation, the system utilization of MME remains unaffected.

A comparison of the results obtained with and without server resource optimization in PS or PS$_g$, shows that the network capacity increases by 32.8% after server resource optimization. This is because C/U plane separation reduces the load on SGW/PGW$_c$, and server resource optimization makes a greater proportion of server resources available for allocation to the bottleneck nodes.

Moreover, regardless of whether server resources are optimized, CN yields a slightly longer bearer establishment time than PS or PS$_g$. This is because, in PS and PS$_g$, propagation delays are reduced by placing the control plane function of SGW in the cloud environment. PS and PSg yield almost the same bearer establishment times because the number of signaling messages and processing of signaling messages are almost identical. However, based on the difference

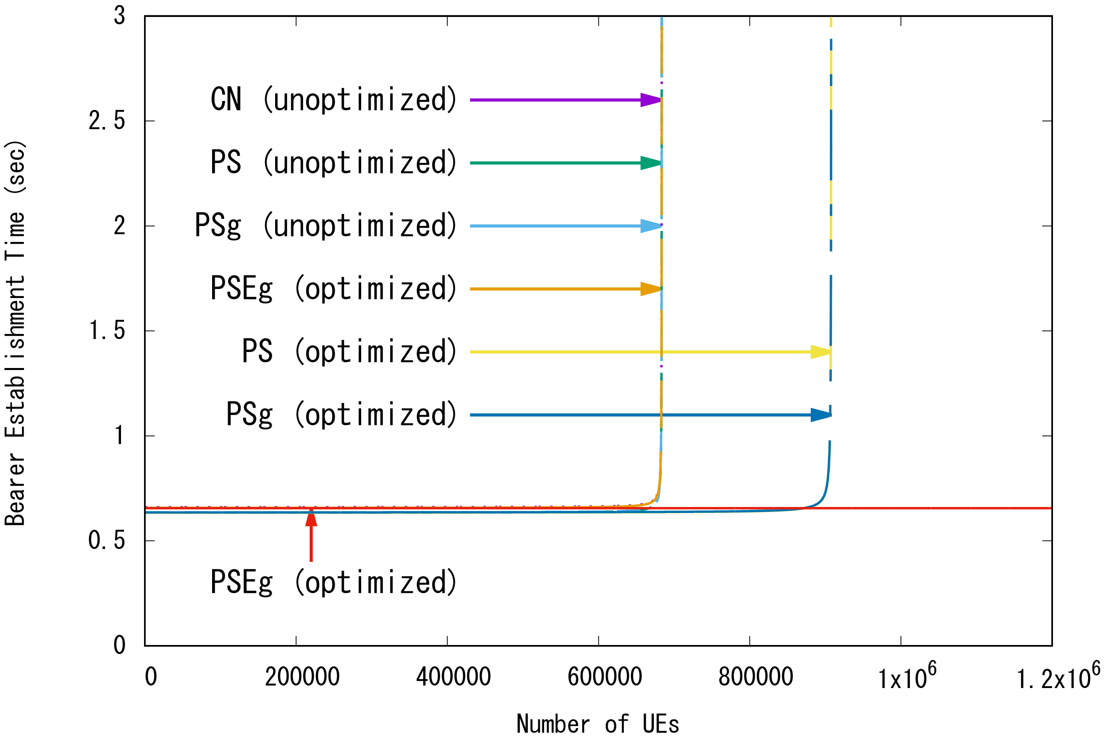


Figure 7: Comparison of network architecture

between Fig. 2(b) and Fig. 3(b),there is one additional signaling message, GTP Setting between the SGW/PGW$_d$ and GTP module, in the case of PS$_g$. Therefore, the total propagation delay of PS$_g$ is slightly larger than that of PS, and this difference cannot be recognized in Fig. 7.

In PSE$_g$, the server resource optimization can be applied to eNodeB$_c$. However, its effect depends on how many eNodeBs an SGW can accomodate, how many UEs an eNodeB can accomodate, and the amount of server resources of eNodeBs compared with EPC nodes, that cannot be determined in reasonable way. Therefore, we discuss about only bearer establishment time in PSE$_g$ where there are enough small number of UEs exist. Fig. 7 shows that the bearer establishment time slightly increases with C/U plane separation at eNodeB. This is because, in PSE$_g$, propagation delays increase by placing the control plane function of eNodeB in the cloud environment and there is only one additional signaling message between eNodeB$_c$ and eNodeB$_d$. We conclude that especially for M2M/IoT application scenarios, this increase can be allowed.

