Joint bearer aggregation and control-data plane separation in LTE EPC for increasing M2M communication capacity

Go Hasegawa
Cybermedia Center
Osaka University
1-32, Machikaneyama-cho, Toyonaka, Osaka 560-0043, Japan
Email: hasegawa@cmc.osaka-u.a.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*—In this paper, we propose a method for increasing the capacity of Machine-To-Machine (M2M) communication in mobile core networks. The proposed method combines two approaches: bearer aggregation inside mobile core networks for decreasing the load of Evolved Packet Core (EPC) nodes, and applying a Software Defined Networking (SDN) architecture to separate the control and data planes and aggregate control plane nodes in a cloud network environment for resource sharing. The combination of these two approaches is meaningful because they have a complementary relationship. We give a mathematical analysis and numerical results of a performance evaluation of the proposed method. The evaluation results show that we can increase the capacity of a mobile core network for M2M communication by around 30% when one of the two approaches is applied, while the performance gain increases up to 124% when both approaches are combined.**

I. Introduction

Due to the continuous increase of mobile phone users and the rapid popularization of rich terminals such as smartphones and tablets, congestion in data plane for transferring user data, and that in control plane for controlling radio access bearers and connections between User Equipments (UEs) and external networks, have become a serious problem in 3G and Long Term Evolution (LTE) mobile networks. There are some solutions for data plane congestion such as offloading to WiFi or other networks [1-5]. However, such solutions cannot alleviate control plane congestion because the number of UEs to be handled in the mobile core network remains unchanged with offloading.

Furthermore, some smartphone applications generate periodic communications to corresponding servers that heavily impact the control plane load [6]. Furthermore, Machine-to-Machine (M2M) communications via 2G/3G/LTE mobile networks are gaining increased attention as a new communication paradigm, driving further mobile network demand. The communication characteristics of M2M terminals significantly differs from those of traditional Human-to-Human (H2H) communications [7], in that the number of M2M terminals is much larger than the number of H2H terminals, while the amount of communication data per each M2M terminal is smaller and communications occur periodically or intermittently. Such communications by rich terminals and M2M terminals have a large impact on the control plane load when traditional method for accommodating H2H terminals. Furthermore, the Average Revenue Per User (ARPU) of M2M terminals would likely be substantially smaller than that of H2H terminals [8], meaning that we cannot recover the cost for accommodating M2M terminals within existing mobile cellular networks under current systems and cost structures.

Existing researches has focused on control plane congestion by M2M communications in mobile core networks [9, 10], and some solutions, such as light-weight signaling protocol, has been proposed [11, 12]. One way to decrease control plane load in a mobile core network is to decrease the number of handling bearers in the network, because the existing 3GPP-based mobile core network persistently maintains a few bearers for each UE, regardless of their traffic amount. Group-based communications [13, 14] can help decrease the number of bearers in a mobile core network because the number of connected UEs to the mobile core network decreases. However, such methods require UEs to have short-range wireless network equipment, such as WLAN or Bluetooth, that should be avoided in order to reduce the cost of M2M terminals.

There has also been recent researches on applying Software Defined Network (SDN) architectures to mobile core networks, aimed at decreasing their operational cost [15-18]. In such architectures, control and data planes on LTE Evolved Packet Core (EPC) nodes are separated and one or both planes are virtualized and located in a cloud network environment. This should allow efficient and flexible operation of mobile core networks, since it provides server and network resource sharing among multiple nodes and among multiple mobile core networks. At present, however, there are almost no detailed system designs and performance evaluations proposed in the literature.

In this paper, we address the control plane congestion problem in mobile core networks by combining bearer aggregation and control-data plane separation. We propose to aggregate bearers for M2M communications at SGW to decrease the number of handled bearers. For control-data plane separation, we separate the planes of a Servicing/Packet data network Gateway (SGW/PGW) and server resources are shared by the control-plane nodes of SGW and PGW and the Mobility Management Entity (MME) in the cloud network environment. As explained below, these two methods should be combined

