Studies on Mobile Communication System Architecture for Accommodating IoT Devices and Mobile Users

Submitted to
Graduate School of Information Science and Technology
Osaka University

January 2021

Shuya ABE

List of publications

Journal Papers

1. Shuya Abe, Go Hasegawa, and Masayuki Murata, "Effects of C/U Plane Separation and Bearer Aggregation in Mobile Core Network," *IEEE Transactions on Network and Service Management*, vol. 15, no. 2, pp. 611-624, June 2018.

International Conference Papers

1. Shuya Abe, Go Hasegawa, and Masayuki Murata, "Design and Performance Evaluation of Bearer Aggregation Method in Mobile Core Network with C/U Plane Separation," in *Proceedings of 2017 IFIP Networking Conference (IFIP Networking) and Workshops*, pp. 1–8, June 2017.

2. Shuya Abe, Go Hasegawa, and Masayuki Murata, "Performance Analysis of Periodic Cellular-IoT Communication with Immediate Release of Radio Resources," in *Proceedings of 2020 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*, pp. 1–6, May 2020.

Non-Refereed Technical Papers

1. Shuya Abe, Go Hasegawa, and Masayuki Murata, "Performance Evaluation of Virtualized and Plane-separated Mobile Core Network Considering Signaling Message Overhead," *Technical Report of IEICE (ICM2015-49)*, vol. 115, no. 507, pp. 41–46, March 2016.

2. Shuya Abe, Go Hasegawa, and Masayuki Murata, "Performance Evaluation of Bearer Aggregation in Mobile Core Network with C/U Plane Separation," *Technical Report of IEICE (NS2016-122)*, vol. 116, no. 382, pp. 25–30, December 2016.

3. Tomoya Adachi, Shuya Abe, Go Hasegawa, and Masayuki Murata, “Efficient Use of Mobile Core Node Resources by Signaling Control Considering IoT Devices,” *Technical Report of IEICE (CQ2019-97)*, vol. 119, no. 298, pp. 47–52, November 2019.

4. Tomoya Adachi, Shuya Abe, Go Hasegawa, and Masayuki Murata, “Adaptive Signaling Control Method for Efficient Resource Utilization of Mobile Core Network Nodes,” *Technical Report of IEICE (CQ2019-141)*, vol. 119, no. 455, pp. 35–40, March 2020.

Preface

Currently, the majority of Internet traffic comes from mobile and IoT devices. This is facilitated by the shift to cloud computing for applications, richer content for applications on user devices such as smartphones, and the increasing functionality and diversity of edge devices.

Cellular networks are attractive in that they can use existing resources and coverage to accommodate mobile and Internet of Things (IoT) communications. With increasing demand for cellular networks owing to the proliferation of IoT terminals, increasing the capacity of mobile core networks is important. However, there are few existing studies that can reduce the signaling overhead between nodes in the core network during accommodation. We summarize the studies on network architectures for efficiently accommodating IoT devices in cellular networks. In the first, a mobile core network architecture with virtualization technologies and C/U plane separation is evaluated using mathematical analysis. In addition, we propose a bearer aggregation method to reduce the load on the control plane to support massive M2M/IoT terminals.

In addition to the existing cellular networks, a new communication network called Low-power Wide-area Network (LPWAN)) is emerging and accommodating mobile and IoT communications. These networks are designed with a focus on low-power and wide-area IoT communications, and include cellular LPWANs called Cellular-IoT (C-IoT) deployed in licensed frequency bands and non-cellular LPWANs deployed in unlicensed frequency bands. We focus on NB-IoT, one of the cellular LPWAN, as a mobile cellular network to support periodic C-IoT communication, and conduct a detailed analysis considering all of random access procedures, radio resource allocation, establishing bearers in the Evolved Packet Core network, and user-data transmissions.

In addition, the demands on the network are becoming more diverse, and 5G slicing is one of the ways to accommodate them. However, 5G slicing built and managed on a per-service basis cannot meet the requirements due to the demand for communication by users moving across cells in cellular and non-cellular networks. In order to accommodate this kind of communication, per-service slicing is not sufficient, and a new mobile communication architecture is required. Therefore, we propose User-Oriented slicing Network Architecture (UONA) as an architecture to accommodate diversified and personalized network services. UONA is a new architecture for maintaining network slices on a per-user basis in order to satisfy diversified specification requirements from users.

We believe that the results of these studies will provide an important contribution to the establishment of a new and flexible mobile communication system architecture that meets a wide variety of requirements.

Acknowledgments

Several persons directly or indirectly contributed to the work presented in this thesis. This section is intended to thank all these persons. I am sorry to the persons that are not listed here. Thank to all of you.

I would like to express my gratitude to Professor Morito Matsuoka for his dedication as my advisor, as well as his support and guidance. I thank him for giving me the opportunity to study.

I am grateful to the members of my thesis committee, Professor Masayuki Murata, Professor Toru Hasegawa, Professor Takashi Watanabe and Professor Teruo Higashino of the Graduate School of Information Science and Technology, Osaka University for reviewing my dissertation and providing many constructive comments in completing this thesis. Especially, Professor Masayuki Murata has provided support and advice throughout my research life.

I would especially like to express my appreciation to Professor Go Hasegawa of Tohoku University for his continuously generous support and encouragement.

I am thankful to all the members of the Matsuoka Laboratory in Graduate School of Information Science and Technology, Osaka University, for their continuous support and friendship.

Finally, I want to thank my parents and friends for their invaluable support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Currently, the majority of Internet traffic comes from mobile and IoT devices. According to Cisco's forecast [2], mobile traffic will exceed 13 billion in 2023 and M2M traffic will reach about 14.7 billion in 2023. This is facilitated by the shift to cloud computing for applications [3], richer content for applications on user devices such as smartphones [4], and the increasing functionality and diversity of edge devices [5].

Cellular networks are attractive in that they can use existing resources and coverage to accommodate mobile and Internet of Things (IoT) communications [6]. With increasing demand for cellular networks owing to the proliferation of IoT terminals, increasing the capacity of mobile core networks is important. The reason is that many IoT communications have characteristics different from rich user terminals such as smartphones—communication may occur periodically or intermittently with small amounts of data, while the number of IoT communications may be enormous. Most of the existing researches can improve the efficiency of using the computational resources of the core network when accommodating IoT communications. However, there are few existing studies that can reduce the signaling overhead between nodes in the core network during accommodation.

In addition to the existing cellular networks, a new communication network called Low-power Wide-area Network (LPWAN)) is emerging and accommodating mobile and IoT communications. These networks are designed with a focus on low-power and wide-area

IoT communications, and include cellular LPWANs called Cellular-IoT (C-IoT) deployed in licensed frequency bands and non-cellular LPWANs deployed in unlicensed frequency bands. Among C-IoT radio technologies, Narrowband-IoT (NB-IoT) is optimized for most IoT communications that transmit small data. Moreover, in standardization activities, the immediate release of radio resources after the completion of data transmissions is considered [7], while radio resources are kept allocated at least 10 seconds in traditional mobile cellular networks and NB-IoT. The immediate release of radio resources would have a considerable effect in accommodating IoT communications with small data transmissions.

In addition, the demands on the network are becoming more diverse, and 5G slicing is one of the ways to accommodate them. However, 5G slicing built and managed on a per-service basis cannot meet the requirements due to the demand for communication by users moving across cells in cellular and non-cellular networks.

1.2 Issues

1.2.1 Accommodation of IoT Devices

Mobile cellular networks are now required for serving numerous IoT communications. However, IoT communications cannot generally be efficiently accommodated because IoT traffic has different characteristics from human communications—communication may occur periodically or intermittently with small amounts of data, while the number of terminals may be enormous. In addition, many IoT terminals have almost no mobility, and most of them only transmit data (i.e., no data are received).

LPWANs are the main way to accommodate these IoT communications. They are mainly divided into cellular LPWANs and non-cellular LPWANs. Non-cellular LPWANs are, for example, LoRa networks using LoRa and LoRaWAN, and Sigfox. Non-cellular LPWANs have the advantage of being able to use unlicensed frequency bands, which generally results in lower costs. On the other hand, the coverage is generally smaller than cellular LPWAN, and the security and radio interference are disadvantageous. In Japan, for example, Softbank and KDDI have LoRa network, and Kyocera has Sigfox service. Cellular

LPWAN, such as NB-IoT, have the advantage of coverage that cannot be obtained with non-cellular systems. Network bandwidth is also high. In Japan, major mobile communication carriers are providing cellular LPWAN services [8, 9].

The demand for such IoT communications will undoubtedly increase in the future. In response, communication standards for C-IoT are being standardized in 5G networks [10]. However, they are mainly aimed at extending the life of 3G/4G/LTE devices, and do not answer the various communication requirements of 5G and beyond.

1.2.2 Accommodation of Mobile Users

The applications of mobile users are becoming more and more diverse, and the quality they demand from the network is also becoming more advanced and diverse. In 5G, the requirements to realize major application QoS such as uRLLC, eMBB, and mMTC are set, and the technologies for the network to realize them are being studied [11]. Among these, network slicing is the most important. In network slicing, a virtual, logically independent, service-specific network (slice) is constructed for each service, and by providing appropriate resources, application traffic generated by user terminals and IoT terminals is accommodated in the slice constructed according to their communication requirements. By providing appropriate resources, application traffic generated by user terminals and IoT terminals is accommodated in the slices constructed according to their communication requirements. On the other hand, the diversification of application services is moving toward personalization. In order to accommodate this kind of communication, per-service slicing is not sufficient, and a new mobile communication architecture is required.

1.3 Contributions

In this thesis, we aim to develop a new mobile communication system architecture, which is essential for realizing the future IoT society and Cyber-Physical System.

First, we summarize the studies on network architectures for efficiently accommodating IoT devices in cellular networks. In the first, a mobile core network architecture with virtualization technologies and C/U plane separation is evaluated using mathematical analysis.

In addition, we propose a bearer aggregation method to reduce the load on the control plane to support massive M2M/IoT terminals. Next, we focus on NB-IoT, one of the cellular LP-WAN, as a mobile cellular network to support periodic C-IoT communication, and conduct a detailed analysis considering all of random access procedures, radio resource allocation, establishing bearers in the Evolved Packet Core network, and user-data transmissions.

Finally, we propose User-Oriented slicing Network Architecture (UONA) as an architecture to accommodate diversified and personalized network services. Specifically, UONA is a new architecture for maintaining network slices on a per-user basis in order to satisfy diversified specification requirements from users.

We believe that the results of these studies will provide an important contribution to the establishment of a new and flexible mobile communication system architecture that meets a wide variety of requirements.

1.4 Outline of the Thesis

The structure of this thesis is as follows.

In Chapter 2, we first evaluate mobile core network architectures with virtualization technologies and C/U plane separation using the mathematical analysis. We also propose a novel bearer aggregation method to reduce the control plane load to accommodate massive 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 IoT terminals.

In Chapter 3, we give the performance analysis results of mobile cellular networks to support periodic C-IoT communications, focusing on the performance differences between LTE and NB-IoT networks. To achieve this, we first construct the analysis model for end-to-end performance of both the control plane and data plane, including random access procedures, radio resource allocation, establishing bearers in the Evolved Packet Core network, and user-data transmissions. In addition, we include the impact of the immediate

release of radio resources proposed in 3GPP. The numerical evaluation results show that NB-IoT can support more IoT devices than LTE up to 8.7 times, but has a significant delay in data transmissions.

In Chapter 4, we propose a novel network architecture, UONA. It has the two major characteristics. One is that network slices is maintained in a *per-user basis*. Such high-resolution of network slices would satisfy diversifying and personal requirements from users. The other is decoupling the process of generating network slices into two subprocesses, providing *subslices* by *subslice providers* and constructing end-to-end user network slices from the subslices by *network slice brokers*. We present the overall design of UONA and explain its advantages, as well as the research challenges to realize it. We also demonstrate the effectiveness of UONA by presenting the numerical evaluation results. We confirm that UONA outperforms the service-based slicing in 5G networks and other methods by selecting control parameters in per-user basis.

In Chapter 5, conclusions of this thesis and future work are described.

Chapter 2

Effects of C/U Plane Separation and Bearer Aggregation in Mobile Core Network

2.1 Introduction

Background. 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 [12], increasing the capacity of mobile core networks is important [13]. 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) [14]". 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) [15–17], 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 chapter, 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 [18–29]. In [30], 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.

Contribution. In this chapter, we evaluate the performance of mobile core networks with node virtualization and control/user (C/U) plane separation. In addition, we propose a detailed mechanism of bearer aggregation. The main contributions of this chapter are as follows.

1. Introduction of 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.

2. Numerical evaluation of the performance of mobile core networks with node virtualization and C/U plane separation with SDN.

3. Proposal and evaluation of bearer aggregation method to concretize the primal idea proposed in [30].

4. Presentation of parameter design according to the characteristics of M2M/IoT terminals.

Outline. In Section 2.2, extant research related to our study is summarized. Section 2.3 presents the architectures of the mobile core network used herein. Section 2.4 introduces the bearer aggregation method. Section 3.4 describes the mathematical analysis of the performance of the mobile core network. In Section 3.5, we provide extensive numerical

Table 2.1: **Comparison of methods:** The characteristics of various existing evaluations and the proposed method.

