Design and performance evaluation of bearer aggregation method in mobile core network with C/U plane separation

Shuya Abe
Graduate School of
Information Science and Technology
Osaka University
1–5 Yamadaoka, Suita,
Osaka 565–0871, Japan
Email: s-abe@ist.osaka-u.ac.jp

Go Hasegawa
Cybermedia Center
Osaka University
1–32, Machikaneyama-cho, Toyonaka,
Osaka 560–0043, Japan
Email: hasegawa@cmc.osaka-u.ac.jp

Masayuki Murata
Graduate School of
Information Science and Technology
Osaka University
1–5 Yamadaoka, Suita,
Osaka 565–0871, Japan
Email: murata@ist.osaka-u.ac.jp

***Abstract*—What is the best way to evaluate the capacity of a mobile core network with virtualization technologies and C/U plane separation using SDN? How can we increase the capacity, especially for accommodating massive M2M/IoT terminals? With increasing demand for cellular networks, enhancing the capacity of the mobile core networks is an urgent issue. In particular, when it comes to accommodating M2M/IoT terminals for cellular networks, the increasing load on the control plane of the mobile core network, as well as user plane, becomes a serious problem. While applying virtualization technologies such as SDN and NFV is one possible solution, there are almost no existing works on numerical or concrete evaluation of such solutions.**

In this paper, on the basis of mobile core networks with virtualized nodes and C/U plane separation, we first propose a bearer aggregation method for decreasing the control plane load to accommodate massive M2M/IoT terminals. We then show our mathematical analysis of the performance of mobile core networks based on a simple queuing theory. Specifically, we focus on the effect of the node virtualization and C/U plane separation and on the design parameters of the bearer aggregation.

The numerical evaluation results show that we can increase the capacity of the mobile core network by up to 32.8% with node virtualization and C/U plane separation, and by an additional 201.4% with bearer aggregation. We also explain that 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 UE is determined on the basis of application characteristics and the number of M2M/IoT terminals to be accommodated.

I. Introduction

Background. What is the best way to evaluate the capacity of a mobile core network with virtualization technologies and C/U plane separation using SDN? How can we increase the capacity, especially for accommodating massive M2M/IoT terminals? With increasing demand for cellular networks by rich user terminals such as smartphones and by massive M2M/IoT terminals [1], enhancing the capacity of the mobile core networks is an urgent issue [2]. Some M2M/IoT communications have different characteristics from rich user terminals—communication may occur periodically and 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 is received). Therefore, as more and more M2M/IoT terminals are accommodated to the cellular networks, the load on the mobile core networks increases, especially on the control plane nodes.

Related work. For these reasons, various methods for improving the capacity of M2M/IoT communications in the mobile core network have been proposed. These existing works, as well as the method proposed in this paper, are listed in Table I. Studies [3], [4] show that applying network function virtualization (NFV) to nodes of the mobile core network decreases costs and signaling traffic. In [5]–[8], applying software defined networks (SDN) to the mobile core network and the virtualization of the nodes in a cloud environment was studied. The effect of virtualization of the mobility management entity (MME) node, which is one of the mobile core network nodes, is evaluated in [5], and the signal processing load in the MME is evaluated in [6].

In these studies, the authors argued that the utilization of server resources can be improved and the cost can be decreased by virtualizing nodes of the mobile core network and applying SDN. However, signaling procedures for virtualized functional modules and SDN control messages may increase in the mobile core network. In particular, when accommodating an enormous number of M2M/IoT terminals, the overhead on the control plane nodes cannot be ignored, especially since such terminals may be synchronized when sending data. However, there has been almost no detailed evaluation of the trade-off between applying SDN and the increased signaling overhead, except that [7] evaluates the additional network traffic by introducing SDN to mobile core network. Also, in these works, the signal processing load is evaluated on the basis of the number and size of messages sent and received by the MME, but in actuality the processing load of a signaling message is

ISBN 978-3-901882-94-4 © 2017 IFIP

determined by many factors not necessarily related to just the message size.

Our previous studies. Our research group analytically evaluated the signaling overhead and the load on the mobile core network nodes in accommodating M2M/IoT terminals when applying SDN and the bearer aggregation method at SGW [9]. However, in [9], no detailed algorithm for the bearer aggregation was determined. Furthermore, the bearer aggregation can be applied to eNodeB, which differently affects the aggregation efficiency and the additional overhead on mobile core networks.

Contribution. The main contribution of this paper is as follows.