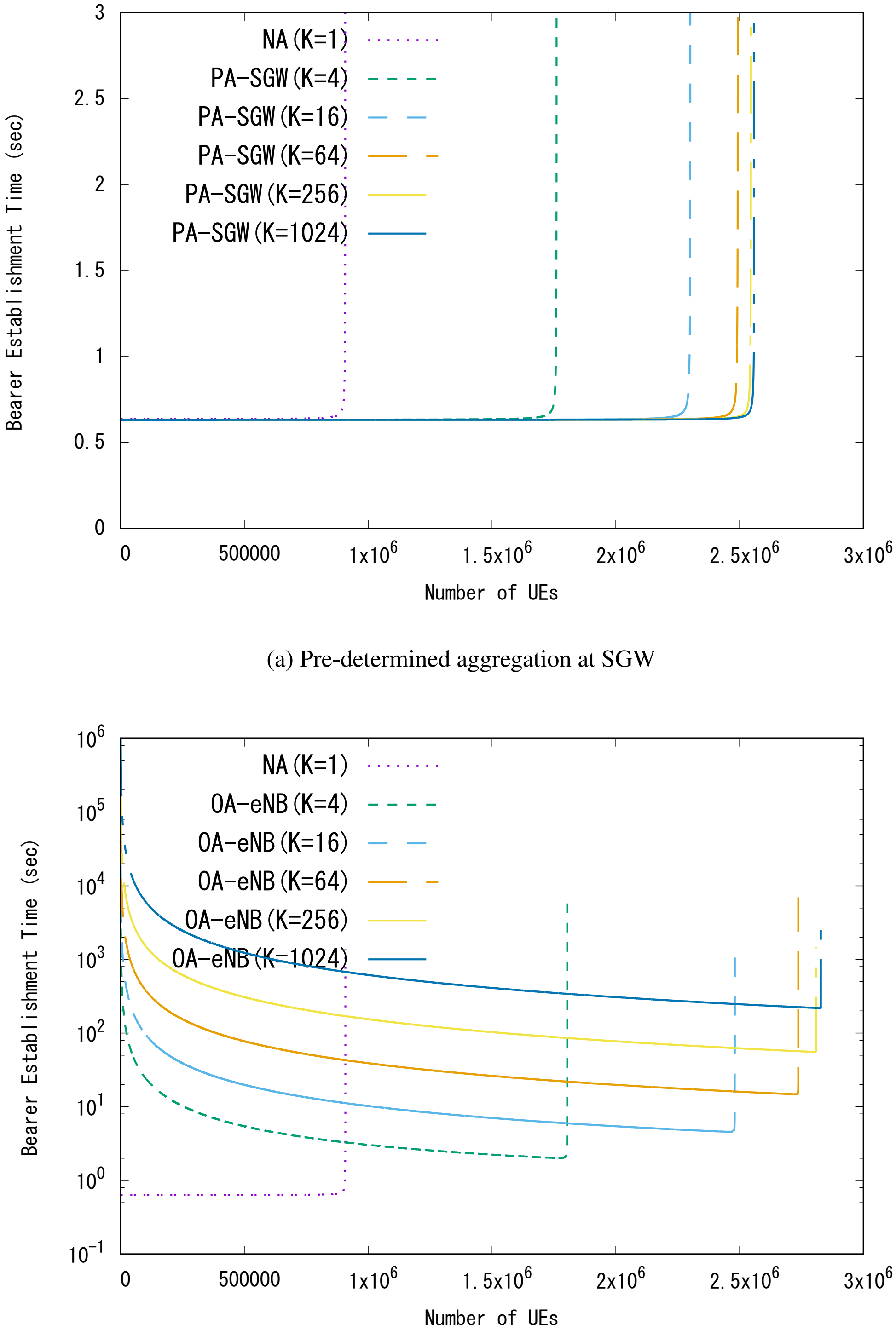
In the following evaluation, PS$_g$ is utilized for exploring the effect of bearer aggregation on the capacity of the network.

6.2.2 Aggregation Level

Figs. 8(a) and 8(b) show the results with server resource optimization when applying pre-determined aggregation at SGW and on-demand aggregation at eNodeB, respectively. In the figure, K = i indicates the results obtained by setting the aggregation level K to i.

We can see from these figures that the network capacity increases further by applying a bearer aggregation method ($K > 1$) regardless of the combination of aggregation point and timing. The performance gain is up to 181.8% when we compare NA K=1 and PA-SGW K=1024. This is because bearer aggregation reduces the number of signaling messages to be processed by MME and SGW/PGW$_c$, which, in turn, reduces the server load. However, when the aggregation level is higher than 64, the network capacity remains almost unchanged. This is because the signaling overhead that can be removed by bearer aggregation is sufficiently small and can be ignored. In the following evaluation, the aggregation level is set to 64.

Fig. 8(b) shows that when applying the on-demand aggregation at eNodeB, the bearer establishment time becomes significantly large. This is caused by the waiting time shown in Fig. 6(b). Equation (10) shows that the waiting time is proportional to the aggregation level and inversely proportional to the number of accommodated UEs. Therefore, when the aggregation level de-



(a) Pre-determined aggregation at SGW

(b) On-demand aggregation at eNodeB (log-scale)

Figure 8: Effect of aggregation level

creases or when the number of accommodated UEs increases, the bearer establishment time decreases.