	[18–22]	[23–27]	[28]	[29]	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				✓	✓

evaluation results and discussions. In Section 2.7, based on the evaluation results, we show the design of parameter settings for M2M/IoT terminals with various characteristics. Finally, in Section 4.6, we conclude this chapter with a brief summary and an outline of future work.

2.2 Related Work

Various methods have been proposed for improving the capacity of M2M/IoT communications in mobile cellular networks [31]. These existing works, as well as the method proposed in this chapter, are listed in Table 3.1.

In [18], the authors showed that efficient resource utilization can be achieved by implementing EPC nodes as software. In addition, [19–21] 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 [22] and [23] showed that applying NFV to EPC nodes decreases the amount of signaling traffic and the cost of devices, infrastructure, and energy consumption. In [23, 24, 27, 28], application of SDN to mobile core networks and virtualization of nodes in a cloud environment were studied.

The authors of [25] 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 [26], 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 [27] where the authors evaluated additional network traffic due to the application of SDN to the mobile core network. In [23, 24, 27], 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 [29], 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) [32], 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.

2.3 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.

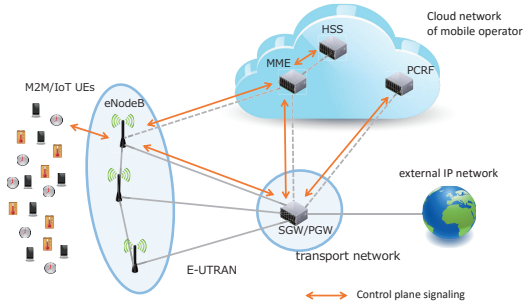
2.3.1 Conventional Architecture (CA)

The conventional mobile core network architecture is shown in Fig. 2.1(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. 2.1(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.

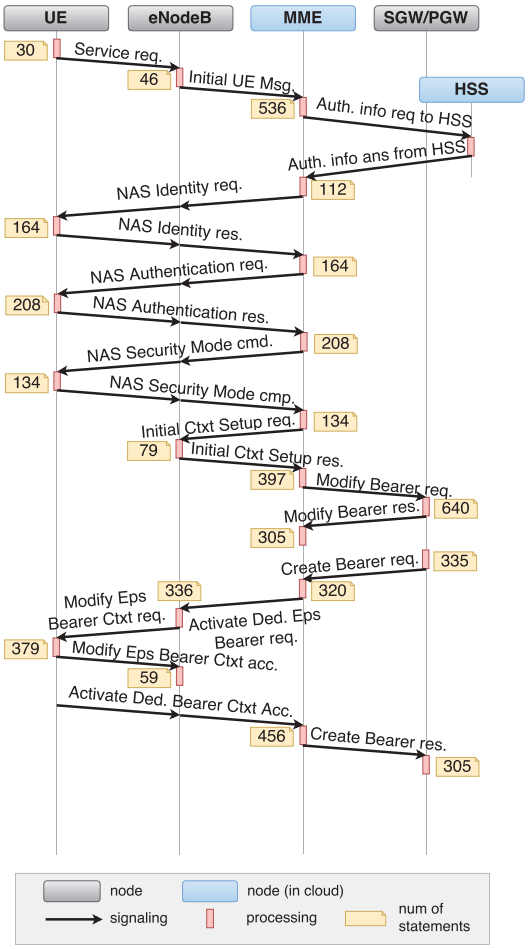
The networks where data is carried by RRC or NAS messages such as small data transmission in NB-IoT [16] can also be evaluated by assuming that the signaling procedure shown in Fig. 2.1(b) is terminated at NAS Security Mode cmp and a pair of a MME and an SGW can perform as a Cellular IoT Serving Gateway Node [16].

2.3.2 C/U Plane-separated Architecture (PSA)

Fig. 2.2(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



(a) Mobile core network model.



(b) Signaling flow for bearer establishment.

Figure 2.1: Conventional Architecture (CA)

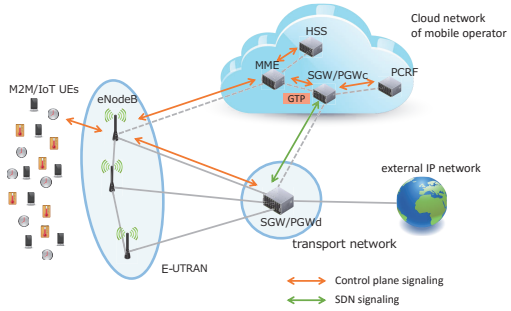
plane functions and SGW/PGW$_d$ for data plane functions. SGW/PGW$_c$ is virtualized and located in the cloud environment, while SGW/PGW$_d$ is in the transport network without virtualization, as in CA. By applying C/U plane separation, the propagation delay between SGW/PGW$_c$ and MME becomes smaller than that in CA.

In this architecture, a GTP module [28], 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 PSA. 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.

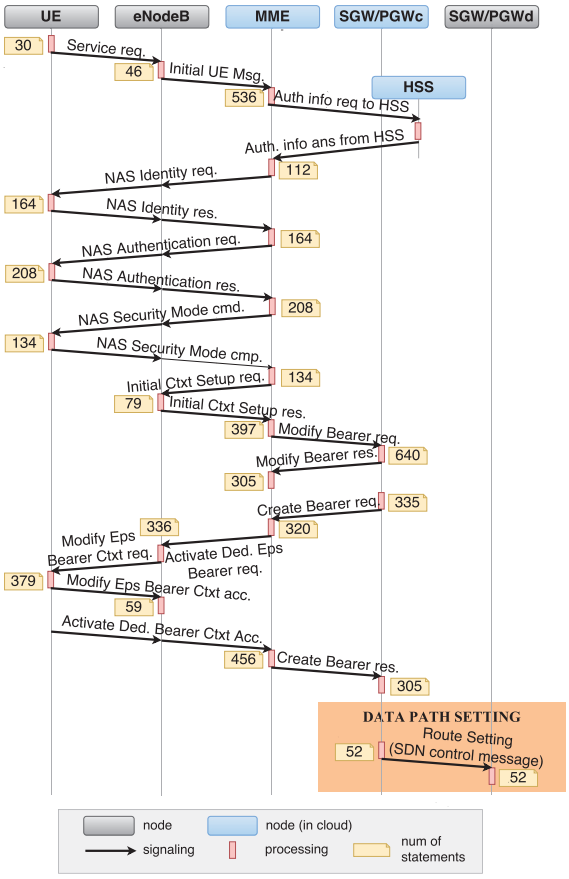
Fig. 2.2(b) shows the signaling flow when a UE changes its state from idle to active and requests the start of communication. Compared with Fig. 2.1(b), a control message related to route setting by SDN is required to be sent from SGW/PGW$_c$ to SGW/PGW$_d$ after signaling bearer establishment is complete. 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 CA since messages between MME and SGW/PGWC are exchanged within the cloud environment.

2.3.3 C/U Plane-separated Architecture with GTP Module Located in Data Plane (PSA$_g$)

Fig. 2.3(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 PSA, while an additional signaling message is required to configure the GTP module for bearer establishment, as shown in Fig. 2.3(b).



(a) Mobile core network model.

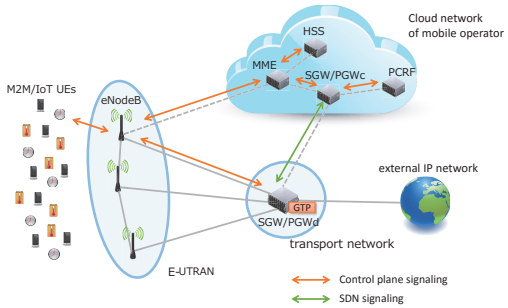


(b) Signaling flow for bearer establishment.

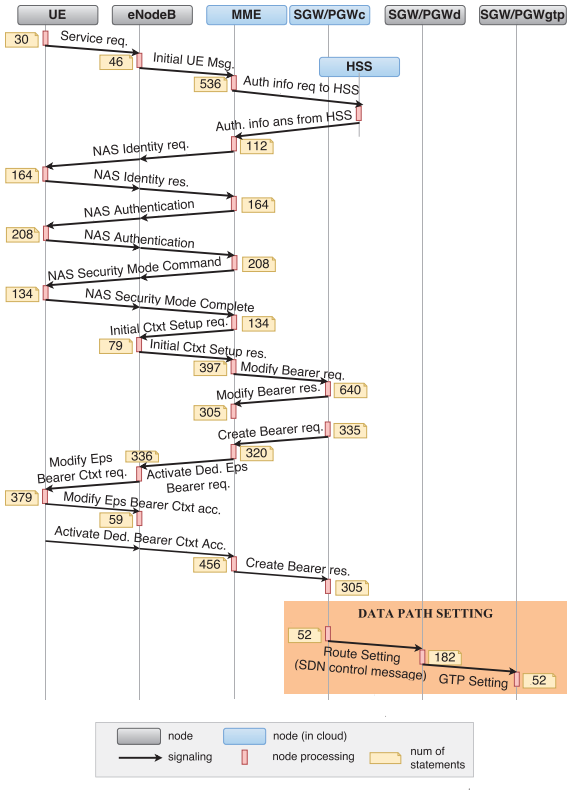
Figure 2.2: C/U Plane-separated Architecture (PSA)

2.4 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



(a) Mobile core network model.



(b) Signaling flow for bearer establishment.

Figure 2.3: C/U Plane-separated Architecture with GTP module located at Data Plane (PSA_g)

core network. In this section, a bearer aggregation method is proposed to decrease the number of concurrent bearers.

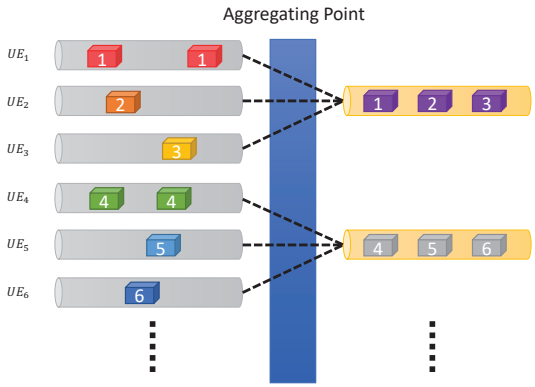
2.4.1 Overview

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. 2.4(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. 2.4(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 chapter, 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 [**?**] 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 signaling procedures such as eMTC. Moreover, the aggregation method can be applied to large data transmission in NB-IoT that utilize bearers.



(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 2.4: (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.

2.4.2 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. 2.4(b)) which represents the current status of bearer aggregation.

In the signaling flows shown in Fig. 2.1(b), Fig. 2.2(b) and Fig. 2.3(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

Table 2.2: **Positive and Negative Effect with SGW or eNodeB Aggregation:** Depending on aggregation point, required modification, network capacity and bearer establishment time are different.

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 2.3: **Positive and Negative Effect with Pre-determined or On-demand Aggregation:** Depending on aggregation timing, wait time and network capacity are different.

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

the IMSI table in Fig. 2.4(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.

2.4.3 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.

Aggregation Point

The bearers are aggregated at SGW or eNodeB. Table 2.2 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.

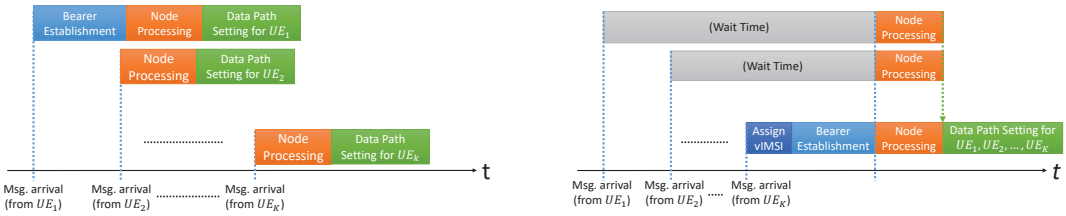
Aggregation at both SGW and eNodeB The combination of aggregations at SGW and eNodeB can be considered to further decrease the number of bearers, in which some of the aggregated S1–u bearers are again aggregated into a single aggregated S5/S8 bearer.

In this aggregation, the IMSI table in Fig. 2.4(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.

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 2.3



(a) Pre-determined aggregation method. (b) On-demand aggregation method.

Figure 2.5: (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.

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. 2.5(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. 2.5(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.

2.5 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 2.3.

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.

2.5.1 Notations

m is one of CA, PSA and PSA$_g$ defined in Section 2.3. 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. 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, M, 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. 2.4). We set K to a constant value in this evaluation for simplicity.