1) Determine the detailed algorithm and signaling procedure for the bearer aggregation method.
2) Propose an analysis method for evaluating the performance of the mobile core networks on accommodating massive M2M/IoT terminals.
3) Numerically evaluate the performance of the mobile core network with node virtualization and C/U plane separation with SDN.
4) Numerically evaluate the effect of the bearer aggregation method.
5) Parameter design according to the characteristics of M2M/IoT terminals.

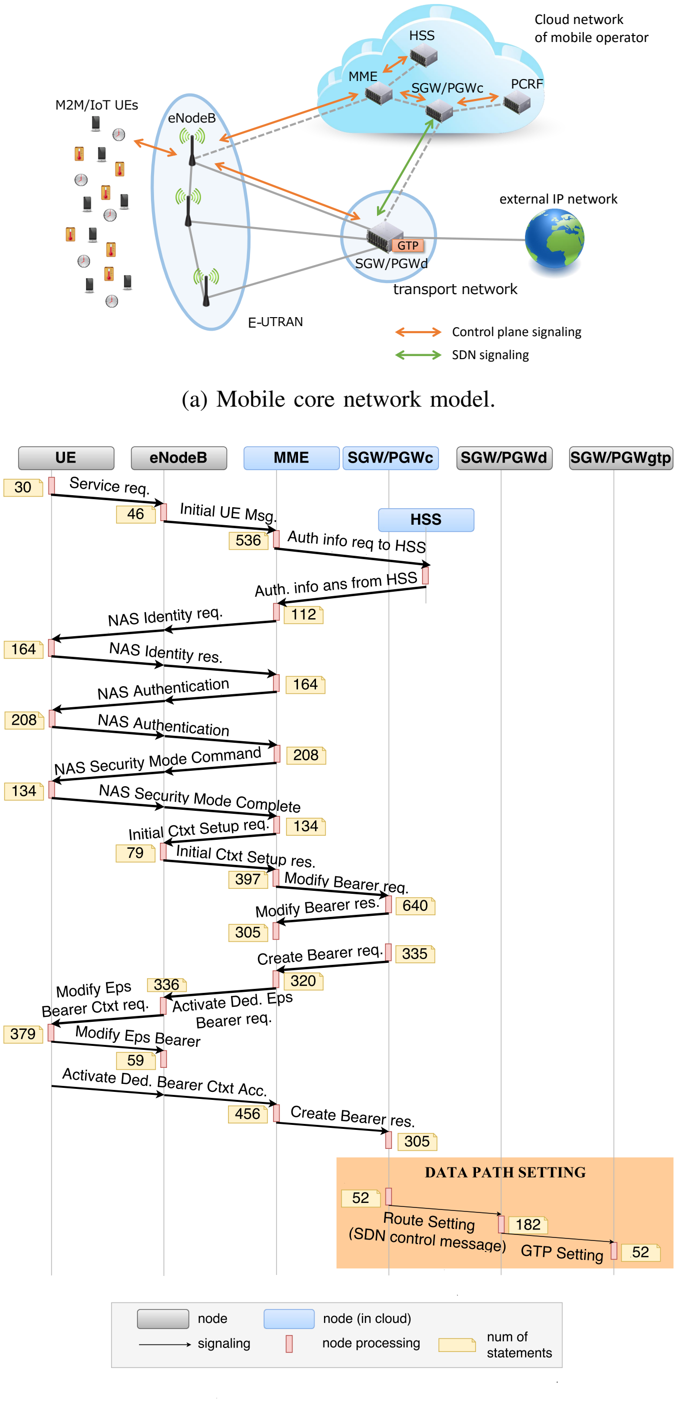
Outline. In Section II of this paper, we explain the model of the mobile core network for performance evaluation. In Section III, we discuss the details of a bearer aggregation method. In Section IV, we present our mathematical analysis of the performance of the mobile core network. In Section V, we provide extensive numerical evaluation results, and in Section VI, we present our discussion based on these results. We conclude in Section VII with a brief summary and mention of future work.

II. Network Model

Here, we explain the network model and signaling flow in the model.