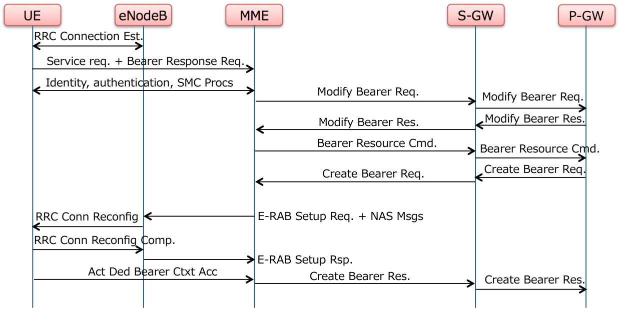


Fig. 1. Signaling procedure when a UE begins its communication

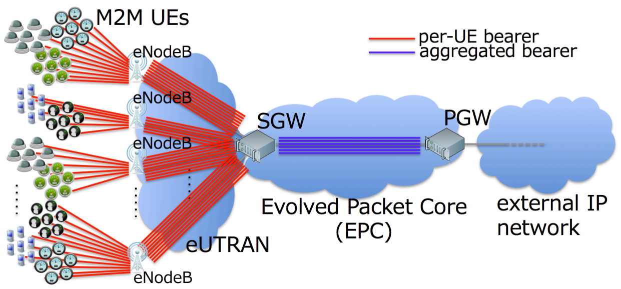


Fig. 2. Bearer aggregation at SGW

since they have a *complementary* relationship. We also give a performance analysis of the proposed method to confirm its effectiveness at increasing capacity in M2M communications.

The remainder of this paper is organized as follows. In Section II, we explain the motivation for this work. In Section III, we present the proposed method for combining bearer aggregation and control-data plane separation. We then give a performance analysis of the proposed method and numerical examples in Section IV. Finally, we conclude this paper with an overview of future work in Section V.

II. Motivation

In Figure 1, we briefly depict the signaling procedure when a UE that has already attached to the network begins its communication, based on [19]. In this figure we only show the signaling steps related to bearer establishment. For each UE, the mobile core network handles two bearers: one between evolved NodeB (eNodeB) and the Serving Gateway (SGW), and another between SGW and Packet data network Gateway (PGW). These bearers are persistently maintained and activated (deactivated) when the UE begins (terminates) the communication. Consequently, when many M2M terminals, that corresponding to UEs in M2M communication, connect to the mobile core network, the number of bearers to be handled increases significantly. Furthermore, signaling overhead on LTE EPC nodes, such as the Mobility Management Entity (MME), SGW, and PGW, also increases, because each periodic or intermittent communication with M2M terminals requires a signaling procedure. Decreasing the number of handled bearers through bearer aggregation is one straightforward approach toward alleviating this problem.

On the other hand, physical and fixed resource allocation of LTE EPC nodes is not desirable, due to the unpredictability of sudden changes in the mobile network traffic and the recent increase of Mobile Virtual Network Operators (MVNOs). Event-driven and intermittent communication demands by M2M terminals make the situation worse. Applying SDN architecture to mobile core networks has been considered as a means to accommodate such dynamic situations. Efficient resource utilization is achieved by separating control and data planes and aggregating control plane nodes in a cloud network environment, since we can dynamically allocate server and network resources to control plane nodes according to the node load.

From the above discussion, we believe that a combination of bearer aggregation and an SDN architecture would decrease the overhead of mobile core networks when accommodating M2M communications. This is because decreasing node overhead by bearer aggregation should allow more efficient resource sharing in SDN-ready mobile core networks, while resource sharing with control-data plane separation becomes more effective when the control-plane load of some EPC nodes is decreased by bearer aggregation.

III. Proposed method

A. *Bearer aggregation at SGW*

Figure 2 briefly depicts a mobile core network. It is composed of UEs, eNodeBs, a SGW, a PGW, and networks (eUTRAN, EPC, and external IP network) connecting them. We omit MME and other nodes from the figure to more simply explain bearer aggregation. The red lines in the figure represent bearers between UEs and eNodeB, and those between eNodeB and SGW. These bearers are established for each UE connecting to the network. In the original EPC network, a bearer is established between SGW and PGW for each UE to relay packets between UE and the external IP network. Therefore, the number of bearers between SGW and PGW equals the number of active UEs accommodated by the network.

Under bearer aggregation at SGW, a single bearer is established for multiple UEs, that is, multiple bearers between eNodeB and SGW. In Figure 2, the blue lines represent the aggregated bearers. Note that the number of blue lines is significantly smaller than the number of red lines between eNodeBs and SGW. We define the number of UEs aggregated to a single bearer between SGW and PGW as the aggregation ratio K. For example, $K = 32$ means that 32 bearers between eNodeB and SGW, corresponding to 32 UEs, are aggregated into a single bearer between SGW and PGW. We can decrease the number of bearers in the mobile core networks and corresponding signaling steps to establish bearers between SGW and PGW.

Such bearer aggregation inside the mobile core networks does not need to modify the behavior of UEs, which is a large advantage compared with group-based communication approaches [13, 14] especially when accommodating a very large number of M2M terminals. On the other hand, it requires more considerations such as realizing paging, IP address management, and traffic charging. In this paper, however, we focus on the performance gain by bearer aggregation in terms of load reduction in EPC nodes.