2.5.2 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

$$E[R] = \frac{\rho^r}{1-\rho} \cdot \frac{E[S^2]}{2E[S]} + \frac{1-\rho^r}{1-\rho} \cdot E[S], \tag{2.1}$$

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:

$$\rho = \lambda \cdot E[S]. \tag{2.2}$$

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

$$\lambda_N = \frac{P_N \cdot n_U}{D \cdot n_N}, \tag{2.3}$$

$$E[S_N] = \sum_{i=1}^{P_N} \frac{L_{N_i}}{A_N \cdot P_N}, \tag{2.4}$$

$$E[S_N^2] = \sum_{i=1}^{P_N} \frac{L_{N_i}^2}{A_N^2 \cdot P_N}. \tag{2.5}$$

2.5.3 Network Capacity

Solving (3.22) for n_U with (3.16) and (2.4), we obtain

$$n_U = D \cdot \rho \frac{n_N \cdot A_N}{\sum_{i=1}^{P_N} L_i}. \tag{2.6}$$

By substituting $\rho = 1$ for (2.6), 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.

$$n_{U_{max}} = \min_{N \in \mathbb{V}_m} \left(D \frac{n_N \cdot A_N}{\sum_{i=1}^{P_N} L_i} \right). \tag{2.7}$$

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 (2.7). 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.

$$A'_N = \left(\sum_{I \in \mathbb{W}} A_I \right) \cdot \frac{\sum_{i=1}^{P_N} L_i}{\sum_{I \in \mathbb{W}} \sum_{i=1}^{P_I} L_i} \cdot \frac{1}{n_N}. \tag{2.8}$$

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 (3.16) gets smaller. On the other hand, $E[S]$ decreases because L_i in (2.4) 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 (2.6), 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.

2.5.4 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 (3.19).

$$\begin{aligned} T &= T_\tau + T_t + T_w \\ &= \sum_{N_1,N_2 \in \mathbb{V}_m} (C_{N_1,N_2} \cdot \tau_{N_1,N_2}) + \sum_{N \in \mathbb{V}_m} (P_N \cdot t_N) + T_w, \end{aligned} \tag{2.9}$$

where T_w is calculated by (2.10) on the basis of the communication period of a UE, the number of UEs attached to the network, and the aggregation level.

$$T_w = \begin{cases} \frac{K \cdot D}{2n_U} & (\text{Aggregation at SGW}) \\ \frac{K \cdot D \cdot n_B}{2n_U} & (\text{Aggregation at eNodeB}) \end{cases} \tag{2.10}$$

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 PSA_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.

$$\begin{aligned}\frac{dT}{dn_U} &= \frac{dT_t}{dn_U} + \frac{dT_w}{dn_U} \\ &= \sum_{N \in \mathbb{V}_m} \left(P_N \frac{dt_N}{dn_U} \right) + \left(-\frac{D \cdot K}{2n_U^2} \right)\end{aligned} \tag{2.11}$$

$\frac{dt_N}{dn_U}$ is expressed as (2.12). This represents the increase in the average signaling processing time t_N at the node N when the number of UEs increases.

$$\begin{aligned}\frac{dt_N}{dn_U} &= \frac{dt_N}{d\rho} \cdot \frac{d\rho}{dn_U} \\ &= \big((2E[S]^2 - E[S^2])(r-1)\rho^r \\ &- (2E[S]^2 - E[S^2])r \cdot \rho^{r-1} + 2E[S]^2 \big) \\ &/2E[S](1-\rho)^2 \frac{d\rho}{dn_U}\end{aligned} \tag{2.12}$$

Assuming that the maximum number r of parallel processing runs corresponding to each node is one, (2.12) can be converted as

$$\begin{aligned}\frac{dt_N}{dn_U} &= \frac{P_N \cdot D \cdot n_N \cdot E[S^2]}{2(D \cdot n_N - P_N \cdot E[S] \cdot n_U)^2} \\ &= \frac{D \cdot n_N \cdot \sum_{i=1}^{P_N} L_i^2}{2\left(D \cdot n_N \cdot A_N - \sum_{i=1}^{P_N} L_i \cdot n_U \right)^2}.\end{aligned} \tag{2.13}$$

Substituting (2.13) for (2.11), we obtain

$$\frac{dT}{dn_U} = \sum_{N \in \mathbb{V}_m} \frac{P_N \cdot D \cdot n_N \cdot \sum_{i=1}^{P_N} L_i^2}{2\left(D \cdot n_N \cdot A_N - \sum_{i=1}^{P_N} L_i \cdot n_U \right)^2} - \frac{D \cdot K}{2n_U^2}. \tag{2.14}$$

From (2.13), 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_{UNmax} in (2.6) with $\rho = 1$. Therefore, with (2.14), by appropriately allocating server resources to EPC nodes as in (2.8), we can maximize the number of accommodated UEs with finite value of the bearer establishment time.

2.6 Numerical Evaluation

In this section, we show the numerical results of the analysis in Section 3.4 for evaluating the effects of the node virtualization and C/U plane separation with SDN discussed in Section 2.3 and the bearer aggregation method proposed in Section 2.4.

2.6.1 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 CA. In PSA and PSA$_g$, SGW/PGW is divided into one SGW/PGW$_c$ and SGW/PGW$_d$. One GTP module exists in the network for PSA and PSA$_g$. 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: 20 msec
- eNodeB–SGW/PGW, SGW/PGW$_d$: 7.5 msec
- eNodeB–MME, SGW/PGW$_c$: 10 msec
- SGW/PGW, SGW/PGW$_d$–MME, SGW/PGW$_c$: 10 msec
- MME–SGW/PGW$_c$: 1 msec
- GTP module–SGW/PGW$_d$: 1 msec

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

- UE: 3,000 statements/sec
- eNodeB: 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 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 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 [**?**].

We assume that in PSA and PSA$_g$, the server resources located in the cloud environment (MME 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 2.4 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.

2.6.2 Evaluation Results

Network Architecture Comparisons

Fig. 2.6 shows the relationship between the number of accommodated UEs and the bearer establishment time when we utilize the mobile core networks based on the three architecture discussed in Section 2.3. 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 three 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$_c$ decreases owing to C/U

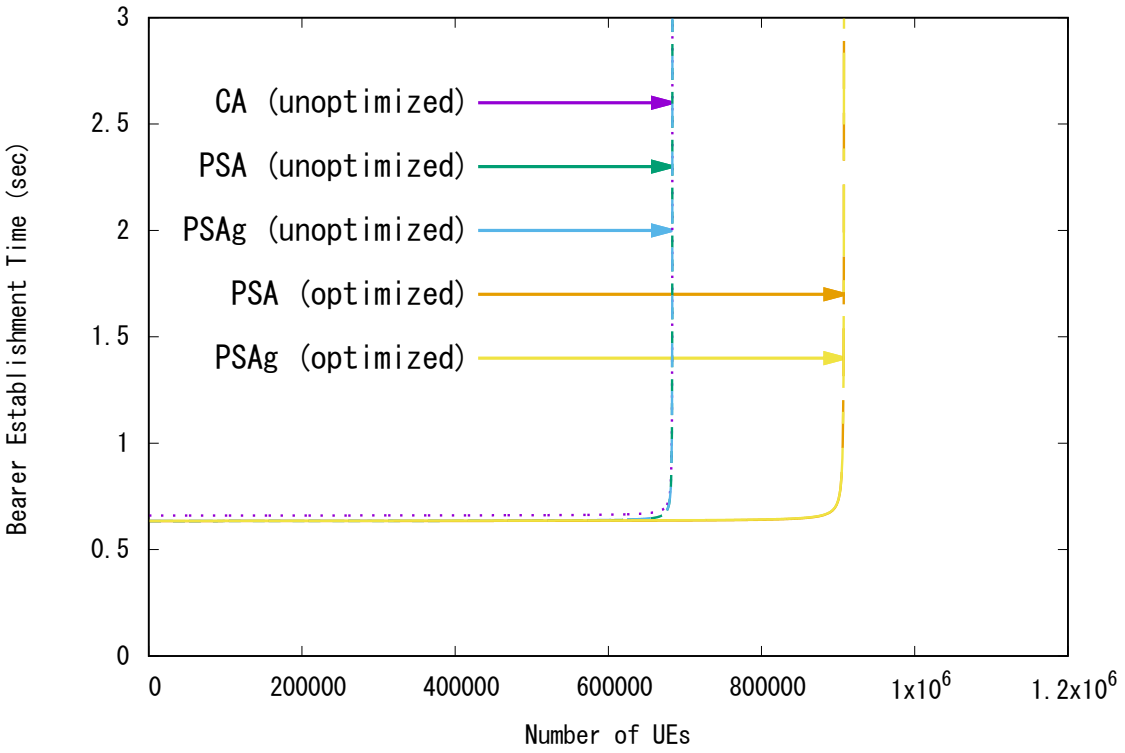
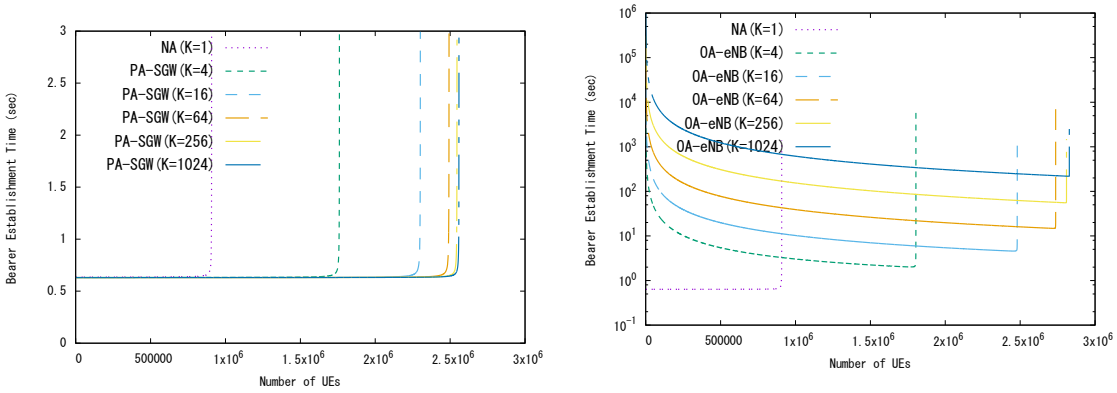


Figure 2.6: **Comparison of network architecture:** The network capacity increases by applying server resource optimization.

plane separation, the system utilization of MME remains unaffected.

A comparison of the results obtained with and without server resource optimization in PSA or PSA$_g$, 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, CA yields a slightly longer bearer establishment time than PSA or PSA$_g$. This is because, in PSA and PSA$_g$, propagation delays are reduced by placing the control plane function of SGW in the cloud environment. PSA and PSAg 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 between Fig. 2.2(b) and Fig. 2.3(b),there is one additional



(a) Pre-determined aggregation at SGW

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

Figure 2.7: **Effect of aggregation level:** The network capacity increases by applying a bearer aggregation method.

signaling message, GTP Setting between the SGW/PGW$_d$ and GTP module, in the case of PSA$_g$. Therefore, the total propagation delay of PSA$_g$ is slightly larger than that of PSA, and this difference cannot be recognized in Fig. 2.6.

In the following evaluation, PSA$_g$ is utilized.

Aggregation Level

Figs. 2.7(a) and 2.7(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. 2.7(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. 2.5(b). Equation (2.10) 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 decreases or when the number of accommodated UEs increases, the bearer establishment time decreases.

Combination of Server Resource Optimization and Bearer Aggregation

Fig. 2.8 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. 2.8, 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.

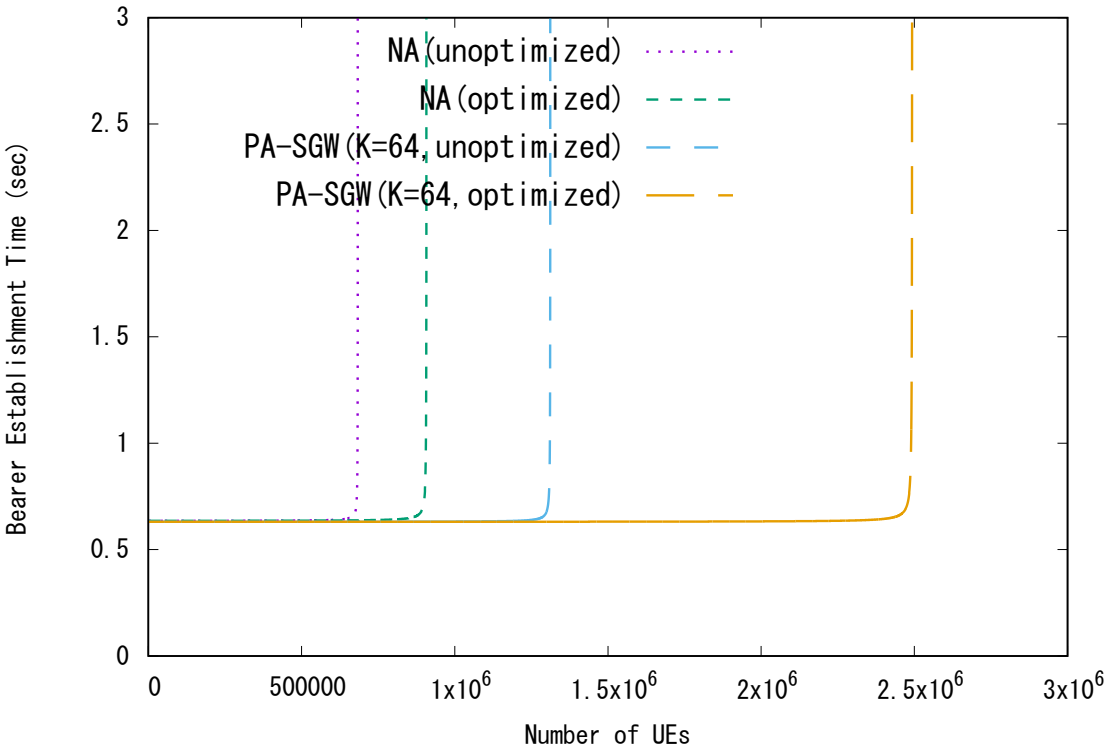
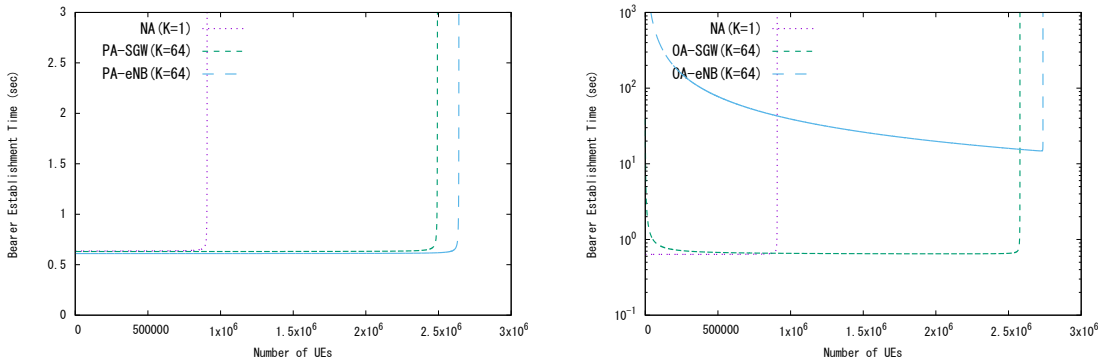


Figure 2.8: **Effect of combination of server resource optimization and bearer aggregation:** In terms of network capacity, the simultaneous application of server resource optimization and bearer aggregation yields better performance than applying them individually.