A model of the mobile core network is shown in Fig. 1(a). This model consists of user equipments (UEs), eNodeBs, SGW/PGW, MME, a home subscriber server (HSS), and a policy and charging rules function (PCRF). We assume that MME, HSS, and PCRF, which are control plane nodes in the mobile core network, are located in the cloud environment. SGW/PGW is separated into SGW/PGWc and SGW/PGWd, which correspond to a control plane node and a user plane node, respectively. SGW/PGWc is installed in the cloud environment and SGW/PGWd is located at the transport network. SGW/PGWd has a GTP module [8] that is a matching function of the GPRS tunneling protocol (GTP) bearers established at the SGW. This module can prevent data packets from/to UEs from passing through the SGW/PGWc in the cloud environment [8]. In these ways, we applied C/U separation to the mobile core network.

Figure 1(b) shows the signaling flow when a UE changes its state from idle to active and makes a request to start



(a) Mobile core network model.

(b) Signaling flow for bearer establishment.

Fig. 1: (a) **Network Model:** Assuming that node virtualization and C/U plane separation with SDN are applied. (b) **Signaling Flow:** Required for a bearer establishment by each UE.

a communication. In this figure, req. and res. mean the signalling message is request and responce messages, respectively. and Msg. stands for “message”. Ctxt, Ded., Acc mean “Context”, “Dedicated”, and “Accept”, respectively. The figure includes the number of statements of programs executed by

TABLE I: **Comparison of methods:** The characteristics of various existing evaluations and the proposed method.

	[5], [6]	[7], [8]	Proposed
Qualitative discussion about applying SDN	✓	✓	✓
Evaluation of overhead due to virtualization and SDN	✓		✓
Evaluation considering signal processing load			✓

each node for processing the signaling messages. The number of statements was obtained by analyzing the source code of OpenAirInterface (OAI) [10], a software application of the LTE/EPC network written in C. The last two signaling messages ("Route Setting" and "GTP Setting") cannot be found in the original signaling flow but are required for data path setting and for achieving the matching of an S1 bearer and an S5/S8 bearer at SGW/PGWd. We determine the number of statements for processing these messages from similar procedures in OAI.

As shown in Fig. 1, when a UE starts the communication, many signaling messages are exchanged between the control plane nodes. As a result, a bearer between eNodeB and SGW (S1–u bearer) and between SGW and PGW (S5/S8 bearer) are established for sending and receiving data packets by the UE. Since these bearers are established for each UE, as the number of M2M/IoT communications increases, the load on the mobile core network would increase.

III. Bearer Aggregation Method

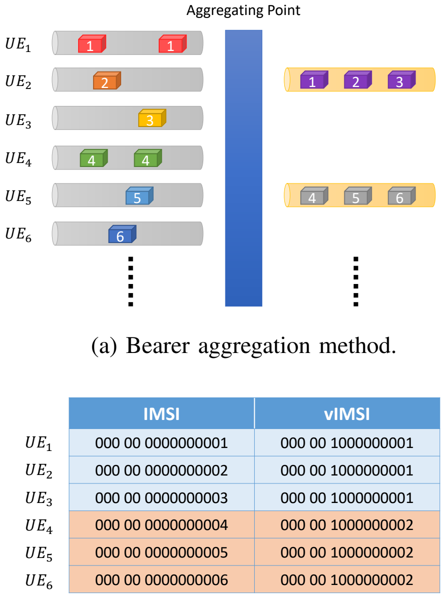
A. Overview

The bearer aggregation method reduces the load on mobile core network nodes by having one bearer shared by multiple UEsa direct contrast to the current mobile core networks, where a single bearer corresponds to a single UE.

An illustration of the bearer aggregation method is given in Fig. 2(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. 2(a), when the aggregation point is a 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 CPU utilization for handling signaling messages and the memory usage of the node are reduced by decreasing the number of concurrent bearers at the node.

B. Virtual IMSI

In order to realize the bearer aggregation method, we introduce the concept of a *virtual International Mobile Subscriber Identity (vIMSI)* that associates with a shared bearer, in contrast to a normal IMSI, which is assigned uniquely to each UE and corresponding bearer. MME handles the matching between IMSIs and vIMSIs by maintaining an *IMSI table* (Fig. 2(b)) that represents the current status of the bearer aggregation.



(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.

Fig. 2: (a) **Bearer Aggregation:** Bearers from a group of UEs are aggregated into a single shared bearer. (b) **IMSI table:** MME handles the matching between IMSIs and a vIMSI.

In the signaling flow shown in Fig. 1(b), from when the flow begins to 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 to locate a vIMSI that corresponds to the UE. Then, in the following signaling flow, signaling messages include both of the normal IMSI for the UE and the vIMSI for the shared bearer. In addition, the MME notifies PCRF of the correspondence between the IMSIs and vIMSIs when it updates the IMSI table.