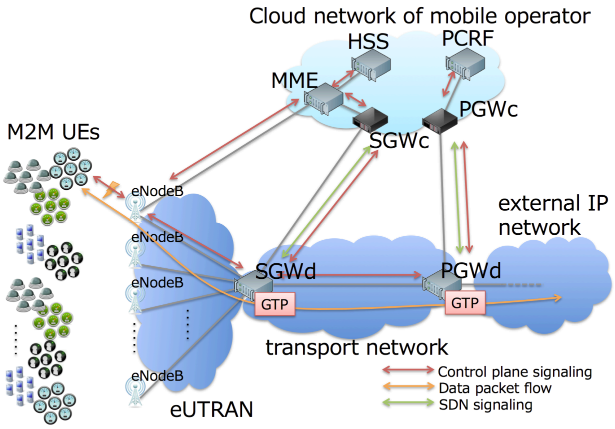


Fig. 3. Network architecture for applying SDN to an LTE EPC network

B. Applying SDN to LTE EPC networks

When applying SDN architecture to EPC nodes, we have a broad choice regarding which node should separate control and data planes, and where to place each plane. Some researchers have focused on plane separation and the plane/function placement problem in mobile core networks [15, 16, 18, 20]. Based on such discussions, we propose a network architecture like that shown in Figure 3. In this architecture, control-data plane separation is applied to SGW and PGW, and the control plane of SGW (SGWc in the figure) and PGW (PGWc) are located in a cloud network environment, as well as MME and other EPC nodes. Data plane nodes of SGW (SGWd) and PGW (PGWd) remain in the transport network. Furthermore, we locate GTP tunnel matching functions at the transport network to avoid additional latency in the data packet path [18]. Note that in the existing architecture without plane separation, control and data planes for SGW and PGW are located at SGWd and PGWd in Figure 3, respectively.

We can expect a decreased delay in signaling steps between MME and SGW/PGW since they are located on the same cloud network. On the other hand, additional signaling is required between SGWc/PGWc and SGWd/PGWd, since GTP tunnel matching information should be updated when a new UE starts communication. Furthermore, server resources can be shared among MME, SGWc, PGWc, and other EPC nodes in the cloud network, increasing overall capacity for accommodating UEs under limited server resources.

C. Impact on signaling procedure

The proposed approaches in Subsections III-A and III-B require some modifications to the signaling procedure. In Figure 4, we briefly depict a possible signaling procedure with two approaches when K UEs start communication, where K is the aggregation ratio. First, MME waits for bearer activation requests from K UEs. Then, MME configures an aggregated bearer between SGW and PGW and allocates it to K UEs (green box in the figure). Blue arrows represent the signaling steps for activating the aggregated bearer. Red arrows represent signaling steps for configuring GTP tunnel matching at GTP modules in SGW (SGWgtp) and PGW (PGWgtp). When we utilize OpenFlow, these signaling messages correspond to FlowMod messages.

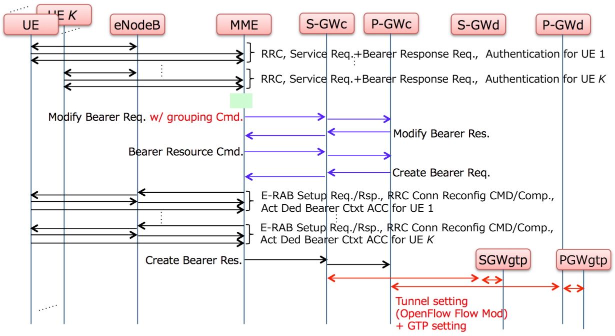

UE K
UE
eNodeB
MME
S-GWc
P-GWc
S-GWd
P-GWd
RRC, Service Req.+Bearer Response Req., Authentication for UE 1
RRC, Service Req.+Bearer Response Req., Authentication for UE K
Modify Bearer Req. w/ grouping Cmd.
Modify Bearer Res.
Bearer Resource Cmd.
Create Bearer Req.
E-RAB Setup Req./Rsp., RRC Conn Reconfig CMD/Comp., Act Ded Bearer Ctxt ACC for UE 1
E-RAB Setup Req./Rsp., RRC Conn Reconfig CMD/Comp., Act Ded Bearer Ctxt ACC for UE K
Create Bearer Res.
SGWgtp
PGWgtp
Tunnel setting
(OpenFlow Flow Mod)
+ GTP setting

Fig. 4. Signaling procedure and data packet path

Comparing Figures 1 and 4, we can observe increased steps and decreased steps for starting communications by K UEs, that affects on the load of EPC nodes. In the next section we present a detailed mathematical analysis to compare the performance of the proposed method with that of the original one.