Aggregation Point

Fig. 2.9 shows a comparison of the performance of bearer aggregation at SGW and that at eNodeB. Figs. 2.9(a) and 2.9(b) plot the results of pre-determined aggregation and on-demand aggregation, respectively.

As shown in Fig. 2.9(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 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



(a) With pre-determined aggregation.

(b) With on-demand aggregation.

Figure 2.9: **Effect of aggregation point:** The aggregation at eNodeB results in larger network capacity than that at SGW.

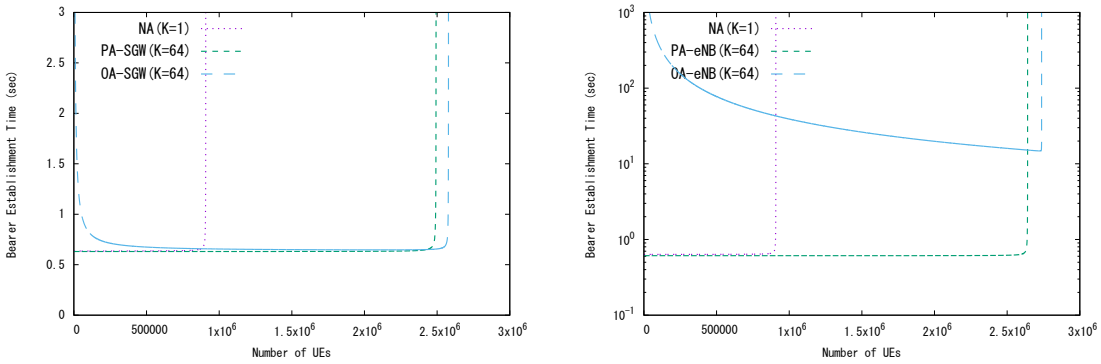
becomes apparent. Fig. 2.9(b) shows that with on-demand aggregation, the aggregation at eNodeB also yields higher network capacity, albeit with a substantially longer bearer establishment time. This is because of the abovementioned long waiting time associated with the on-demand aggregation.

Aggregation Timing

Fig. 2.10 shows similar results for the sake of comparing the pre-determined aggregation and on-demand aggregation schemes. Figs. 2.10(a) and 2.10(b) show plots of the results of aggregation at SGW and eNodeB, respectively.

As shown in Fig. 2.10(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. 2.10(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 2.6.2.



(a) With the bearer aggregation at SGW (b) With the bearer aggregation at eNodeB

Figure 2.10: **Effect of aggregation timing:** The on-demand aggregation scheme outperforms the pre-determined aggregation scheme in terms of network capacity.

Aggregation at both SGW and eNodeB

Fig. 2.11 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. 2.11 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 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.

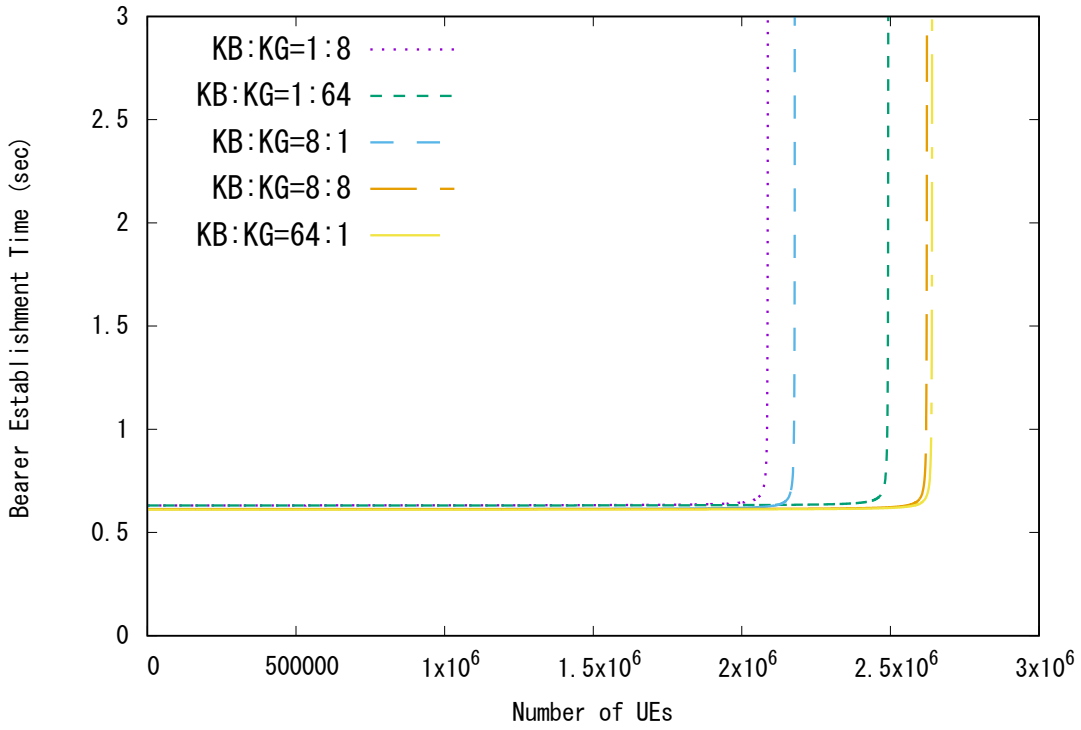


Figure 2.11: **Effect of aggregation at both eNodeB and SGW:** The performance gain becomes different depending on the aggregation level at each aggregation point.

2.7 Discussion

2.7.1 Server Resource Optimization

From the results in Subsection 2.6.2, 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.

2.7.2 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

Table 2.4: **Recommended setting and obtained performance:** Varies according to the characteristics of UEs.

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

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.

2.7.3 Aggregation Point

The results described in Subsection 2.6.2 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. 2.9(a), when applying the pre-determined aggregation scheme, the bearer establishment time with the bearer aggregation at eNodeB is smaller than that at SGW. For supporting these discussions, Fig. 2.12 shows plots of changes in the total processing times of signaling messages (T_t in (3.19)) 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 (3.19)) on the bearer establishment time is stronger, as shown in Fig. 2.9(b).

2.7.4 Aggregation Timing

The results described in Subsection 2.6.2 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

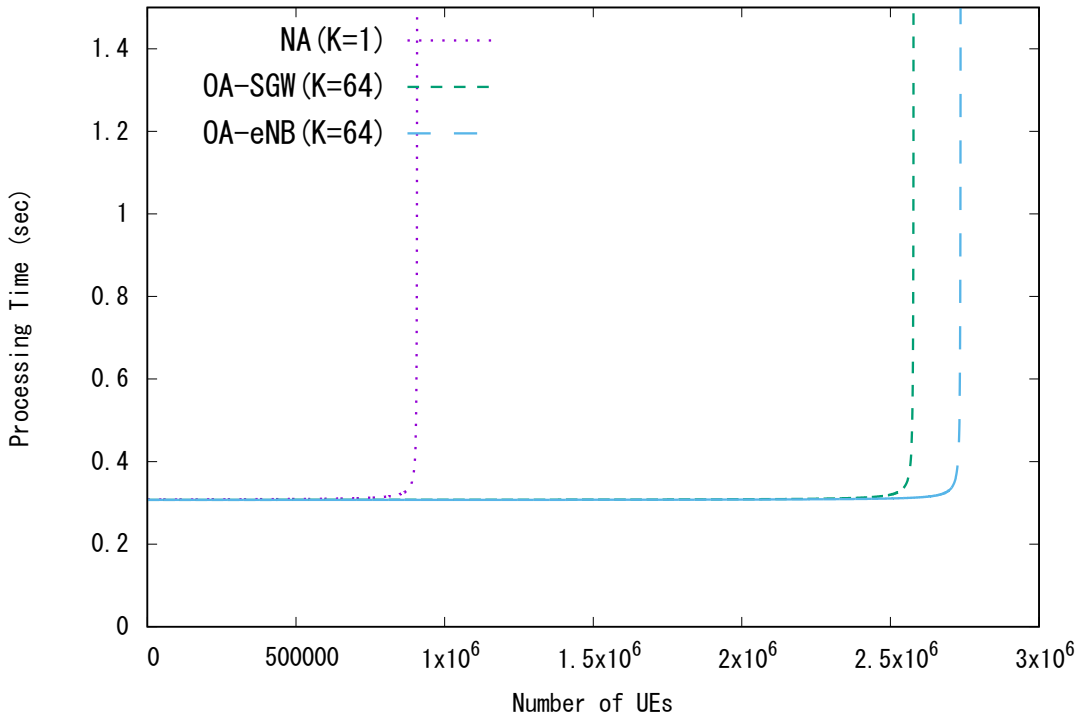


Figure 2.12: **Processing time comparison:** The on-demand aggregation at eNodeB outperforms the on-determined aggregation at SGW in terms of processing time.

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.

2.7.5 UE's mobility

In this chapter, 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.

2.7.6 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 2.4 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.

2.7.7 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.

2.7.8 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 chapter can reduce the load on C-plane.

2.8 Conclusion

In this chapter, 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 2.4).

2. We presented an analysis for evaluating the performance of mobile core networks (Section 3.4).

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. 2.7(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. 2.9, 2.10 and 2.8).

5. We discussed appropriate settings for the aggregation method in accordance with the characteristics of M2M/IoT terminals (Table 2.4).

The result of this chapter can be applied to the optimization of packet core subslices in User Oriented Network slicing Architecture proposed in Chapter 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 experimentation where massive UEs connects to the mobile network simultaneously. In future work, we plan to evaluate the effect of virtualization and C/U plane separation at eNodeB, and the relationship between an aggregation level and the performance in U-plane such as latency. 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.

Chapter 3

Performance Analysis of Mobile Cellular Networks Accommodating Cellular-IoT Communications with Immediate Release of Radio Resources

3.1 Introduction

Traditionally, mobile networks mainly accommodate human devices, such as smartphones and feature phones. Therefore, the control- and data-plane mechanisms of the current mobile cellular networks, such as contention-based random access, radio resource allocation, and bearer establishment procedure, are predominantly optimized for human communications. On the other hand, mobile cellular networks are now required for serving numerous Internet of Things (IoT) communications. However, IoT communications cannot generally be efficiently accommodated because IoT traffic has different characteristics from human communications. In particular, the Third-Generation Partnership Project (3GPP) assumes that most cellular IoT traffic results from periodic data transmissions of Mobile Autonomous Reporting (MAR) such as sensors and smart meters [17, 33]. In [17], a typical example of an inter-arrival time distribution for MAR is presented as: 1 day (40%), 2 hours (40%), 1

hour (15%), and 30 minutes (5%). It is clear that conventional network designs for human devices are not suitable for efficiently accommodating such communications.

For this reason, standardization activities for connecting IoT devices to cellular networks, such as Cellular-IoT (C-IoT) [34], have emerged. C-IoT can be deployed over an existing infrastructure, and is advantageous in terms of maintaining security and managing radio interference. Among C-IoT radio technologies, Narrowband-IoT (NB-IoT) is optimized for most IoT communications that transmit small data. Moreover, in standardization activities, the immediate release of radio resources after the completion of data transmissions is considered [7], while radio resources are kept allocated at least 10 seconds in traditional mobile cellular networks and NB-IoT. The immediate release of radio resources would have a considerable effect in accommodating IoT communications with small data transmissions. However, few existing studies have evaluated the performance of C-IoT communications with periodic data transmissions in terms of the immediate release of radio resources.