C. Design Options

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

1) Aggregation Points:

a) Aggregation at SGW: Multiple S1–u bearers between an eNodeB and a SGW are aggregated into a single S5/S8 bearer between the SGW and a PGW. The number of Modify Bearer req/res. messages and Create Bearer req/res. messages for creating S5/S8 bearers then decreases.

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

b) Aggregation at eNodeB: Multiple radio bearers between UEs and an eNodeB are aggregated into a single S1–u bearer between the eNodeB and a SGW. The number of S5/S8 bearers is also reduced, as an S1–u bearer and an S5/S8 bearer have a one-to-one relationship. Consequently, the reduction of the signaling overhead by the aggregation at 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 corresponding radio bearer are released simultaneously when a UE becomes idle. In contrast, when the aggregation at eNodeB is applied, a shared S1–u bearer should be maintained until all UEs in the group for the shared bearer become idle.

2) Aggregation Timing:

a) 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.

Figure 3(a) shows a timeline of the signal processing at the MME with a pre-determined aggregation method. Vertical dashed lines represent the 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 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 is necessary for each UE.

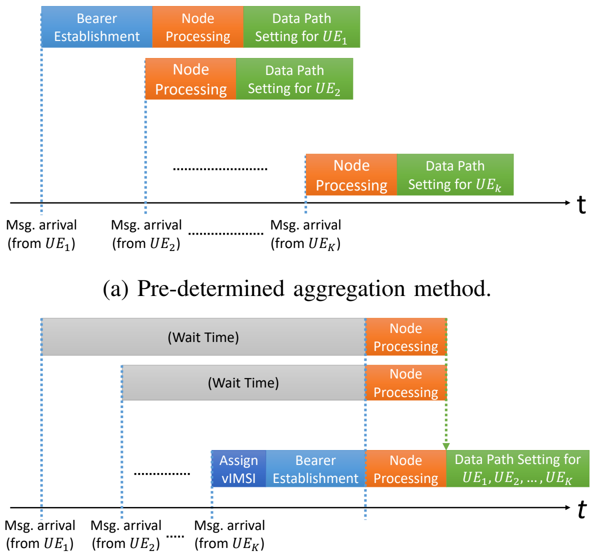
b) On-demand Aggregation: The group of a UE for bearer aggregation is determined when the UE becomes active and the communication request is issued, not when the UE attaches to the network. Therefore, notification to the PCRF occurs every time the UE begins the communication. Figure 3(b) shows an example of the on-demand aggregation method in action. Each UE waits for the communication requests from all UEs in the group to arrive. Then the assignment of vIMSI, notification to PCRF, and the establishment of the shared bearer are conducted. Note that this method requires only one data path setting procedure for the whole group of UEs. However, UEs experience a waiting time from when the communication request occurs to when the shared bearer is established.

IV. Performance Analysis

In our performance analysis, we calculate the average of *the bearer establishment time*, which is defined as the time when the signaling flow starts to when it ends, as shown in Fig. 1(b), assuming the network model in Fig. 1(a).

A. Notations

N is a general representation of a node and $\mathbb{V}$ is a set of nodes in the network model. For the individual node, we abbreviate UE, eNodeB, MME, and SGW/PGW as U, B, M, and G, respectively. 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 in Fig. 1(b). 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 performance of node N in terms of the number of statements that can be processed per second. L_{N_i} is the number of statements for processing the ith signaling message at node N. Note that N, N_1, and N_2 mean one of U, B, M, G_c, G_d, and G_g.


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 $UE_1, UE_2, ..., UE_K$
t
Msg. arrival (from UE_1)
Msg. arrival (from UE_2)
Msg. arrival (from UE_K)

(b) On-demand aggregation method.

Fig. 3: (a) **Pre-determined Aggregation:** The first UE (UE_1) establishes a shared bearer. (b) **On-demand Aggregation:** A shared bearer is established when the communication request from all UEs in a group arrives.

We assume that each UE starts the communication at regular intervals of D, which is called a communication period of a UE. K represents aggregation level in the bearer aggregation, which means the number of UEs in each group ($K = 3$ in Fig. 2).