6.2.3 Combination of Server Resource Optimization and Bearer Aggregation

Fig. 9 shows the results without both bearer aggregation and server resource optimization (NA (unoptimized)), with only server resource optimization (NA (optimized)), with only bearer aggregation (PA–SGW (K=64, unoptimized)) and with both bearer aggregation and server resource optimization (PA–SGW (K=64, optimized)). Note that in this evaluation, we utilize only PA–SGW as the bearer aggregation method. As can be seen from Fig. 9, when comparing NA (unoptimized) and NA (optimized), performance is improved by 32.8% with server resource optimization. Comparing NA (unoptimized) and PA–SGW (K = 64, unoptimized) indicates that the bearer aggregation increases network capacity by 91.8%. On the other hand, by comparing NA (unoptimized) and PA–SGW (K = 64, unoptimized), we can observe that combining server resource optimization and bearer aggregation improves the network capacity by 264.5%, which is much greater than the performance improvement by one of both methods. This means that a higher performance gain can be achieved by combining these methods. This difference arises from the amount of resources that can be allocated to the bottleneck node. Server resource optimization without aggregation cannot greatly reduce the load on the bottleneck nodes, and the server resources that can be allocated to the bottleneck nodes are limited. However, the combination of server resource optimization and bearer aggregation greatly reduces the load on the nodes in the network and increases the amount of resources that can be allocated to the bottleneck node.

In the following evaluations, server resource optimization is applied.

6.2.4 Aggregation Point

Fig. 10 shows a comparison of the performance of bearer aggregation at SGW and that at eNodeB. Figs. 10(a) and 10(b) plot the results of pre-determined aggregation and on-demand aggregation, respectively.

As shown in Fig. 10(a), with pre-determined aggregation, the aggregation at eNodeB outperforms that at SGW in terms of network capacity and bearer establishment time. This is because the aggregation at eNodeB can largely reduce the number of bearers and the corresponding signaling

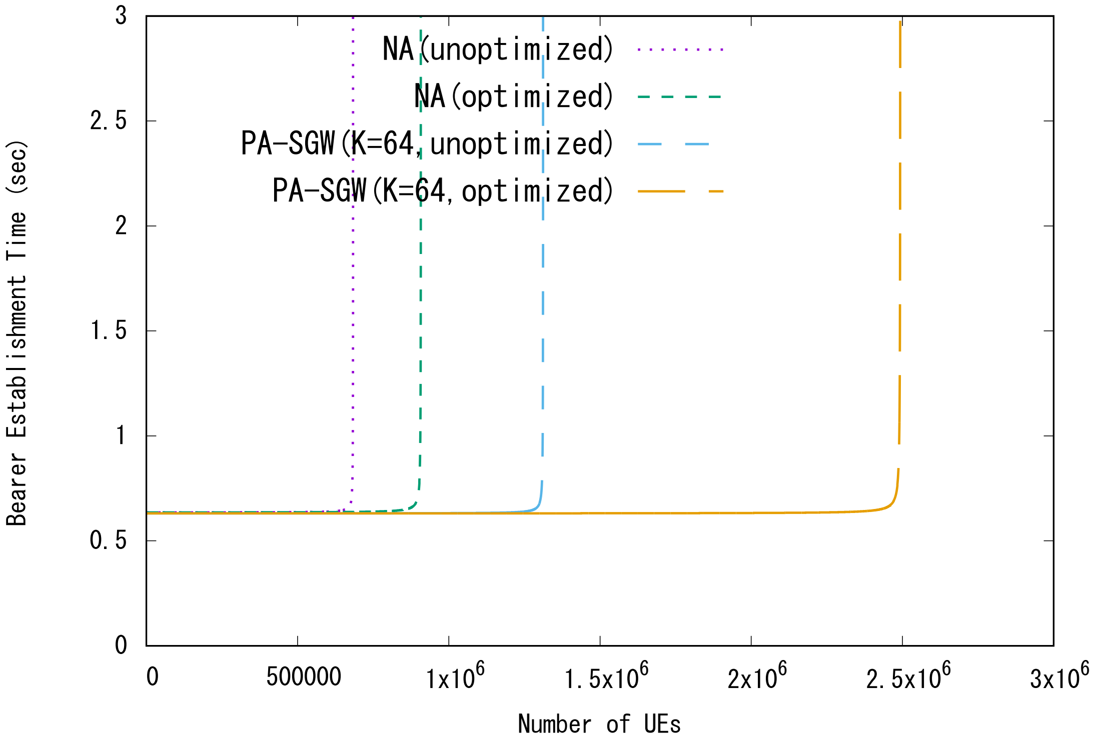
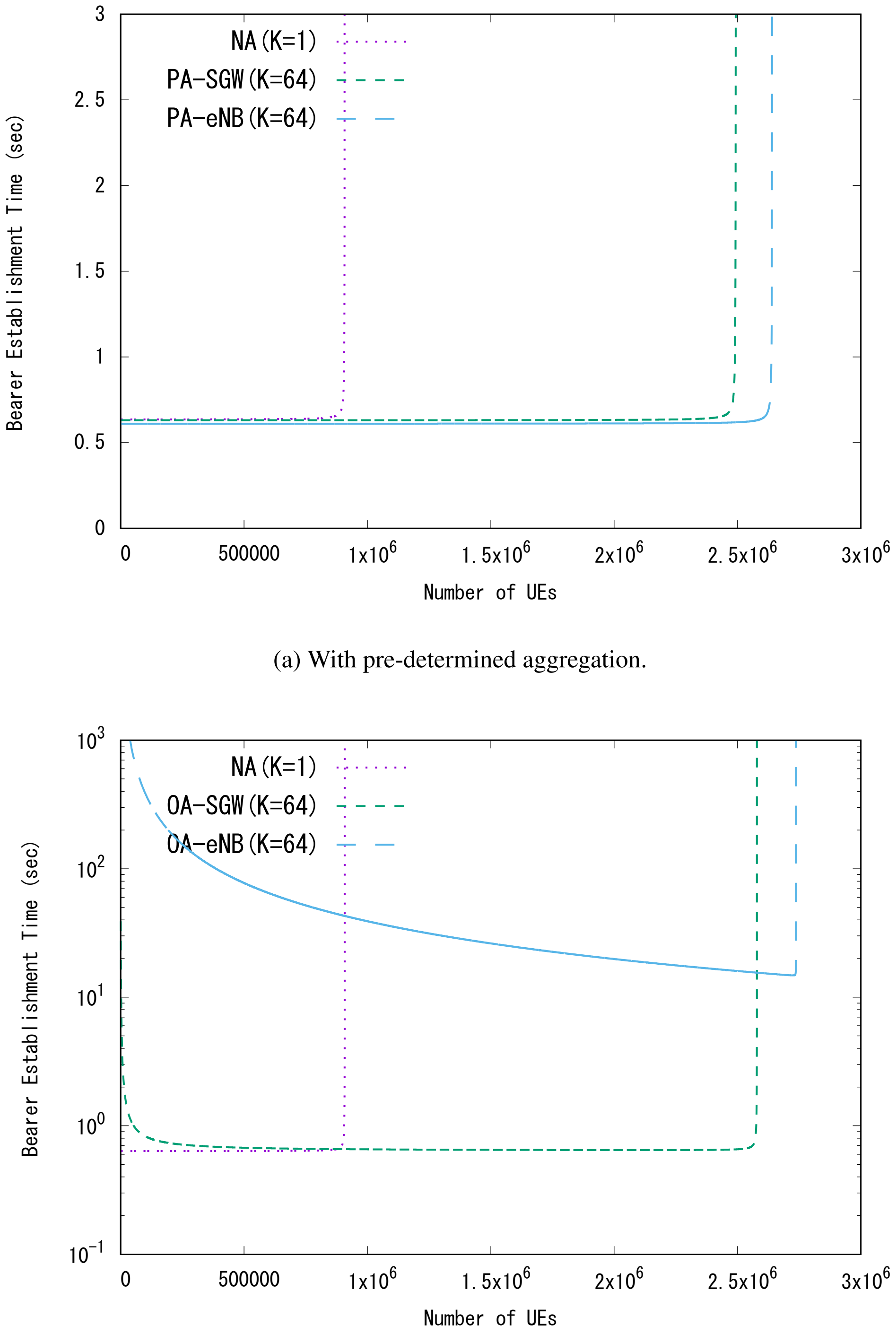