IV. Performance analysis

A. Analysis overview

In this section we derive data transfer times by mathematical analysis. Data transfer time is defined as the time duration from when a UE starts to establish its bearers for packet transmission to when the data is completely transferred. It is divided into bearer establishment time and data transmission time.

For bearer establishment time, we apply a simple queuing model to derive the time required for processing each signaling message. We then follow the signaling steps in Figures 1 and 4 to derive the bearer establishment time.

For data transmission time, we assume that the destination server is located at the external IP network in Figure 3, where we ignore the propagation delay between PGW and the destination server. We consider both TCP and UDP to confirm the effect of transport-layer protocol overhead.

In the following analysis, we first derive the time in the case where only control-data plane separation is applied. We then extend the analysis to accommodate bearer aggregation.

B. Notations

The propagation delay of signaling messages between EPC nodes is denoted as $\tau_{\mathrm{NODE1,NODE2}}$, where NODE1 and NODE2 represent one of a terminal or a node in Figures 1 and 4, that is, UE, eNodeB, MME, SGW, PGW, SGWc, SGWd, PGWc, PGWd, SGWgtp, and PGWgtp. For example, $\tau_{\mathrm{SGW,PGW}}$ represents the propagation delay between SGW and PGW. We also denote the delay for processing a single signaling message at each node by t_{NODE}, where NODE represents the same meaning as NODE1 and NODE2 in $\tau_{\mathrm{NODE1,NODE2}}$. For example, t_{MME} means the processing delay for a single signaling message at MME.

C. Bearer establishment time

The bearer establishment time is the sum of propagation delays between nodes and the node processing time for all signaling messages. For the original signaling procedure without the proposed approaches, we follow the signaling steps in Figure 1 and sum up the delays to derive the bearer establishment delay T_{org}, as follows.

$$\begin{aligned} T_{\mathrm{org}} \quad = \quad & L_{\mathrm{RRC}} \\ & +(6t_{\mathrm{UE}} + 4t_{\mathrm{eNodeB}} + 9t_{\mathrm{MME}} + 5t_{\mathrm{SGW}} + 3t_{\mathrm{PGW}}) \\ & +(10\tau_{\mathrm{UE,eNodeB}} + 10\tau_{\mathrm{eNodeB,MME}} + 5\tau_{\mathrm{MME,SGW}} + 5\tau_{\mathrm{SGW,PGW}}) \end{aligned}$$

In the above equation, L_{RRC} represents the time for establishing a Radio Resource Control (RRC) connection between UE and eNodeB.

When applying only the SDN architecture without bearer aggregation, additional signaling steps are required between SGWc and SGWd and between PGWc and PGWd. Furthermore, some control messages are exchanged between SGWd/PGWd and SGWgtp/PGWgtp for configuring GTP modules. We denote the number of required message exchanges between SGWc/PGWc and SGWd/PGWd and the number of required message exchanges between SGWd/PGWd and SGWgtp/PGWgtp as N_{FM} and N_{GTP}, respectively. By following the signaling steps in Figure 4, we can derive the bearer establishment time when applying only the SDN architecture T_{SDN} as follows.

$$\begin{aligned} T_{\mathrm{SDN}} \quad = \quad & L_{\mathrm{RRC}} + (6t_{\mathrm{UE}} + 4t_{\mathrm{eNodeB}} + 9t_{\mathrm{MME}} + 5t_{\mathrm{SGWc}} + 3t_{\mathrm{PGWc}}) \\ & +(10\tau_{\mathrm{UE,eNodeB}} + 10\tau_{\mathrm{eNodeB,MME}} + 5\tau_{\mathrm{MME,SGWc}} + 5\tau_{\mathrm{SGWc,PGWc}}) \\ & + \max(2N_{\mathrm{FM}}\tau_{\mathrm{SGWc,SGWd}} + (N_{\mathrm{FM}} + 1)t_{\mathrm{SGWc}} + N_{\mathrm{FM}}t_{\mathrm{SGWd}} \\ & +2N_{\mathrm{GTP}}\tau_{\mathrm{SGWd,SGWgtp}} + (N_{\mathrm{GTP}} + 1)t_{\mathrm{SGWd}} + N_{\mathrm{GTP}}t_{\mathrm{SGWgtp}}, \\ & 2N_{\mathrm{FM}}\tau_{\mathrm{PGWc,PGWd}} + (N_{\mathrm{FM}} + 1)t_{\mathrm{PGWc}} + N_{\mathrm{FM}}t_{\mathrm{PGWd}} \\ & +2N_{\mathrm{GTP}}\tau_{\mathrm{PGWd,PGWgtp}} + (N_{\mathrm{GTP}} + 1)t_{\mathrm{PGWd}} + N_{\mathrm{GTP}}t_{\mathrm{PGWgtp}}) \end{aligned}$$