B. Bearer Establishment Time

The bearer establishment time T is the sum of propagation delay of all signaling messages T_τ, the processing time for all messages T_t, and the waiting time required when using on-demand aggregation T_w. We derive the bearer establishment time by Eq. (1), as Equation (1).

$$\begin{aligned} T &= T_\tau + T_t + T_w \\ &= \sum_{N_1,N_2 \in \mathbb{V}} (C_{N_1,N_2} \tau_{N_1,N_2}) + \sum_{N \in \mathbb{V}} (P_N t_N) + T_w, \end{aligned} \quad (1)$$

where T_w is calculated by Eq. (2) on the basis of the communication period of a UE, the number of UEs attached to the network, and the aggregation level.

$$T_w = \begin{cases} \frac{KD}{2n_U} & \text{(Aggregation at SGW)} \\ \frac{KDn_B}{2n_U} & \text{(Aggregation at eNodeB)} \end{cases} \tag{2}$$

C. Processing Time

To derive the processing time t_N, we exploit the M/G/1/PS queuing model. It means that UEs have ON period in which they send a communication request and OFF period in which they do not send it. The ON periods occur periodically, and the distribution with which the signaling messages arrive at the nodes follows a Poisson process. In the M/G/1/PS model, the mean sojourn time $E[R]$ can be derived as

$$E[R] = \frac{\rho^r}{1-\rho}\frac{E[S^2]}{2E[S]} + \frac{1-\rho^r}{1-\rho}E[S], \tag{3}$$

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, and $\rho = \lambda E[S]$ is the system utilization.

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

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

V. Numerical Evaluation

In this section, we show the numerical results of the analysis in Section IV for evaluating the effect of the bearer aggregation method discussed in Section III.

A. Evaluation Candidates and Parameter Settings

We evaluate the performance of four different bearer aggregation methods that combine aggregation point and aggregation timing. For comparison purposes, we also 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

The communication period of a UE is set to 600 seconds. The network model has 2,000 eNodeBs, one MME, one SGW/PGWc, one SGW/PGWd, and one GTP module. We change the number of UEs to be attached to the network while each eNodeB has an identical number of UEs to be accommodated. 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: 20 msec
- eNodeB–SGW/PGWd: 7.5 msec
- eNodeB–MME, SGW/PGWc: 10 msec
- SGW/PGWd–SGW/PGWc: 10 msec
- MME–SGW/PGWc: 1 msec
- GTP module–SGW/PGWd: 1 msec

The default values of the server performance of nodes in the network model are as follows.

- UE: 3,000 statements/sec
- eNodeB: 1,500 statements/sec
- MME: 3,000,000 statements/sec
- SGW/PGWc: 3,000,000 statements/sec
- SGW/PGWd: 3,000,000 statements/sec
- GTP module: 600,000 statements/sec

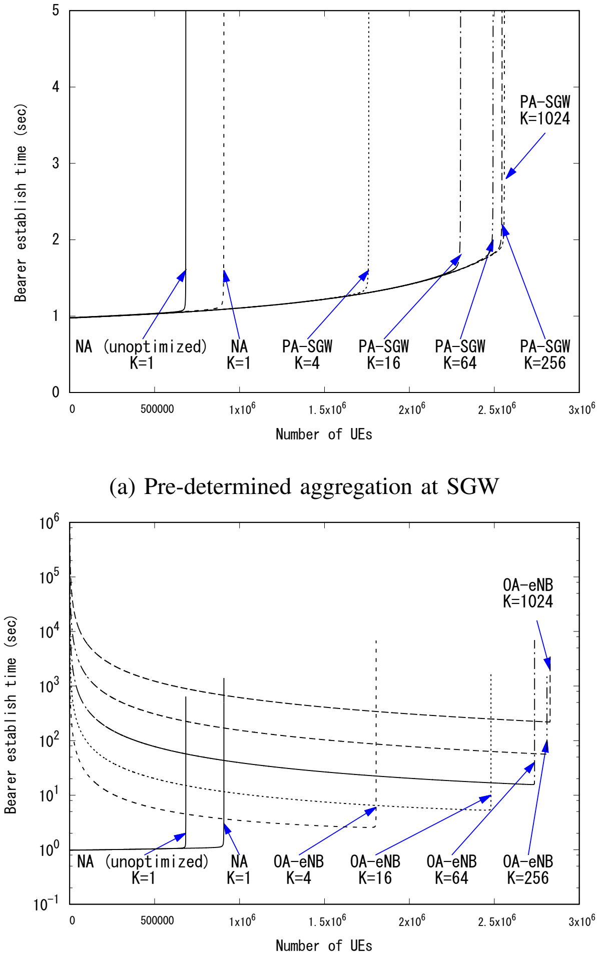
These values were determined on the basis of discussions with researchers from one of the mobile network operators in Japan, assuming a nation-wide mobile core network. Note that the performance of the UE is set to be enough large to ignore the influence on the bearer establishment time.