In this chapter, we perform an analysis of mobile cellular networks to support periodic C-IoT communications, focusing on the performance differences between LTE and NB-IoT networks. In detail, we show the mathematical analysis of both data- and control-plane performance, including random access procedures, radio resource allocation, bearer establishment, and user-data transmission. We adopted a two-dimensional Markov chain model to analyze the behavior of User Equipments (UEs) in a random access procedure. We also exploited the queuing theory for the derivation of the failure probability of radio resource allocation and for the time required for bearer establishment. We also include the effect of the immediate release of radio resources to the analysis. We give numerical evaluation results of the analysis in order to compare the performance of the LTE and NB-IoT networks, clarify the effect of data transmission cycle, and reveal the effect of immediate release of radio resources.

In [35], our research group performed the mathematical analysis of both data- and control- plane performance, including random access procedures, radio resource allocation, bearer establishment, and user-data transmission. In this chapter, we extend our analysis to obtain the capacity of the mobile cellular network for accommodating IoT communications as closed formulas. In addition, in the numerical evaluations, we added the discussions on the effect of the communication cycle and transmission data size.

The remainder of this chapter is organized as follows. In Section 4.5, we summarize related studies. In Section 3.3, we explain the model of the mobile cellular network and the detailed behavior of UEs for performance analysis. In Section 3.4, we present the mathematical analysis. In Section 3.5, we provide numerical evaluation results of the analysis and present our discussion. In Section 3.6, we explain the effectiveness of our analyses, results and discussions in 5G network. We conclude this chapter in Section 4.6 with a brief summary and perspectives for future work.

3.2 Related Work

In order to improve the capacity of mobile cellular networks for IoT communications, various methods have been proposed. These existing works, as well as the proposed method in this chapter, are summarized in Table 3.1.

Studies [36–38] focus on accommodating periodic data transmissions in the mobile cellular network. In [36], the power consumption with applying the immediate release of radio resources for NB-IoT was mainly studied for comparing the performance with other LPWAN technologies. However, they did not consider the collisions of preamble transmissions and radio resource allocation. [38] evaluated the capacity and coverage performance of LTE-M and NB-IoT considering the effect of both of the collision of preamble transmissions and radio resources allocation. However, the immediate release of radio resources was out of their focus. [39] compared the performance of NB-IoT with Long Term Evolution (LTE) in rural and urban areas while considering various data size. However, the effect of immediate release of radio resources and periodic communication are not considered. The authors of [40–42] evaluated the collision of preamble transmissions in detail with numerical analysis or simulation. However, these works do not compare the multiple candidates for C-IoT networking technologies, and there is no consideration of periodic communications and immediate release of radio resources. In summary, few existing studies evaluate the performance of the C-IoT networks with periodic data transmissions such as MAR traffic and the effect of the immediate release of radio resources.

In [35], our research group analytically evaluated periodic C-IoT communication with immediate release of radio resources. However, in [35], the types of parameters in the

Table 3.1: Summary of existing research

	[36]	[37]	[38]	[39]	[40–42]	[35]	Proposed
Periodic communications	✓	✓	✓			✓	✓
Immediate RRC connection release	✓					✓	✓
Preamble collision			✓		✓	✓	✓
Radio resource allocation		✓	✓			✓	✓
Joint analysis of core and radio access networks						✓	✓
Comparison of LTE and NB-IoT			✓	✓		✓	✓
Closed formulas of network capacity for IoT communications							✓

numerical evaluations were limited and these evaluations were based on simple analyses with severe assumptions. In the numerical evaluations in this chapter, we added the discussions on the effect of the communication cycle and transmission data size. In addition, we extend our analysis to obtain the capacity of the mobile cellular network for accommodating IoT communications as closed formulas.

3.3 Network Model

The mobile network architecture used in this chapter is shown in Fig. 3.1. It consists of two networks, an evolved universal terrestrial radio access network (E-UTRAN) and Evolved Packet Core (EPC). When a UE transmits data, it first conducts a random access procedure in order to obtain radio resources. Then a bearer establishment procedure is executed to set up logical transmission paths in the data plane. Finally, the UE sends user data. In what follows, we describe the detailed behavior of the UE and mobile core network in detail.

3.3.1 Random Access Procedure and Radio Resource Allocation

E-UTRAN consists of user equipments (UEs) and eNodeBs. An eNodeB implements a remote radio head and accommodates multiple IoT UEs, which transmit data periodically. We assume that all UEs transmit data of the same size with identical periodic intervals.

When the UE starts data transmission, it conducts a contention-based random access, termed the Radio Resource Control (RRC) connection setup procedure, in both the LTE and NB-IoT [34, 43]. This process consists of the following four steps:

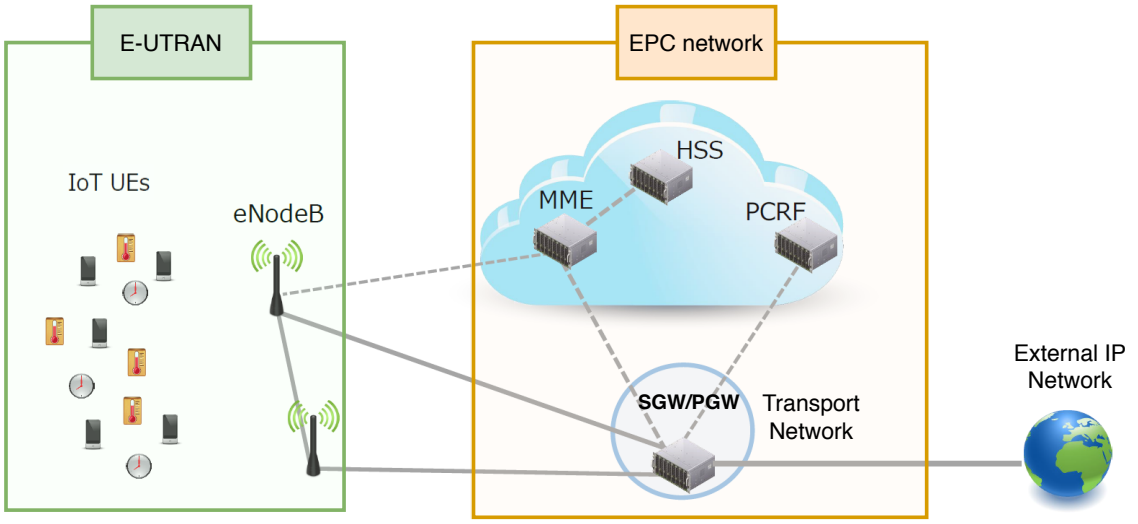


Figure 3.1: Mobile network architecture.

1. The UE transmits a preamble to the eNodeB, which is randomly selected from multiple preambles reserved in the cell. The number of preambles is 54 in LTE and 48 in NB-IoT.

2. When the eNodeB receives the preamble, it transmits a Random Access Channel (RACH) response message to the corresponding UE.

3. The UE that has received the RACH response message sends an RRC connection request message to the eNodeB. This request message contains the identity of the UE.

4. After receiving the RRC connection request message, the eNodeB transmits an RRC connection setup message to the UE that includes cell setting information.

When the UE receives the RRC connection setup message including its identity, the random access procedure has successfully finished. Otherwise, the random access has failed and the procedure is performed again after a back off time.

When the UE completes the RRC connection setup, radio resources are allocated to the UE. Radio resources and RRC connection are kept allocated until the inactivity timer expires. The typical value of an inactivity timer is 10 s [44]. In 3GPP, the immediate release of radio resources after data transmissions is considered [34], such that the radio

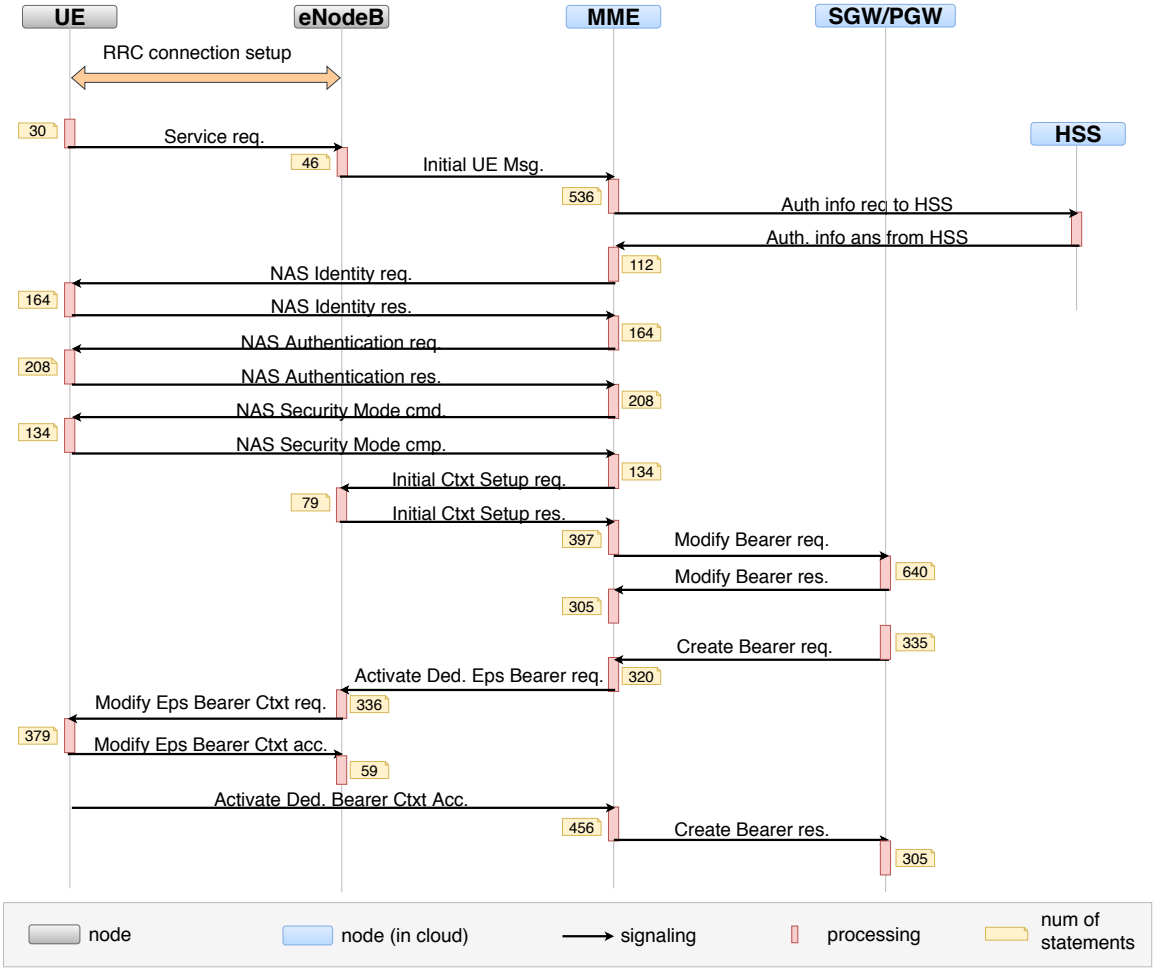


Figure 3.2: Signaling flow for bearer establishment.

resource and RRC connection are immediately released when the UE completes the data transmission. It would be effective for IoT communications with small data transmissions, that is evaluated in this chapter.

3.3.2 Bearer Establishment Procedure

After the random access procedure, the UE performs the user-plane bearer establishment procedure. Figure 3.2 shows the signaling flow to establish user-plane bearers in EPC. EPC consists of 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). In this figure, req. and res. represent a request and a response message, respectively. Msg. stands for "message". Ctxt, Ded., and Acc mean "Context", "Dedicated", and "Accept", respectively. The figure includes the number of statements executed by each EPC node for processing each signaling message in the programs. These numbers are obtained by analyzing the source code of OpenAirInterface [32]. Note that each processing of signaling messages has a different number of statements, meaning that each message imposes a different load on the corresponding EPC node.

3.3.3 User Data Transmission

After the bearer establishment procedure, the UE transmits its user data, using the radio resources allocated in RRC connection setup procedure. In this chapter, we assume that the amount of radio resources allocated to the UE is minimal regardless of the transmission data size.

3.4 Analysis

We give the analysis for the average of *service time*, which is defined as the time between the moment at which the a UE starts a random access procedure and when it completes a bearer establishment procedure and data transmission. The service time $t_{service}$ is the sum of *random access time* t_r, *bearer establishment time* t_b and *data transmission time* t_d, as presented Eq. (3.1).

$$t_{service} = t_r + t_b + t_d \tag{3.1}$$

In Subsections 3.4.1, 3.4.2, and 4.3, we derive t_r, t_b, and t_d, respectively.

We also analyze the network capacity which is defined as the number of IoT devices accommodated in the mobile network. The network capacity n_M is the smaller of *radio access network capacity* n_{M_r} and *EPC network capacity* n_{M_E}, as presented Eq. (3.2).

$$n_M = \min\left(n_{M_r}, n_{M_E}\right) \tag{3.2}$$

In Subsections 3.4.4 and 3.4.5, we derive n_{M_r} and n_{M_E}, respectively.