To derive node processing time, t_{NODE}, we exploit the M/G/1/PS queuing model. In the M/G/1/PS model, the mean sojourn time $E[R]$ can be derived as

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

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, and $\rho = \lambda \cdot E[S]$ is the system utilization. By assuming that an overhead for handling a signaling message is identical for all nodes, we obtain t_{NODE} for a given node using Equation (1). In detail, we use the arrival rate of signaling messages at the node for λ. The arrival rate of signaling messages can be calculated from the number of signaling messages per one communication, the communication frequency, and the data size of UE and transport-layer protocol overhead. Tables I and II respectively summarize the number of signaling messages to be processed at each node per one communication by TCP and UDP, where N_{P} is the number of data packets to be transmitted. We present the results of the original network without the proposed approaches and those of the network with control-data plane separation.

TABLE I. NUMBER OF SIGNALING MESSAGES PER UE COMMUNICATION (TCP)

	Original	Control-data plane separation
UE	6+$(3/2)N_{\mathrm{P}} + 2$	6+$(3/2)N_{\mathrm{P}} + 2$
eNodeB	4+$(3/2)N_{\mathrm{P}} + 2$	4+$(3/2)N_{\mathrm{P}} + 2$
MME	9	9
SGW	5+$(3/2)N_{\mathrm{P}} + 2$	
SGWc		5+2N_{FM}
SGWd		$N_{\mathrm{FM}} + 2N_{\mathrm{GTP}} + 2((3/2)N_{\mathrm{P}} + 2)$
SGWgtp		$N_{\mathrm{GTP}} + (3/2)N_{\mathrm{P}} + 2$
PGW	3+$(3/2)N_{\mathrm{P}} + 2$	
PGWc		3+2N_{FM}
PGWd		$N_{\mathrm{FM}} + 2N_{\mathrm{GTP}} + 2((3/2)N_{\mathrm{P}} + 2)$
PGWgtp		$N_{\mathrm{GTP}} + (3/2)N_{\mathrm{P}} + 2$

TABLE II. NUMBER OF SIGNALING MESSAGES PER UE COMMUNICATION (UDP)

	Original	Control-data plane separation
UE	6+N_{P}	6+N_{P}
eNodeB	4+N_{P}	4+N_{P}
MME	9	9
SGW	5+N_{P}	
SGWc		5+2N_{FM}
SGWd		$N_{\mathrm{FM}} + 2N_{\mathrm{GTP}} + 2N_{\mathrm{P}}$
SGWgtp		$N_{\mathrm{GTP}} + N_{\mathrm{P}}$
PGW	3+N_{P}	
PGWc		3+2N_{FM}
PGWd		$N_{\mathrm{FM}} + 2N_{\mathrm{GTP}} + 2N_{\mathrm{P}}$
PGWgtp		$N_{\mathrm{GTP}} + N_{\mathrm{P}}$

The workload distribution $S(x)$ corresponds to the distribution of processing time for signaling messages by the node. For simplicity, we assume that the processing time distribution of signaling messages follows an exponential distribution. The mean workload $E[S]$ is determined by the processing power. Finally, we obtain t_{NODE} by calculating $E[R]$ in Equation (1).

D. Data transmission time

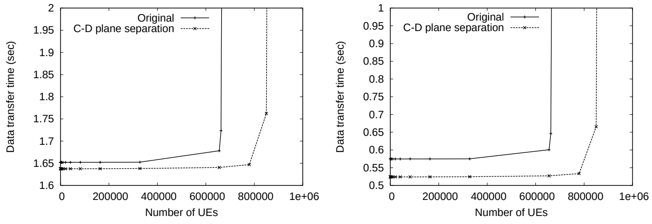
After establishing bearers the UE starts sending data packets to the destination server. We calculated data transmission time based on the protocol overheads of TCP and UDP. For TCP transmission, we assume that the transmission data size is sufficiently small to be transmitted in a slow start phase. The data transmission time with TCP $C_{\mathrm{TCP}}(S)$ can be obtained as

$$\begin{aligned} C_{\mathrm{TCP}}(S) \quad \approx \quad & T + 2\left(O + \frac{P_{\mathrm{header}}}{B_{\mathrm{wireless}}} + \frac{P_{\mathrm{header}}}{B_{\mathrm{core}}}\right) \\ & +2\left(\log_2\left(\left\lfloor \frac{S}{P - P_{\mathrm{header}}} \right\rfloor + 1\right) + 1\right) \\ & \cdot \left(O + \frac{P}{B_{\mathrm{wireless}}} + \frac{P}{B_{\mathrm{core}}}\right), \end{aligned} \tag{2}$$

where S is the transmission data size, P is the data packet size, P_{header} is the sum of the TCP/IP header sizes, B_{wireless} is the bandwidth of the wireless network between UE and eNodeB, B_{core} is the bandwidth of the mobile core network between eNodeB and PGW, and O is the one-way delay from the UE to the destination server, which is calculated by the propagation delays and node processing delays on a data packet path from the UE to the destination server.