We assume that when we apply server virtualization, the server performance located in the cloud environment (MME and SGW/PGWc) is optimized so that the load of the servers becomes identical, while the sum of the server performance is kept unchanged from the default values mentioned above.

The number of statements for processing each signaling message in the signaling flow in Fig. 1(b) 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.

B. Evaluation Results

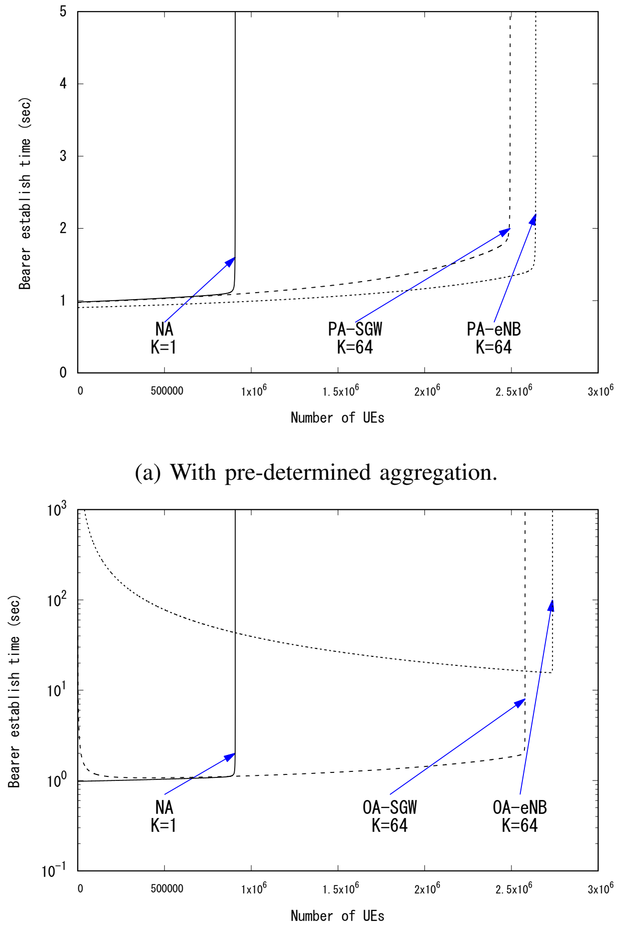
1) Effect of Aggregation Level: Figure 4 shows the relationship between the number of accommodated UEs and the bearer establishment time when the pre-determined aggregation at SGW and on-demand aggregation at eNodeB are applied. In the figure, K = i indicates the results when the aggregation level K is set to i. NA (unoptimized) means that the server performance of MME and SGW/PGWc is set to the default values and is not optimized. Other results are obtained with the optimization of server performance. 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 of one of the nodes in the network becomes 100%. In what follows, we use that number of UEs as the capacity of the network. When we compare NA (unoptimized) K=1 and NA K=1, we see that the network capacity increases by 32.8% with server performance optimization.



(a) Pre-determined aggregation at SGW

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

Fig. 4: **Evaluation Results:** The network capacity increases by applying a bearer aggregation method.



(a) With pre-determined aggregation.

(b) With on-demand aggregation.

Fig. 5: **Effect of aggregation point:** The aggregation at eNodeB increases the network capacity more than the aggregation at SGW.