Figure 9: Effect of combination of server resource optimization and bearer aggregation



(a) With pre-determined aggregation.

(b) With on-demand aggregation.

Figure 10: Effect of aggregation point

messages compared to the aggregation at SGW. Note that in this result, there is almost no difference of bearer establishment time between them. However, we confirmed that especially when the server resources of eNodeB are small, this difference becomes apparent. Fig. 10(b) shows that with on-demand aggregation, the aggregation at eNodeB also yields higher network capacity, albeit with a substantially longer bearer establishment time. This is because of the abovementioned long waiting time associated with the on-demand aggregation.

6.2.5 Aggregation Timing

Fig. 11 shows similar results for the sake of comparing the pre-determined aggregation and on-demand aggregation schemes. Figs. 11(a) and 11(b) show plots of the results of aggregation at SGW and eNodeB, respectively.

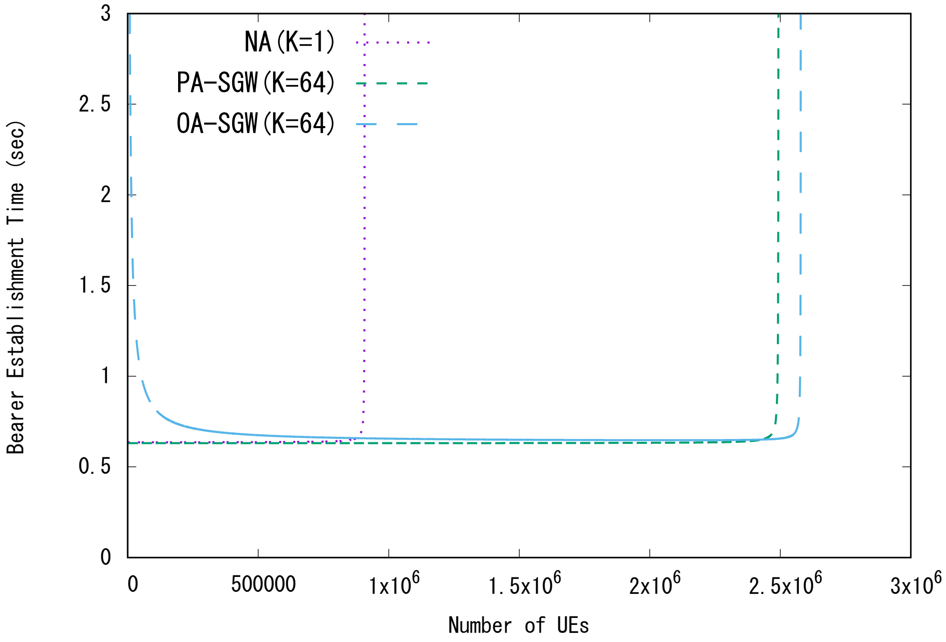
As shown in Fig. 11(a), the pre-determined aggregation scheme yields a slightly shorter bearer establishment time at the cost of lower network capacity.