We similarly obtain the data transmission time with UDP $C_{\mathrm{UDP}}(S)$ as follows.

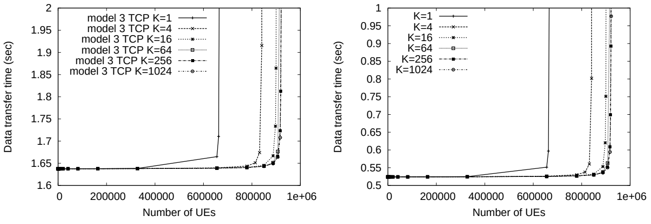
$$\begin{aligned} C_{\mathrm{UDP}}(S) \quad \approx \quad & T + \left(O + \frac{P_{\mathrm{header}}}{B_{\mathrm{wireless}}} + \frac{P_{\mathrm{header}}}{B_{\mathrm{core}}}\right) \\ & + \left(\left\lfloor \frac{S}{P - P_{\mathrm{header}}} \right\rfloor + 1\right) \\ & \cdot \min(B_{\mathrm{wireless}}, B_{\mathrm{core}}) \end{aligned} \tag{3}$$



(a) TCP traffic

(b) UDP traffic

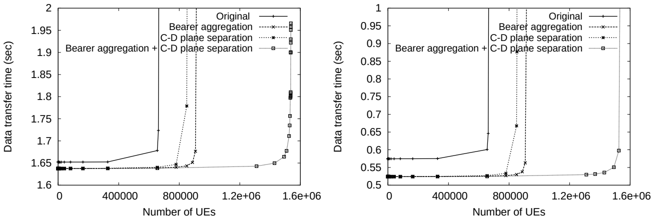
Fig. 5. Effect of control-data plane separation



(a) TCP traffic

(b) UDP traffic

Fig. 6. Effects of bearer aggregation



(a) TCP traffic

(b) UDP traffic

Fig. 7. Combination of bearer aggregation and control-data plane separation

E. Effect of bearer aggregation

By applying bearer aggregation, we can decrease the number of signaling messages per UE communication in Tables I and II according to the aggregation ratio K. Specifically, signaling messages for activating a bearer between SGWd and PGWd (blue arrows in Figure 4) are required only once for K UEs. Considering this, the number of signaling messages to be processed is decreased from 9 to $6.5 + 2.5/K$ at MME, from 5 to $1 + 4/K$ at SGWc, and from 3 to $3/K$ at PGWc. This means that increasing K would decrease the node overhead significantly, but the positive effect would converge with too large a K.

F. Numerical results and discussion

We set B_{wireless} = 10 Mbps, B_{core} = 1 Gbps, and P = 128 bytes. P_{header} is set to 40 bytes for TCP transmission, and 28 bytes for UDP transmission. Each UE initiates a communication in 600 sec intervals and the transmission data size for each communication is set to 1,000 bytes. The propagation delays between nodes and terminals are set as $\tau_{\text{UE,eNodeB}}$ = 20 msec, $\tau_{\text{eNodeB,MME}}$ = $\tau_{\text{eNodeB,SGW}}$ = $\tau_{\text{SGW,PGW}}$ = $\tau_{\text{SGWd,PGWd}}$ = 7.5 msec, $\tau_{\text{MME,SGW}}$ = $\tau_{\text{SGWc,SGWd}}$ = $\tau_{\text{PGWc,PGWd}}$ = 10 msec, $\tau_{\text{eNodeB,SGWc}}$ = $\tau_{\text{SGWc,PGWc}}$ = 1 msec, $\tau_{\text{SGWd,SGWgtp}}$ = $\tau_{\text{PGWd,PGWgtp}}$ = 1 msec, and L_{RRC} = 20 msec. For node processing power, we set t_{UE} = t_{eNodeB} = 1,000 messages/sec, t_{MME} = t_{SGW} = t_{PGW} = t_{SGWc} = t_{PGWc} = t_{SGWd} = t_{PGWd} = 10,000 messages/sec, and t_{SGWgtp} = t_{PGWgtp} = 100 Gbps. Note that these parameters are configured as typical values based on discussions on the current status of mobile core networks operated by the mobile carrier company in Japan.