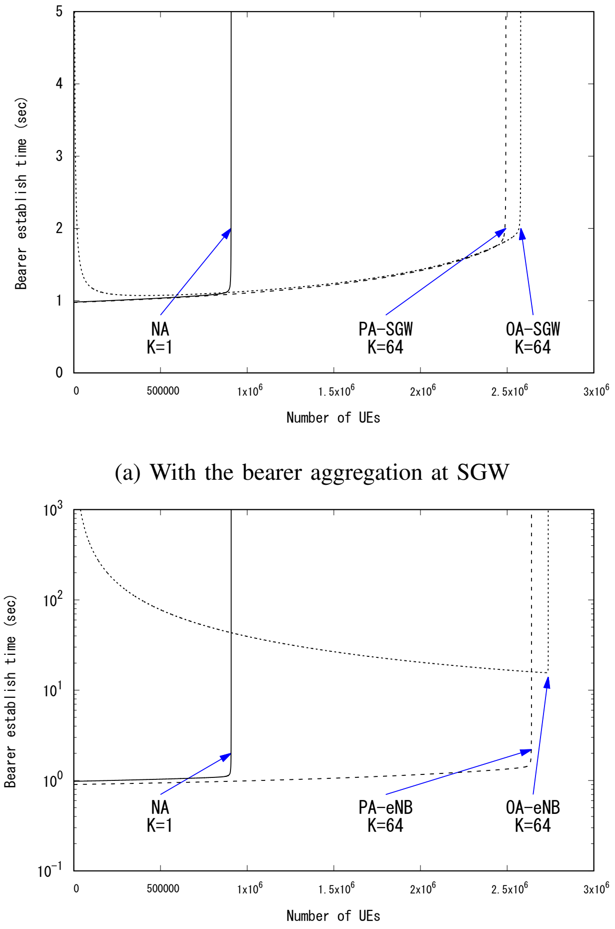
We can also see from the figure that the network capacity further increases by applying a bearer aggregation method ($K > 1$). The performance gain is up to 181.8% when we compare NA K=1 and PA-SGW K=1024. This is because the bearer aggregations reduce the number of signaling messages to be processed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

Figure 4(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. 3(b). Equation (2) 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 decreases or when the number of accommodated UEs increases, the bearer establishment time decreases.

2) Effect of Aggregation Point: Figure 5 shows the relationship between the number of accommodated UEs and the bearer establishment time to compare the performance of the bearer aggregation at SGW and the bearer aggregation at eNodeB. Figures 5(a) and 5(b) plot the results of the pre-determined aggregation and the on-demand aggregation, respectively.

As shown in Fig. 5(a), with the pre-determined aggregation, the aggregation at eNodeB outperforms the aggregation at SGW in terms of the network capacity and the bearer establishment time. This is because the aggregation at eNodeB can reduce the number of bearers and corresponding signaling messages more than the aggregation at SGW. Figure 5(b) shows that with the on-demand aggregation, the aggregation at eNodeB also gives higher network capacity, but with a much larger bearer establishment time. This is because of the long waiting time with the on-demand aggregation.

3) Effect of Aggregation Timing: Figure 6 shows similar results to compare the pre-determined aggregation and the on-



(a) With the bearer aggregation at SGW

(b) With the bearer aggregation at eNodeB

Fig. 6: **Effect of aggregation timing:** The on-demand aggregation outperforms the pre-determined aggregation in terms of network capacity.

demand aggregation. Figures 6(a) and 6(b) plot the results of the aggregation at SGW and at eNodeB, respectively.

As shown in Fig. 6(a), the pre-determined aggregation gives slightly smaller bearer establishment time but at the cost of a smaller network capacity. In addition, when the number of accommodated UEs is small, the bearer establishment time of the on-demand aggregation is large due to the waiting time. In contrast, Fig. 6(b) shows that with the aggregation at eNodeB, the negative effect of the waiting time becomes apparent when the on-demand aggregation is applied. This is because of the larger waiting time, as discussed in Subsection V-B2.

VI. Discussion

The results described in Subsection V-B2 demonstrate that the bearer aggregation at eNodeB outperforms the bearer aggregation at SGW in terms of the network capacity. This is because the bearer aggregation at SGW only reduces the number of S5/S8 bearers, while the bearer aggregation at eNodeB reduces the number of both S1–u and S5/S8 bearers. Therefore, the load of the network nodes is smaller when applying the bearer aggregation at eNodeB than when applying it at SGW. For the same reason, as shown in Fig. 5(a), with the pre-determined aggregation, the bearer establishment time is smaller when applying the aggregation at SGW than when applying it at eNodeB. For supporting these discussions, Fig. 7 shows the change in the total processing time for signaling messages (T_t in Eq. (1)) with the on-demand aggregation as a function of the number of accommodated UEs. We can see from the figure that the aggregation at eNodeB has a smaller total processing time than the aggregation at SGW. However, especially when the number of UEs is small in the on-demand aggregation, the effect of the waiting time has a larger effect on the bearer establishment time.