When the number of accommodated UEs is small, the bearer establishment time with the on-demand aggregation scheme is long owing to the waiting time. In contrast, Fig. 11(b) shows that with the aggregation at eNodeB, the negative effect of the waiting time becomes apparent when the on-demand aggregation scheme is applied. This is because of the longer waiting time, as discussed in Subsection 6.2.4.

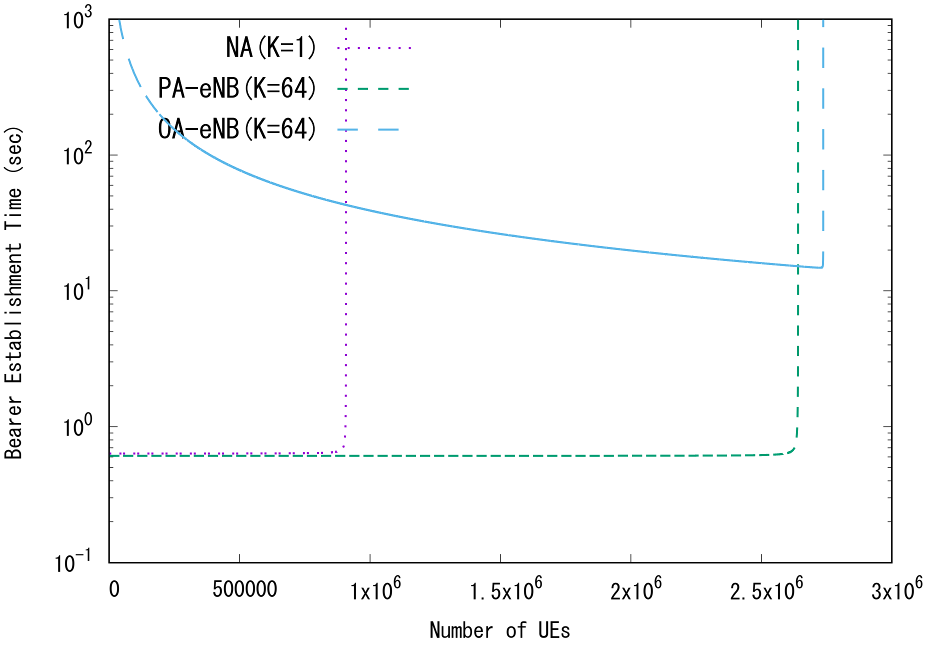
6.2.6 Aggregation at both SGW and eNodeB

Fig. 12 shows the effects of bearer aggregation at both SGW and eNodeB when applying the pre-determined aggregation scheme. In the figure, KB:KG=i:j indicates the results when the aggregation level at eNodeB is set to i and that at SGW is set to j.

According to the above results, to increase the network capacity, we would increase the aggregation level at eNodeB or SGW. For example, assume that the current aggregation level at eNodeB and SGW is 1:8. When we increase the aggregation level at either eNodeB or SGW by eight times, the aggregation level changes to 8:8 or 1:64, respectively. Fig. 12 plots these cases. We can observe that the performance achieved with 8:8 is better than that achieved with 1:64. On the other hand, when the current aggregation level is 8:1 and we want to increase the aggregation level at only one node, the performances achieved with 8:8 and 64:1 are similar. This is caused mainly