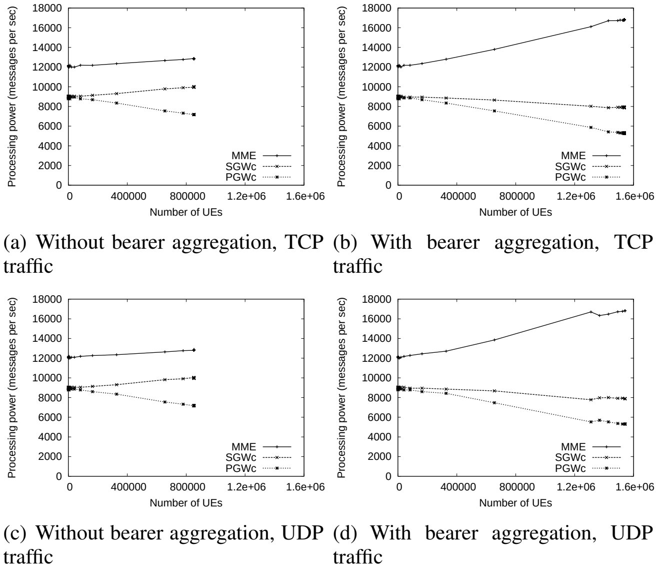
For the control-data plane separation in the proposed method, we set the processing power of MME, SGWc, and PGWc so that we obtain the minimum value of data transmission time, while the total processing power remains unchanged at 30,000 messages/sec.

Figure 5 shows the changes in data transmission time as a function of the number of accommodated UEs when we utilize TCP (Figure 5(a)) and UDP (Figure 5(b)). We plot the results of the original method and those of the proposed method with only control-data plane separation. From these figures we can observe that regardless of the transport-layer protocol, the data transmission time of the proposed method is smaller than that of the original method. This means that when applying the SDN architecture, the positive effect of decreasing propagation delays of signaling messages is larger than the negative effect of the additional signaling messages for SDN control. Furthermore, the number of accommodated UEs without divergence of the data transmission time, which is defined as the network capacity, can be increased by using the proposed method. In this case the network capacity increases by around 30%.

Figure 6 present the results of the proposed method with only bearer aggregation. K = 1 corresponds to the original method. From these figures we can observe that larger values of K result in higher network capacity, up to a certain level at which the capacity converges with further increases of K. In this case the network capacity increases by 37% for K = 64. We also confirm that the data transmission delay remains almost unchanged with a smaller number of UEs regardless of K. This means that the negative effect of the bearer aggregation can be ignored.

We finally show the results of a combination of the control-data plane separation and bearer aggregation in Figure 7. We set K = 64 for bearer aggregation. For comparison, we also plot the results of the original method, those of the proposed method with only control-data plane separation, and those of the proposed method with only bearer aggregation. From this figure we can see that combining the two approaches significantly increases the network capacity. Specifically, the network capacity increases by 124% by combining two approaches, but only 30% and 37% with only one approach. This result clearly shows the effectiveness of combining the two approaches, as predicted in Section II.

In Figure 8, we plot the processing power of MME, PGW, and SGW nodes in the proposed method, to confirm the effect of the server resource sharing in an SDN architecture. In the figure we plot the results with and without bearer aggregation for comparison. From this figure, we can see that more processing power is allocated to MME, while decreasing the processing power of SGW and PGW. This is because the bottleneck is the signaling message processing at MME. On



(a) Without bearer aggregation, TCP traffic

(b) With bearer aggregation, TCP traffic

(c) Without bearer aggregation, UDP traffic

(d) With bearer aggregation, UDP traffic

Fig. 8. Node processing power of the proposed methods

the other hand, the processing power of SGW increases as the number of UEs increases, because the signaling overhead of SGW increases. Furthermore, when applying bearer aggregation (Figures 8(b) and (d)) the more processing power is allocated to MME, while SGW and PGW have less processing power. This is because the signaling overhead at SGW and PGW decreases with bearer aggregation and more processing power can be allocated to MME to further decrease the node processing time. These results again show that a combination of bearer aggregation and control-data plane separation is very meaningful.

V. Conclusion

We proposed a method for increasing the capacity of M2M communications in mobile core networks. The proposed method combines two approaches, bearer aggregation and control-data plane separation. We performed mathematical analysis to evaluate the performance of the proposed method. From extensive evaluation results we confirm that the network capacity, in terms of the number of accommodated M2M terminals without divergence of the data transfer time, increases by 124% by combining two approaches, but only 30% and 37% with only one approach. These results clearly show that the combination of these two approaches is meaningful since they have complementary relationship.