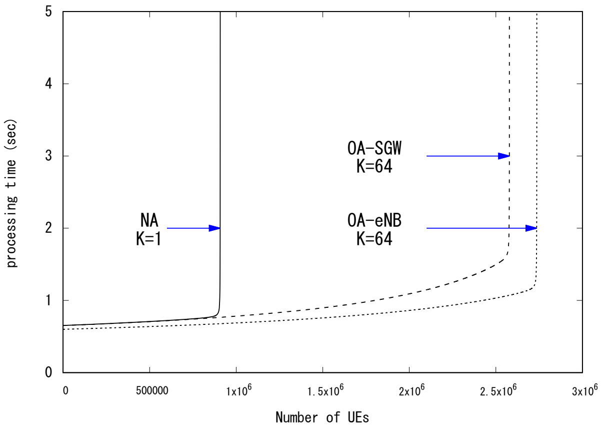


Fig. 7: **Processing time comparison:** The on-demand aggregation at eNodeB outperforms the on-determined aggregation at SGW in terms of processing time.

The results described in Subsection V-B3 demonstrate that the on-demand bearer aggregation has larger network capacity than the pre-determined bearer aggregation. This is because the pre-determined aggregation requires a data path setting for each UE, while the on-demand aggregation requires only one setting for a group of UEs. On the other hand, the on-demand aggregation method increases the MME load due to the process of determining vIMSI and 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 of the nodes located in the cloud environment (MME and SGW/PGWc) decreases.

The difference of the pre-determined and the on-demand aggregation affects the efficiency of the shared bearers. Since UEs in a certain group do not always communicate at the same time, the efficiency of the shared bearers with the pre-determined aggregation varies according to UE's communication frequency. In contrast, with the on-demand aggregation, we can achieve high efficiency since the shared bearer is established only when the number of UEs reaches the aggregation level.

TABLE II: **Recommended setting and obtained performance:** Varies according to the characteristics of UEs.

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/no mobility	eNodeB	pre-determined	large (UE, eNodeB and MME)	small	high

In this paper, we assume that UEs do not have any mobility and no handover occurs. When we consider the mobility of UEs, additional signaling messages are required for leaving the current shared bearer, re-assigning a new shared bearer, and handling corresponding vIMSIs for the handover UEs. From this viewpoint, the aggregation at SGW is preferable since it does not affect the handover procedure, while it significantly affects the signaling procedure of the aggregation at eNodeB, since UE's handover changes which eNodeB to connect. Also, the efficiency of the shared bearer would degrade, since the number of UEs in the shared bearer decreases due to the handover.

From the above discussion, we can determine the recommended combination of aggregation points and timing depending on the number of UEs and the mobility of UEs. Table II summarizes the relationships among the characteristics of UEs, preferable aggregation points and timing, required modification to the mobile core network nodes, and resulting bearer establishment time and network capacity. Note that this table does not include the aggregation level, since the desired value remains unchanged regardless of the aggregation method.

VII. Conclusion

In this paper, we evaluated the performance of a mobile core network with node virtualization and C/U plane separation with SDN. We further proposed a bearer aggregation method that decreases the signaling overhead, which is of particular value for massive M2M/IoT terminals.

The conclusions of this paper are as follows.

1) Determined a detailed algorithm and signaling procedure for the bearer aggregation method (Section III).
2) Presented an analysis for evaluating the performance of the mobile core networks (Section IV).
3) Exhibited numerical results showing that the network capacity is increased by up to 32.8% with node virtualization and C/U plane separation with SDN (Fig. 4(a)).
4) Numerically revealed that the network capacity is further enhanced by 201.4% by the bearer aggregation with appropriate aggregation point and timing (Figs. 4, 5, 6).
5) Discussed appropriate settings for the aggregation method in accordance with the characteristics of M2M/IoT terminals (Table II).

In future work, we plan to evaluate the effect of virtualization and C/U plane separation at eNodeB. We will also extend our discussion to compare the conventional bearer-based mobile core network with packet-routing-based networks that do not use bearers. We also plan to evaluate the effect of buffering overflow with on-demand aggregation and that of the size of buffer of the nodes, which are out of scope in this paper. Additionally, we are currently carrying out experimental evaluation to support the results mentioned in this paper.

Acknowledgment

The authors 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] 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, Jul. 2014.

[4] 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.

[5] I. Widjaja, P. Bosch, and H. La Roche, "Comparison of MME Signaling Loads for Long-Term-Evolution Architectures," in *Proceedings of IEEE Vehicular Technology Conference*. IEEE, Sep. 2009, pp. 1–5.

[6] 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.

[7] 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.

[8] 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.

[9] G. Hasegawa and M. Murata, "Joint bearer aggregation and control-data plane separation in lte epc for increasing m2m communication capacity," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2015, pp. 1–6.

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