(a) With the bearer aggregation at SGW



(b) With the bearer aggregation at eNodeB

Figure 11: Effect of aggregation timing

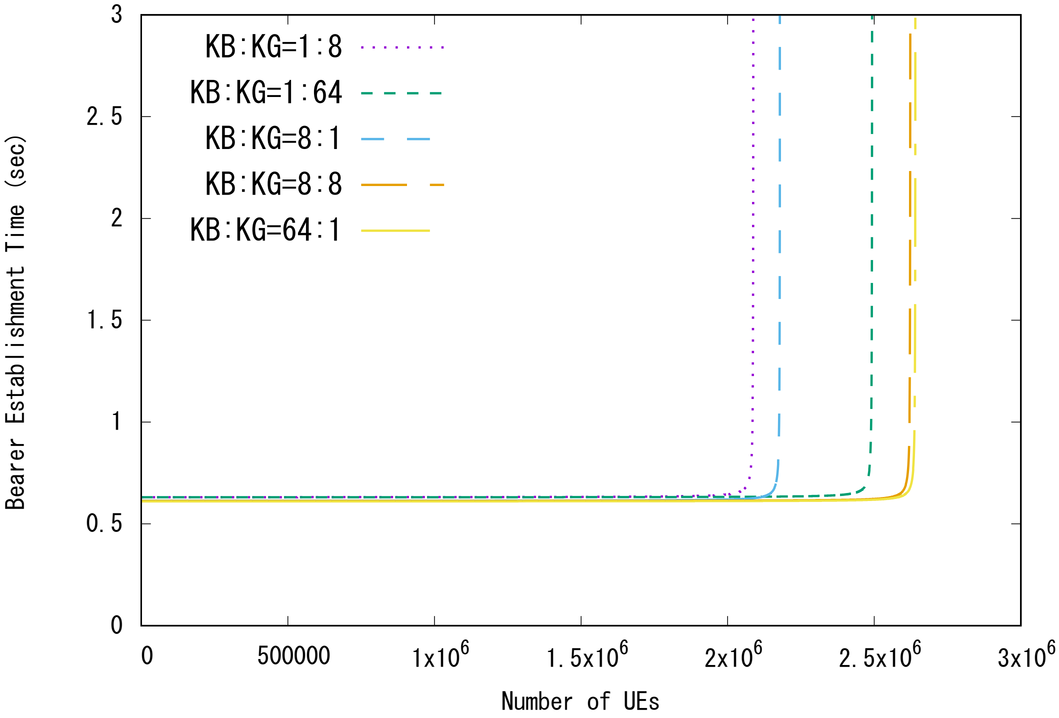


Figure 12: Effect of aggregation at both eNodeB and SGW

by the difference in the effect of bearer aggregation at eNodeB and SGW. That is, the aggregation at eNodeB decreases the number of S1-u and S5/S8 bearers, while that at SGW decreases only the number of S5/S8 bearers. The second reason is that the performance gained by increasing the aggregation level from 1 to 8 is significantly larger than that gained by increasing the aggregation level from 8 to 64.

7 Discussion

7.1 Server Resource Optimization

From the results in Subsection 6.2.3, server resource optimization can improve network capacity, regardless of whether bearer aggregation is applied. This means server resource optimization in a cloud environment is fundamentally advantageous.

7.2 Aggregation Level

In the aggregation at eNodeB, an aggregated S1–u bearer remains established until all UEs in the aggregated bearer become idle. Also, when the S1–u bearer is released, the additional signaling procedure is required. From this perspective, to decrease the signaling procedure in the mobile core network, the aggregation level should be determined so that at least one UE in the aggregated bearer being active, that depends on the communication characteristics of UEs such as communication cycle and the degree of communication synchronization of UEs.

7.3 Aggregation Point

The results described in Subsection 6.2.4 demonstrate that bearer aggregation at eNodeB outperforms that at SGW in terms of network capacity. This is because bearer aggregation at SGW reduces only the number of S5/S8 bearers, while that at eNodeB reduces the numbers of both S1–u and S5/S8 bearers. For the same reason, as shown in Fig. 10(a), when applying the predetermined aggregation scheme, the bearer establishment time with the bearer aggregation at eNodeB is smaller than that at SGW. For supporting these discussions, Fig. 13 shows plots of changes in the total processing times of signaling messages (T_t in (9)) with the on-demand aggregation scheme as a function of the number of accommodated UEs. The figure shows that the aggregation at eNodeB has a smaller total processing time than that at SGW. However, especially when the number of UEs is small in the on-demand aggregation scheme, the effect of waiting time (T_w in (9)) on the bearer establishment time is stronger, as shown in Fig. 10(b).