3.4.1 Random Access Time

As explained in Subsection 3.3.1, when the UE starts a random access procedure, it conducts a contention-based random access, and radio resource allocation is performed if it is done successfully. After data transmission with the allocated radio resource, the UE waits for the next communication timing. On the other hand, when a collision occurs in the preamble transmission or when the radio resource allocation fails, the UE restarts preamble transmission after waiting for a *MAC-level back off* whose length is randomly determined within a range of up to B_M *slots*. Note that the length of a slot is equal to 0.5 *ms* in both LTE and NB-IoT [34]. This procedure is repeated up to $N_M - 1$ times until both the preamble transmission and radio resource allocation have been successfully completed. In this chapter, we also consider the *application-level back off* after $N_M - 1$ times of MAC-level back off. For simplicity, the length of the application back off is fixed at B_A *slots*.

State Transition Diagram

We exploit a two dimensional Markov chain to model the above-described behavior of a UE, depicted in Fig. 3.3. This is based on the models in [41,42], which we have extended to include MAC-level back off, application-level back off, and wait time after data transmission for periodic communications. All state transitions take 1 *slot*. In Fig. 3.3, T_c is the data transmission cycle of UEs. B_M *slots* is the maximum time for the MAC-level back off. B_A *slots* is the time for the application-level back off. p_f is the probability that the preamble transmission collides or succeeds without collision but the radio resource allocation fails nonetheless. $p_{i,j}$ represents the probability of transition from state $(i-1, 0)$ to state (i, j). No description of transition probability indicates that the transition probability is 1.0. We assume that data transmission is completed prior to beginning the transmission of the next cycle. Therefore, we consider that $T_c \geq t_r$ is always satisfied.

When a UE is in states such as $(1, 0), \cdots, (N_M, 0)$, depicted in orange ellipses in Fig. 3.3, the UE transmits a preamble. When the preamble transmission and radio resources allocation have been conducted successfully, the UE waits for the next data transmission, depicted by blue arrows. On the other hand, when the UE fails the preamble transmission or radio resources allocation, it begins the MAC-level back off, depicted by orange arrows.

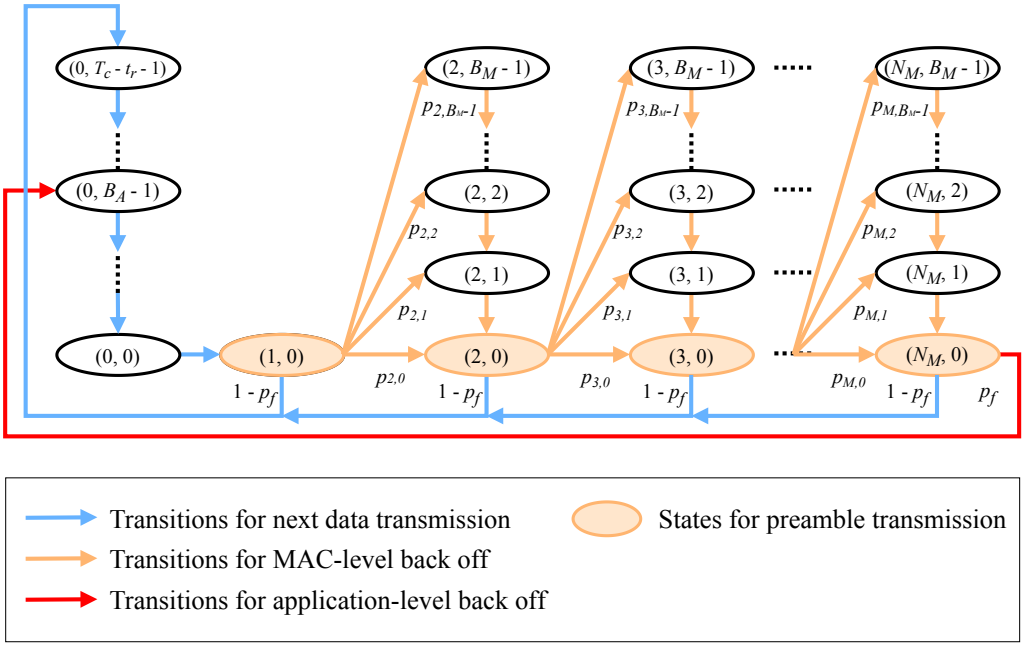


Figure 3.3: State transition diagram for random access procedure

When these failures are repeated N_M times, an application-level back off occurs, depicted by a red arrow.

By calculating the probability of the UE being in each state, defined as the state probability, we obtain the average time required for the random access procedure. According to Fig. 3.3, when the probability of state (i, j) is denoted by $\pi_{(i,j)}$, the following equations are satisfied for $i = 0, 1, \cdots, (T_c - t_r - 1)$, $j = 2, 3, \cdots, N_M$, and $k = 0, 1, \cdots, (B_M - 1)$.

$$\pi_{(0,i)} = \pi_{(0,0)} \tag{3.3}$$

$$\pi_{(1,0)} = \pi_{(0,0)} \tag{3.4}$$

$$\begin{aligned} \pi_{(j,k)} &= \sum_{i=k}^{B_M-1} \pi_{(j-1,0)} p_{j,i} \\ &= \frac{B_M - k}{B_M} {p_f}^{j-1} \pi_{(0,0)} \end{aligned} \tag{3.5}$$

Since the sum of the probabilities of all states must equal 1.0, the following equation is

satisfied:

$$\sum_{i=0}^{T_c-t_r-1} \pi_{(0,i)} + \pi_{(1,0)} + \sum_{j=2}^{N_M} \sum_{k=0}^{B_M-1} \pi_{(j,k)} = 1 \tag{3.6}$$

According to Eqs. (3.3)–(3.6), the state probability of the initial state (0,0) is derived by Eq. (3.7).

$$\pi_{(0,0)} = \frac{2(1-p_f)}{2(T_c - t_r + 1)(1-p_f) + (B_M+1)p_f(1-{p_f}^{N_M-1})} \tag{3.7}$$

In order to calculate the state probabilities, it is necessary to obtain p_f and t_r; these are not given as system parameters. p_f is derived using Eq. (3.8), where p_r is the failure probability of the radio resource allocation and p_c is the collision probability of the preamble transmission.

$$p_f = 1 - (1-p_r)(1-p_c) \tag{3.8}$$

p_r and p_c are given below. In order to calculate t_r, we consider the detailed behavior of MAC- and application-level back offs. Since an application-level back off occurs after a MAC-level back off is repeated $(N_M - 1)$ times, t_r is given by the following equations, where t_M is the average time for MAC-level back offs with successful preamble transmission, and t_A is the average time for an application-level back off with $(N_M - 1)$ times of MAC-level back offs.

$$t_r = \sum_{k=1}^{n_A} \left({p_f}^{N_M}\right)^{k-1} ((k-1)t_A + t_M) \tag{3.9}$$

$$t_M = 1 + \frac{B_M+1}{2} \sum_{i=1}^{N_M-1} {p_f}^i \tag{3.10}$$

$$t_A = 1 + \frac{B_M+1}{2}(N_M - 1) + B_A \tag{3.11}$$

Preamble Transmission Probability

In this study, we assume that when two or more UEs transmits the same preamble at the same time, a collision occurs and that all colliding preamble transmissions fail. In this case,

the UEs perform MAC-level back off.

Therefore, p_c, i.e., the collision probability of preamble transmission, can be obtained using Eq. (3.12), where n_U is the number of UEs accommodated in the eNodeB, P is the number of preambles and p_t is the probability at which the UE is in a state of preamble transmission.

$$p_c = 1 - \left(\frac{P-1}{P}\right)^{(n_U-1)p_t} \tag{3.12}$$

According to Fig. 3.3, p_t is derived by Eq. (3.13).

$$p_t = \sum_{i=1}^{N_M} \pi_{(i,0)} \tag{3.13}$$

With Eqs. (3.5) and (3.7), this equation can be expanded as follows.

$$p_t = \frac{2(1-p_f^{N_M})}{2(T_c - t_r + 1)(1-p_f) + (B_M+1)p_f(1-p_f^{N_M-1})} \tag{3.14}$$

Radio Resource Allocation

Even when a UE succeeds in preamble transmission, data transmission can be made only when radio resource allocation has been successfully conducted, otherwise, the UE goes into a MAC-level back off.

The number of resource blocks per each frame in LTE and NB-IoT with a given radio bandwidth is denoted by n_R. For simplicity, it is assumed that the number of resource blocks per each frame allocated to each UE is fixed at n_{R_U}. The maximum number of UEs to which the network can concurrently allocate the radio resource is then $\left\lfloor \frac{n_R}{n_{R_U}} \right\rfloor$. We also assume that the radio resource is kept allocated until the inactivity timer expires. Furthermore, for simplicity, we assume that a Poisson arrival process of UEs that successfully transmit their preambles. We then employ the $M/D/K/K$ queuing model [45] to derive the failure probability for radio resource allocation. In the $M/D/K/K$ model, the steady-state

distribution of the number of jobs in a system is represented by Eq. (3.15),

$$P(L=n) = \frac{\frac{(\lambda\tau)^n}{n!}}{\sum_{i=0}^{K} \frac{(\lambda\tau)^i}{i!}} \qquad (0 \leq n \leq K) \tag{3.15}$$

where λ is the job arrival rate, τ is the serving time, K is the number of servers, and L is the number of jobs in the system. In our model, $P(L = K)$ corresponds to the failure probability of radio resource allocation, p_r, which has the following parameter calculations:

$$\lambda = n_s n_U p_t (1 - p_c) \tag{3.16}$$

$$\tau = T_i \tag{3.17}$$

$$K = \left\lfloor \frac{n_R}{n_{R_U}} \right\rfloor \tag{3.18}$$

where n_s is the number of slots per second and T_i is the inactivity timer discussed in Subsubsection 3.4.1. When we employ the immediate release of radio resources, we set the value of T_i to $t_{service}$. This means that $t_{service}$ and p_r are dependent upon one another. Therefore, when obtaining numerical results, we make iterations for the calculation of $t_{service}$ and p_r until both values converge.

p_f is derived by (3.8), (3.12), (3.14) and (3.15). Then, we obtain t_r from substituting p_f for (3.9).

3.4.2 Bearer establishment time

Based on the signalling flow in Fig. 3.2, the bearer establishment time t_b is the sum of the propagation delays of all signaling messages, denoted by t_τ, and the processing times of all messages, denoted by t_p.

$$t_b = t_\tau + t_p \tag{3.19}$$

t_τ is obtained as the following equation, by calculating the sum of the propagation delays in the signaling flow depicted in Figure 3.2.

$$t_\tau = \sum_{N_1,N_2 \in V} (C_{N_1,N_2} \tau_{N_1,N_2}) \tag{3.20}$$

where C_{N_1,N_2} is the number of the signaling message and τ_{N_1,N_2} is the average propagation delay between the EPC nodes N_1 and N_2.

For obtaining t_p, we use the analysis results of the bearer establishment time in our previous study [46], while we modify the calculation of the average processing time.

To derive the average processing time at a mobile core node, we employ the M/G/1/PS queuing model [47] and assume that the arrival of signaling messages at a node follows a Poisson distribution. In the M/G/1/PS model, the mean sojourn time $E[R]$ can be derived as follows Eq. (3.21):

$$E[R] = \frac{\rho^r}{1-\rho} \frac{E[S^2]}{2E[S]} + \frac{1-\rho^r}{1-\rho} E[S] \tag{3.21}$$

where λ is job arrival rate, $S(x)$ is workload distribution, $E[S]$ is the mean workload, and r is the maximum number of parallel processing runs. ρ is the system utilization, and is given by the following equation.

$$\rho = \lambda E[S] \tag{3.22}$$

In our analysis, a job to be processed at a server corresponds to a signaling message to be processed at a mobile core node. The workload of each job corresponds to the number of statements in the program executed at the mobile core node for processing the signaling message, and this is presented numerically in Fig. 3.2. The job arrival rate at node N, denoted by λ_N, corresponds to the number of signaling messages arriving per unit time at node N. Note that the jobs for the bearer establishment arrive only from UEs that successfully finish the preamble transmission and radio resource allocation. Consequently, the time distribution for processing signaling messages at node N, denoted by S_N, corresponds to the distribution of the workload. Accordingly, 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 resource of node N. λ_N, $E[S_N]$, and $E[S_N^2]$ may therefore

be derived as Eqs. (3.23)–(3.25),

$$\lambda_N = \frac{P_N n_U n_E (1 - p_f) p_t}{T_c} \tag{3.23}$$

$$E[S_N] = \sum_{i=1}^{P_N} \frac{L_{N_i}}{R_N P_N} \tag{3.24}$$

$$E[S_N^2] = \sum_{i=1}^{P_N} \frac{L_{N_i}^2}{R_N^2 P_N} \tag{3.25}$$

where P_N is the number of messages processed at node N in the signaling flow, n_E is the number of eNodeBs accommodated in the EPC, R_N is the amount of server resource of node N in terms of the number of statements that can be processed per unit time, and L_{N_i} is the number of the statements of the program involved in the processing of a i th signaling message at node N.