In future work we plan to extend the proposed method to accommodate wide-area mobile core networks, in which the location of EPC nodes and distribution of UEs should be taken into account. Also important is enhancing the analysis model in Section IV by determing the detailed procedure for bearer aggregation and by considering the overhead of each signaling message in detail.

Acknowledgment

The authors thank Wataru Takita, Kazuaki Obana, and Motoshi Tamura from NTT DOCOMO, and Takashi Shimizu from NTT Network Innovation Laboratories for their helpful comments. The authors also appreciate the feedback offered by Kohei Shiomoto, Noriaki Kamiyama, and Yuusuke Nakano from NTT Network Technology Laboratories, and Hideo Mihayara from Osaka University.

References

[1] A. de la Oliva, C. J. Bernardos, M. Calderon, T. Melia, and J. C. Zuniga, “IP flow mobility: Smart traffic offload for future wireless networks,” *IEEE Communication Magazine*, vol. 49, pp. 124–132, Oct. 2011.

[2] Cisco Systems, Inc., “Architecture for mobile data offload over Wi-Fi access networks,” *Available from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/service-provider-wi-fi/white_paper_c11-701018.pdf*.

[3] 3GPP TS 24.312, “Access network discovery and selection function (ANDSF) management object (MO),” 2014.

[4] A. Balasubramanian, R. Mahajan, and A. Venkataramani, “Augmenting mobile 3G using WiFi,” in *Proceedings of MobiSys 2010*, pp. 209–222, June 2010.

[5] K. Lee, I. Rhee, J. Lee, S. Chong, and Y. Yi, “Mobile data offoading: How much can WiFi deliver?,” in *Proceedings of CoNEXT 2010*, Nov. 2010.

[6] F. Qian, Z. Wang, Y. Gao, J. Huang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, “Periodic transfers in mobile applications: Network-wide origin, impact, and optimization,” in *Proceedings of WWW 2012*, Apr. 2012.

[7] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, “A first look at cellular machine-to-machine traffic - large scale measurement and characterization,” in *Proceedings of ACM SIGMETRICS 2012*, June 2012.

[8] A. Daj, C. Samoila, and D. Ursutiu, “Digital marketing and regulatory challenges of Machine-to-Machine (M2M) communications,” in *Proceedings of REV 2012*, pp. 1–5, July 2012.

[9] D. Bouallouche, “Congestion control in the context of machine type communications in 3GPP LTE networks,” *Master thesis internship report, University of Rennes*, Aug. 2012.

[10] R. Vaidya, C. Yadav, J. Kunkumath, and P. Yadati, “Network congestion control: Mechanisms for congestion avoidance and recovery,” in *Proceedings of ACWR 2011*, Dec. 2011.

[11] Y. Chen and W. Wang, “Machine-to-machine communication in LTE-A,” in *Proceedings of VTC2010-Fall*, pp. 1–4, Sept. 2010.

[12] K. Jun, “Enabling massive machine-to-machine communications in LTE-Advanced,” in *Proceedings of GPC 2013*, pp. 563–569, May 2013.

[13] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, “Device-to-device communication as anunderlay to LTE-Advanced networks,” *IEEE Communication Magazine*, vol. 47, pp. 42–49, Dec. 2009.

[14] S.-Y. Lien, K.-C. Chen, and Y. Lin, “Toward ubiquitous massive accesses in 3GPP machine-to-machine communications,” *IEEE Communication Magazine*, vol. 49, pp. 66–74, Apr. 2011.

[15] A. Khan, D. Jurca, K. Kozu, W. Kellerer, and M. Yabusaki, “The reconfigurable mobile network,” in *Proceedings of ICC 2011*, pp. 1–5, June 2011.

[16] L. E. Li, Z. M. Mao, and J. Rexford, “Toward software-defined cellular networks,” in *Proceedings of EWSDN 2012*, pp. 7–12, Oct. 2012.

[17] A. Khan, W. Kellerer, K. Kozu, and M. Yabusaki, “Network sharing in the next mobile network: TCO reduction, management flexibility, and operational independence,” *IEEE Communication Magazine*, vol. 49, pp. 134–142, Oct. 2011.

[18] 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 SDN4FNS 2013*, pp. 8–14b, Nov. 2013.

[19] V. S. Rao and R. Gajula, “Protocol signaling procedures in LTE,” *White Paper, Radisys Corporation*, Sept. 2011.

[20] 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 AllThingsCellular 2014 Workshop*, Aug. 2014.