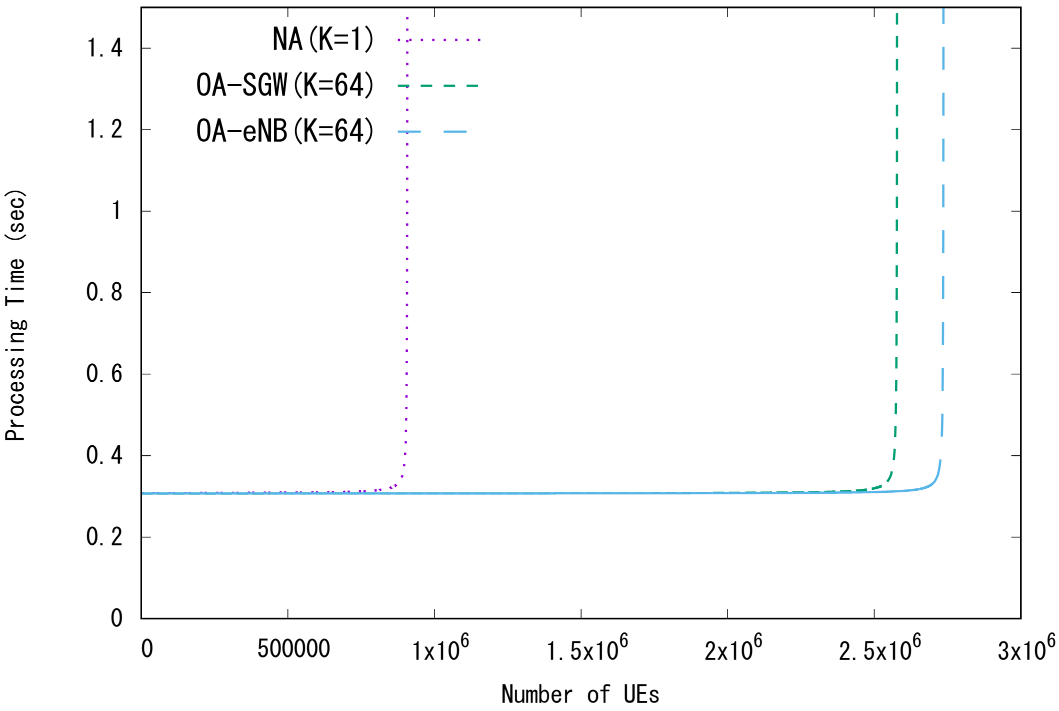


Figure 13: Processing time comparison

7.4 Aggregation Timing

The results described in Subsection 6.2.5 demonstrate that the on-demand bearer aggregation scheme yields a larger network capacity than the pre-determined bearer aggregation scheme. The main reason is that the pre-determined aggregation scheme requires a data path setting for each UE, while the on-demand aggregation scheme requires only one setting for a group of UEs. On the other hand, the on-demand aggregation method increases the MME load owing to the process of determining vIMSI and the corresponding shared bearer for a group of UEs at the start of communication. However, because the amount of the overhead is inversely proportional to the aggregation level, when the aggregation level exceeds a certain value, the total load on the nodes located in the cloud environment (MME and SGW/PGW$_c$) decreases.

The difference between the pre-determined and on-demand aggregation schemes affects the utilization of the shared bearers. Given that UEs in a certain pre-determined group do not always communicate simultaneously, utilization of the shared bearers with the pre-determined aggregation scheme varies according to the UEs' communication frequency. By contrast, with the on-demand aggregation scheme, utilization of each shared bearer is always high because shared bearers are established only for active UEs.

7.5 UE's mobility

In this thesis, we assume that UEs do not have any mobility and no handover occurs. When the mobility of UEs is considered, additional signaling messages are required for UEs leaving the current shared bearer, re-assigning a new shared bearer, and handling vIMSIs for handover UEs. From this viewpoint, the aggregation at SGW is preferable because it does not affect the handover procedure while the aggregation at eNodeB significantly affects the signaling procedure. Furthermore, utilization of the shared bearer would degrade owing to UE handover because the number of UEs in the shared bearer decreases.

7.6 Preferred bearer aggregation settings

From the above discussions, we can determine the recommended combinations of aggregation point and timing depending on delay constraints of the M2M/IoT applications, the number of UEs and their mobility. Table 4 summarizes the relationships among the characteristics of UEs, preferable aggregation point and timing, modifications required for EPC nodes, and resulting bearer establishment times and network capacities. Note that the aggregation level is not included in this table because the ideal aggregation level is not affected by the total number of UEs and their mobility.

7.7 Aggregation at both SGW and eNodeB

Finally, the bearer aggregation at both SGW and eNodeB inherits the characteristics of the aggregations at SGW and eNodeB. The performance gain differs depending on the current aggregation level at each aggregation point.

7.8 M2M/IoT communication

C/U plane separation and bearer aggregation are not specialized for M2M/IoT communication and we can apply the methods to conventional user communication. However, while UEs communicate large data in the user communication, many M2M/IoT UEs transmit and receive small data. Because of that, for such M2M/IoT communication, processing of the C-plane are significant in terms of the load on a mobile core network. We think that C/U plane separation and bearer aggregation are more suitable for M2M/IoT communication because these methods used in this thesis can reduce the load on C-plane.

Table 4: Recommended setting and obtained performance

UEs' characteristics	Aggregation point	Aggregation timing	Required modification	Bearer establishment time	Network capacity
high mobility	SGW	pre-determined	small (MME)	large	low
massive, high mobility	SGW	on-demand	small (MME)	large	medium
low latency, low/no mobility	eNodeB	pre-determined	large (UE, eNodeB and MME)	small	high

8 Conclusion

In this thesis, we evaluated the performance of a mobile core network with node virtualization and C/U plane separation based on SDN. We proposed a bearer aggregation method that decreases the signaling overhead, which is very important from the viewpoint of using massive M2M/IoT terminals.

The main results of this study are as follows.

(1) We developed a detailed algorithm and a signaling procedure for the bearer aggregation method (Section 4).

(2) We presented an analysis for evaluating the performance of mobile core networks (Section 5).

(3) We presented numerical results showing that the network capacity can be increased by up to 32.8% with node virtualization and C/U plane separation (Fig. 8(a)).

(4) We showed the simultaneous application of server resource optimization and bearer aggregation with appropriate aggregation point and timing yields 201.4% larger network capacity than applying only server resource optimization (Figs. 10, 11 and 9).

(5) We discussed appropriate settings for the aggregation method in accordance with the characteristics of M2M/IoT terminals (Table 4).

We consider that implementation experiments of the proposed methods are required to confirm their effectiveness on the actual environment. We are currently constructing the proof-of-concept implementation of the proposed methods. In detail, we plan to apply C/U plane separation and bearer aggregation methods to OAI and conduct experimentations where massive UEs connects to the mobile network simultaneously. In future work, we plan to evaluate the delay of LTE random access which is ignored in this thesis and the results brings the possibility of evaluating the capacity of the mobile core network with virtualization and C/U plane separation at eNodeB. Moreover, we will evaluate LPWANs which utilize the existing mobile network infrastructure such as eMTC and NB-IoT. In addition, we will extend our discussion to compare conventional bearer-based mobile core networks with packet-routing-based, i.e., GTP-less networks that do not use bearers.