Then, t_p is calculated as follows.

$$t_t = \sum_{N \in \mathbb{V}} (P_N t_N) \tag{3.26}$$

where t_N is the average processing delay of the signaling message at node N, that is obtained by substituted Eqs. (3.22), (3.24), (3.25) into Eq. (3.21).

3.4.3 Data Transmission Time

When the link speed between nodes N_1 and N_2 is W_{N_1,N_2} and the message size is C, a time $t_{d_{N_1,N_2}}$ is required to transmit the message; this is derived by Eq. (3.27).

$$t_{d_{N_1,N_2}} = \frac{C}{W_{N_1,N_2}} \tag{3.27}$$

The data transmission time t_d is the sum of all data and signaling message transmissions and is derived as Eq. (3.28),

$$t_d = \sum_{N_1,N_2 \in \mathbb{V}} \left(n_{d_{N_1,N_2}} \cdot t_{d_{N_1,N_2}} \right) \tag{3.28}$$

where $\mathbb{V}$ represents a set of nodes which transmit data and signaling messages and $n_{d_{N_1,N_2}}$ is the number of data and signaling messages transmitted from N_1 to N_2.

3.4.4 Radio access network capacity

The bottleneck of the radio access network capacity is determined by either the collisions of the preamble transmissions or the shortage of the radio resources, as described in Eq. (3.8).

However, in this chapter we consider only the shortage of the radio resources since p_r is quite small compared with p_c in Eq. (3.8), meaning that the effect of the preamble transmission collisions can be ignored with standard configurations of LTE and NB-IoT.

Therefore, we calculate n_{M_r} as the number of IoT devices that gives nearly equal to 1 for the failure probability of radio resource allocation described in Subsection 3.4.1. That is, n_{M_r} is obtained, using Eqs. (3.15)–(3.18), as n_U that satisfies $P(L = K) = 1 - \epsilon$, where ϵ is a margin parameter for avoiding the overload of the radio access network.

3.4.5 EPC network capacity

The maximum number of UEs that can be accommodated in the EPC n_{M_E} is equal to the number of UEs when the utilization of any node in the EPC reaches 100%.

From Eqs. (3.22)–(3.25), n_{M_E} can be expressed as Eq. (3.29),

$$n_{M_E} = \min_{N \in \mathbb{V}} \left(\frac{D n_N A_N}{n_E p_s p_t \sum_{i=1}^{P_N} L_{N_i}} \right) \tag{3.29}$$

where, $\mathbb{V}$ is a set of EPC nodes.

3.5 Numerical Evaluation

3.5.1 Parameter Settings

The parameter settings are summarized in Table 3.2. Note that the number of RBs per each frame (n_R) is determined by the radio bandwidth of both LTE and NB-IoT, assuming that NB-IoT is deployed in In-band mode.

Table 3.2: Parameter Settings for Numerical Evaluation

Variable	Description	LTE	NB-IoT
T_c	Data transmission cycle	1, 10, 7,715 *s*	1, 10, 7,715 *s*
T_i	Inactivity timer	10 *s*	10 *s*
B_M	Maximum length of MAC-level back off	20 *ms*	20 *ms*
n_M	Maximum times of MAC-level back off	10	10
P	The number of preambles	54	48
B_A	Length of application-level back off	1 *s*	1 *s*
n_R	The number of resource brocks per frame	100 *RBs/frame*	100 *RBs/frame*
n_{R_U}	The number of resource brocks allocated for each UE	6 *RBs/UE/frame*	2/3 *RBs/UE/frame*
C	Transmission data size	50, 200, 2,000, 20,000 *Bytes*	50, 200, 2,000, 20,000 *Bytes*

The propagation delays and bandwidths of links between EPC nodes are determined as follows. Note that the propagation delays do not include the processing time of signaling messages. The bandwidth of the link between UE and eNodeB is determined based on n_R and n_{R_U}.

- UE–eNodeB: 20 *ms*, 22.8 *Mbps* (for LTE) / 0.106 *Mbps* (for NB-IoT)
- eNodeB–SGW/PGW: 7.5 *ms*, 1000 *Mbps*
- eNodeB–MME: 10 *ms*, 1000 *Mbps*
- SGW/PGW–MME: 10 *ms*, 1000 *Mbps*

The amount of resources of EPC nodes are set as follows.

- UE: 1.5×10^6 *statements/s*
- eNodeB: 3.0×10^6 *statements/s*
- MME: 1.0×10^7, 8.0×10^7, 3.2×10^8 *statements/s*
- SGW/PGW: 1.0×10^7, 8.0×10^7, 3.2×10^8 *statements/s*

We assume that SGW and PGW are implemented as a single node, as in the state-of-the-art implementation design of EPC. We denote the integrated node as a SGW/PGW in what follows. We also assume that a MME is virtualized and located in the cloud environment owned by the mobile network operator, whereas eNodeBs and a SGW/PGW are located in the transport network without virtualization.

3.5.2 Results and Discussions

Comparison of LTE and NB-IoT

Figures 3.4 and 3.5 plot the results of $t_{service}$, that is the service time in Eq. (3.1), as a function of the number of accommodated UEs in the network. Figure 3.4 is for the case of LTE and Fig. 3.5 is for NB-IoT. Each figure has three graphs, each of which has a different setting of the amount of server resources of MME and SGW/PGW (1.0×10^7, 8.0×10^7, and 3.2×10^8 *statements/s*). Also, each graph has the results for different values of transmission data size, C (50, 200, 2,000, and 20,000 *Bytes*).

In all graphs in Figs. 3.4 and 3.5, we can observe that the service time increases sharply when the number of accommodated UEs reaches a certain value. That number means the upper limit on the number of UEs that can be accommodated in the network. We define that number as a system capacity.

By comparing Figs. 3.4 and 3.5, we can observe that the service time of LTE is smaller than that of NB-IoT. Also, in LTE, the transmission data size has quite small effect on the service time, while in NB-IoT, the increase of the transmission data size significantly affects the service time. This is because of the difference in the number of resource blocks per each frame for a UE, that is 6 *RBs* for LTE and 2/3 *RBs* for NB-IoT.

On the other hand, when comparing Figs. 3.4(b) and 3.5(b), and Figs. 3.4(c) and 3.5(c), NB-IoT outperforms LTE in terms of the system capacity. This is also because of the difference of the number of resource blocks for a UE per each frame. However, comparing Figs. 3.4(a) and 3.5(a), the system capacity of both networks are identical. This is because the amount of resources of EPC nodes is the bottleneck of the system.

Effect of the amount of resources of EPC nodes

By comparing Figs. 3.4(a) and 3.4(b), we can see that the system capacity increases significantly when the amount of resources of EPC nodes increases. However, by comparing Figs. 3.4(b) and 3.4(c), the system capacity remains unchanged. The reasons of these results are as follows. In Fig. 3.4(a), the bottleneck of the system is at the EPC nodes. Therefore, The system capacity increases when the amount of resources of EPC nodes increases. On the other hand, in Fig. 3.4(b), the bottleneck of the system is at the radio access network,

not at EPC nodes. Consequently, the system capacity does not increase by increasing the amount of resources of EPC nodes.

On the other hand, in Fig. 3.5, the system capacity increases as the amount of resources of EPC nodes increases, since the bottleneck of the system remains at the EPC nodes. This is because of the small amount of resource blocks allocated to each UE in NB-IoT, that avoids the radio access network to become the bottleneck of the system.

Data transmission cycle

Figure 3.6 plots the result for different settings of T_c, the data transmission cycle. Note that 7,715 *s* of the data transmission cycle is calculated from the typical distribution of transmission cycle of IoT devices described in Section 3.1. In this figure, we set the amount of resources of EPC nodes to 6.4×10^8 *statements/s*, so that EPC nodes do not become the bottleneck of the system. Therefore, the system capacity found in this figure is determined by the wireless network capacity. This figure clearly shows that the system capacity increases as the data transmission cycle increases. This is simply because the concentration of transmission request from UEs is decreased by increasing the transmission data cycle.

Immediate release of radio resources

Figure 3.7 is the results for assessing the effect of the immediate release of radio resources. Figures 3.7(a) and 3.7(b) are for 10 *s* and 1 *s* of the data transmission cycle (T_c), respectively. The amount of resources of EPC nodes is set to 6.4×10^8 *statements/s*. In each graph, T_i indicates the value of the inactivity timer and $T_i = \mathsf{immediate}$ represents the result with immediate release of radio resources. In both figures, we can confirm that the system capacity significantly increases by introducing the immediate release of radio resources. This is because the holding time of the allocated radio resource decreases.

Furthermore, by comparing Figs. 3.7(a) and 3.7(b), we observe that the effect of the immediate release of radio resources is large when T_c is small. In detail, when T_c is 10 *s* (Fig. 3.7(a)), the system capacity increases by 11.9 times by introducing the immediate release of radio resources, while 20.7 times for 1 *s* of T_c (Fig. 3.7(b)). This is because the

released radio resources are reused by other UEs in shorter time when the concentration of transmission requests is higher with smaller T_c.

3.6 Applicability to 5G networks

The mathematical analysis in this chapter is based on RAN and EPC architecture for LTE networks. However, the authors consider that the analysis can be applied to RAN and EPC in 5G networks.

5G networks have almost the same mechanism for random access procedure and radio resource allocation [48], while the configuration parameters may be different. For example, some of the preambles may be reserved for Ultra-Reliable and Low Latency Communications (uRLLC) [49,50]. Therefore, the analysis for RAN in Subsection 3.4.1 can be directly applied by modifying parameters in Table 3.2 and the state transition diagram in Figure 3.3.

For core networks, 5G networks would employ the different network configuration based on service-based architecture, the native C/U plane separation, and network slicing technologies [51]. However, the fundamental architecture remain unchanged, meaning that the core network has multiple nodes, and signaling messages are processed at the nodes and exchanged each other. Therefore, the analysis for EPC in Subsection 3.4.2 can be applied to 5G core networks by modifying the signaling flow depicted in Figure 3.2.

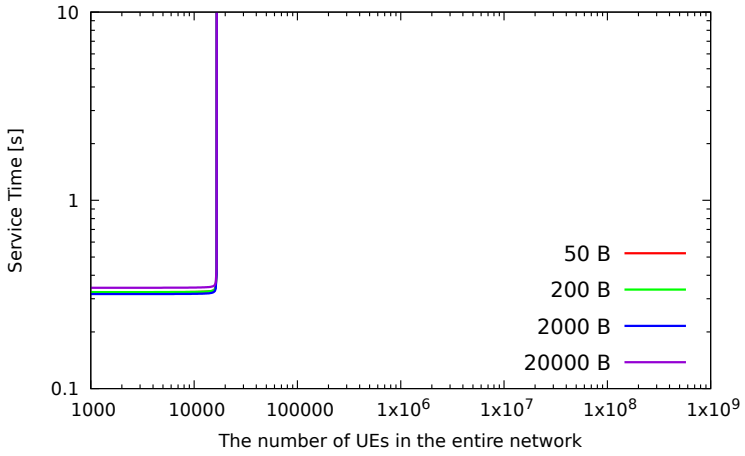
3.7 Conclusion

In this chapter, we presented the performance analysis results of the performance of mobile cellular networks for accommodating C-IoT communications. We conducted the mathematical analysis for deriving the upper limit on number of accommodated UEs in the network, and the time for completing the data transmission for each UE. We also assessed the effect of the immediate release of radio resources, and comparison between LTE and NB-IoT. Through numerical examples of the mathematical analysis, we confirmed that NB-IoT can accommodate larger number of UEs while gives larger service time, when the amount of resources of EPC nodes is enough large. We also presented that the immediate release of

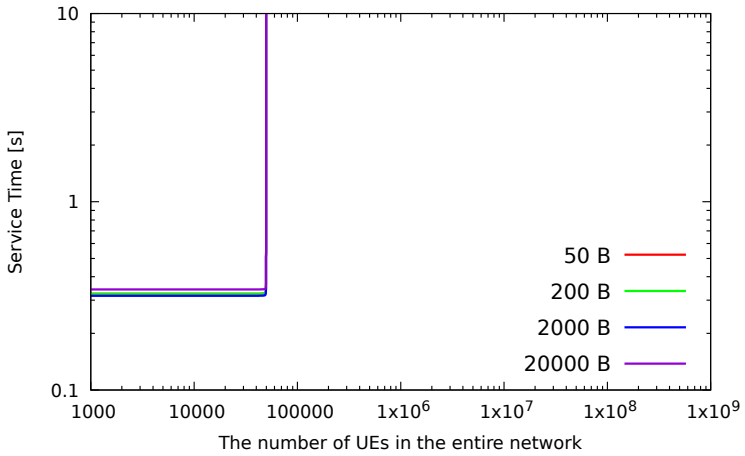
radio resources has a significant impact on the system capacity, especially when there are many and concentrated transmission requests on the network.

The result of this chapter can be applied to the optimization of RAN subslices in User Oriented Network slicing Architecture proposed in Chapter 4.