Acknowledgment

I want to thank Professor Morito Matsuoka. He taught my attitude towards research and also gave me support and useful comments in various situations. And I would like to show my greatest appreciation to Professor Masayuki Murata. He gives me insightful comments and suggestions. And in my research, Associate Professor Go Hasegawa always gives me constructive comments and warm encouragement. It is thanks to him that I remember my interest and positive feelings in research and I am motivated to develop the research. I sincerely thank you. I would like to express my gratitude to Assistant Professor Yuya Tarutani. I sometimes asks him about my research and life in the laboratory. He always gives me useful comments. Special thanks also to students of Matsuoka Laboratory for their support of my laboratory life. I thank Wataru Takita, Kazuaki Obana, Motoshi Tamura, and Ryosuke Kurebayashi from NTT DOCOMO, and Takashi Shimizu from NTT Network Innovation Laboratories for their helpful comments.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2347–2376, Jun. 2015.

[2] F. Ghavimi and H. H. Chen, "M2M communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 525–549, Oct. 2015.

[3] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, Jan. 2017.

[4] "3GPP Low Power Wide Area Technologies (LPWA)," White Paper, GSMA, pp. 1–19, 2016.

[5] J. Schlienz and D. Raddino, "Narrowband Internet of Things Whitepaper," White Paper, Rohde&Schwarz, pp. 1–42, 2016.

[6] Third Generation Partnership Project, TS45.820, V13.1.0, *Cellular System Support for Ultra-low Complexity and Low Throughput Internet of Things (CIoT)*, Nov. 2015.

[7] F. Yousaf, J. Lessmann, P. Loureiro, and S. Schmid, “SoftEPC, Dynamic Instantiation of Mobile Core Network Entities for Efficient Resource Utilization,” in *Proceedings of Communications (ICC), 2013 IEEE International Conference on*, Jun. 2013, pp. 3602–3606.

[8] X. An, F. Pianese, I. Widjaja, and U. G. Acer, “DMME: Virtualizing LTE Mobility Management,” in *Proceedings of 2011 IEEE 36th Conference on Local Computer Networks*, Oct. 2011, pp. 528–536.

[9] ——, “DMME: A Distributed LTE Mobility Management Entity,” *Bell Labs Technical Journal*, vol. 17, no. 2, pp. 97–120, Sep. 2012.

[10] G. Premsankar, K. Ahokas, and S. Luukkainen, “Design and Implementation of a Distributed Mobility Management Entity on OpenStack,” in *Proceedings of 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, Nov. 2015, pp. 487–490.

[11] Z. A. Qazi, V. Sekar, and S. R. Das, “A Framework to Quantify the Benefits of Network Functions Virtualization in Cellular Networks,” *CoRR*, vol. abs/1406.5634, pp. 1–6, Jul. 2014.

[12] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, “NFV: State of the Art, Challenges, and Implementation in Next Generation Mobile Networks (vEPC),” *IEEE Network*, vol. 28, no. 6, pp. 18–26, Nov. 2014.

[13] M. R. Sama, S. Ben, H. Said, K. Guillouard, and L. Suciu, “Enabling Network Programmability in LTE / EPC Architecture Using OpenFlow,” in *Proceedings of Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014 12th International Symposium on*. IEEE, May 2014, pp. 389–396.

[14] A. Jain, N. Sadagopan, S. K. Lohani, and M. Vutukuru, “A Comparison of SDN and NFV for Re-designing the LTE Packet Core,” in *Proceedings of Network Function Virtualization and Software Defined Networks (NFV-SDN), IEEE Conference on*. IEEE, Nov. 2016, pp. 74–80.

[15] A. Tawbeh, H. Safa, and A. R. Dhaini, “A Hybrid SDN/NFV Architecture for Future LTE Networks,” in *Proceedings of 2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[16] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE Mobile Core Gateways, The Functions Placement Problem," in *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, & Challenges*. ACM New York, NY, USA, Aug. 2014, pp. 33–38.

[17] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E.-D. Schmidt, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," in *Proceedings of 2013 IEEE SDN for Future Networks and Services (SDN4FNS)*, Nov. 2013, pp. 8–14b.

[18] V. Nagendra, H. Sharma, A. Chakraborty, and S. R. Das, "LTE-Xtend: Scalable Support of M2M Devices in Cellular Packet Core," in *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, Oct. 2016, pp. 43–48.

[19] G. Hasegawa and M. Murata, "Joint Bearer Aggregation and Control-Data Plane Separation in LTE EPC for Increasing M2M Communication Capacity," in *Proceedings of 2015 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2015, pp. 1–6.

[20] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, "SDN/NFV-Based Mobile Packet Core Network Architectures: A Survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1567–1602, thirdquarter 2017.

[21] "OpenAirInterface," available at http://www.openairinterface.org/.

[22] Y. Tian, K. L. Lee, C. Lim, and A. Nirmalathas, "Performance Evaluation of CoMP for Downlink 60-GHz Radio-over-fiber Fronthaul," in *Proceedings of 2017 International Topical Meeting on Microwave Photonics (MWP)*, Oct. 2017, pp. 1–4.

[23] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks – A Technology Overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[24] Third Generation Partnership Project, TS23.720, V13.0.0, *Study on Architecture Enhancements for Cellular Internet of Things*, Mar. 2016.

[25] "Information and Communications in Japan," White Paper, Ministry of Internal Affairs and Communications of Japan, pp. 1–86, 2016.