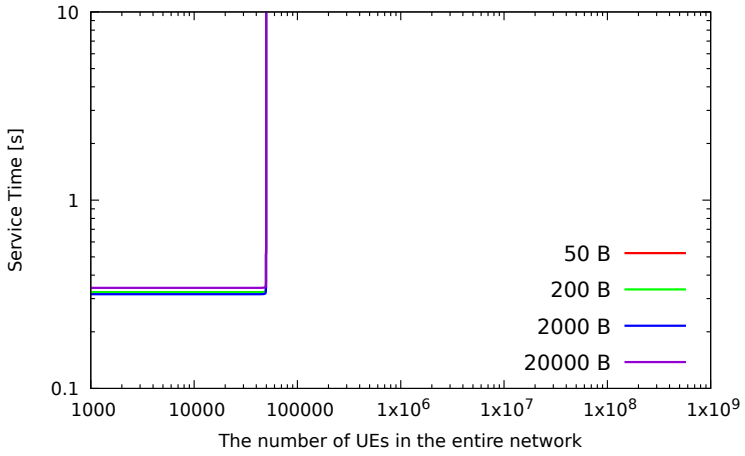
In future work, we plan to evaluate the methods that reduce the overhead of bearer establishment procedure as in [46] for effectively accommodating C-IoT communications to mobile cellular networks. Additionally, we plan for experimental evaluation to support the results mentioned in this chapter.



(a) 1.0×10^7 *statements/s* of server resource

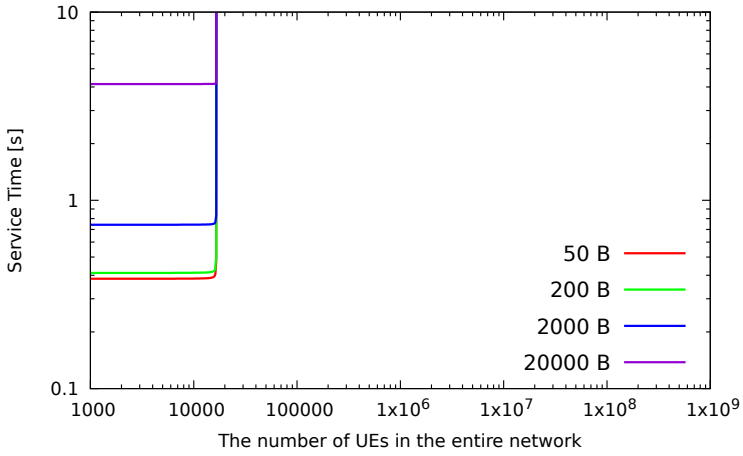


(b) 8.0×10^7 *statements/s* of server resource

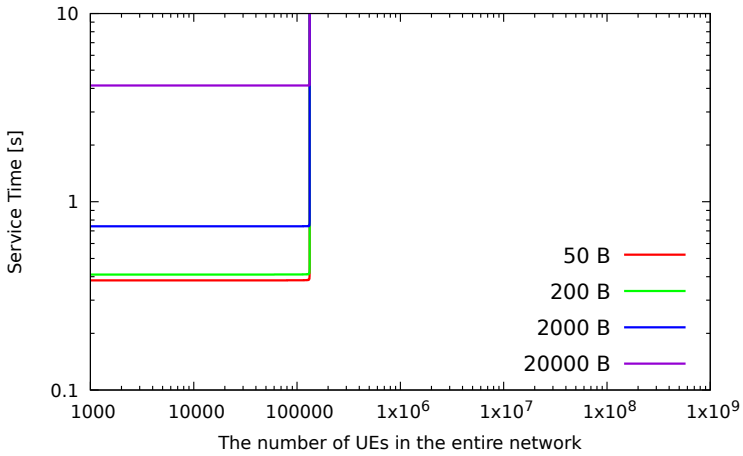


(c) 3.2×10^8 *statements/s* of server resource

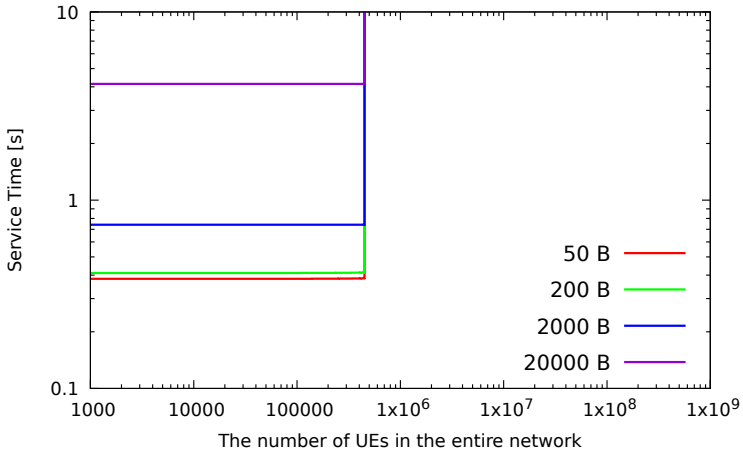
Figure 3.4: Evaluation results for LTE



(a) 1.0×10^7 *statements/s* of server resource



(b) 8.0×10^7 *statements/s* of server resource



(c) 3.2×10^8 *statements/s* of server resource

Figure 3.5: Evaluation results for NB-IoT

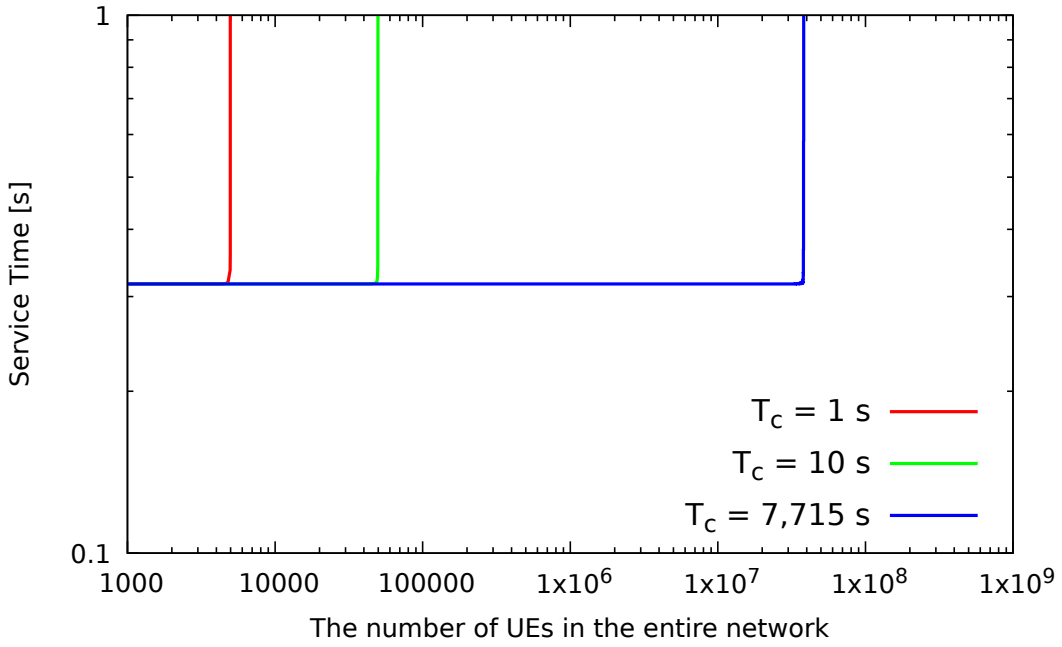
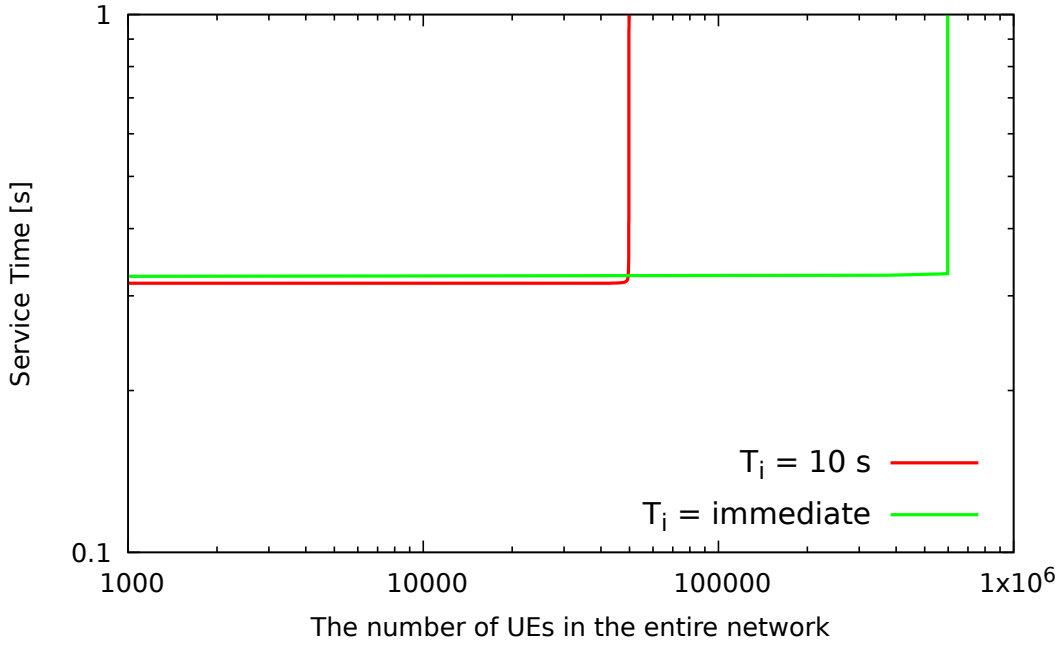
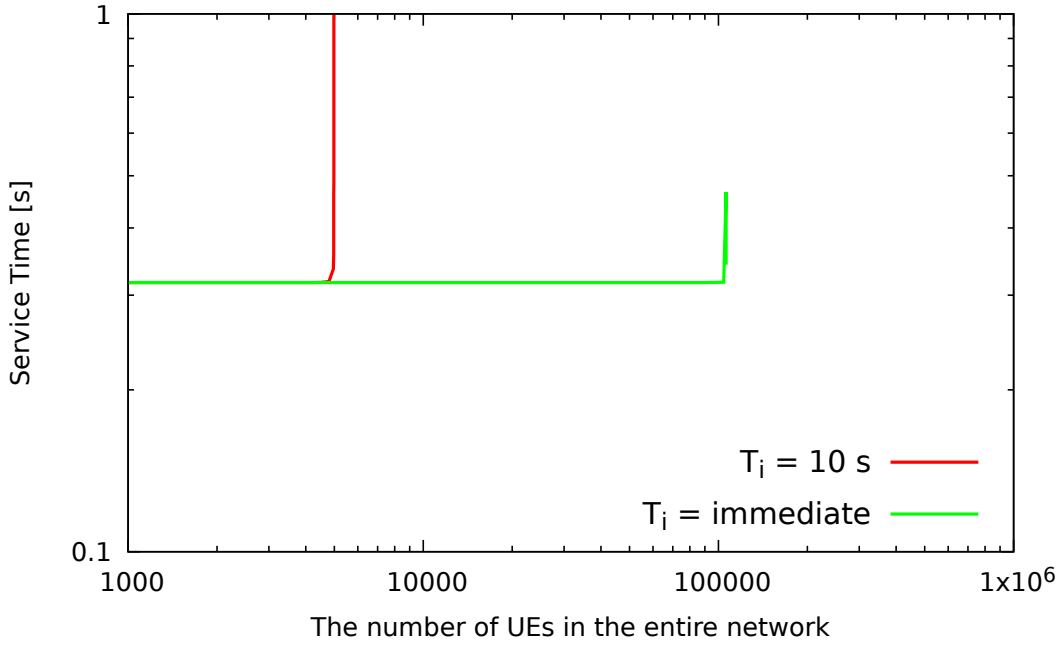


Figure 3.6: Effect of data transmission cycle



(a) 10 *s* for data transmission cycle



(b) 1 *s* for data transmission cycle

Figure 3.7: Effect of immediate release of radio resources

Chapter 4

User-oriented network slicing architecture for beyond-5G networks

4.1 Introduction

In 5G networks, the variety of network services such as enhanced Mobile Broadband (eMBB), massive Machine Type Communications (mMTC), and Ultra Reliable and Low Latency Communications (URLLC) would be supported. One of the key technologies to achieve such objectives is *network slicing* [52], where a virtualized network is generated from physical network and computing resources for accommodating application services.

In future beyond 5G networks, the cell size for higher datarate services with higher radio frequency would be significantly small [53]. Further, the heterogeneous networks [54], which is the mixture of macro/micro/pico/femto cells and other types of wireless networks, such as wireless LANs and Low Power Wide Area (LPWA) networks [55], would be typical environment for last one-hop networks. On the other hand, user applications on the wireless networks will also be diversified and personalized, including X-reality and digital twins. Therefore, the requirements for network quality would be on a per-user basis.

Furthermore, the resource demand for each network slice would change spatio-temporally, because of the various factors such as network load changes, user mobility, and unexpected utilization of application services. The personalization of user application services makes such changes more largely. However, most of existing researches on network slicing for

5G network explicitly or implicitly assume the service-based slice maintenance, meaning that multiple users (user equipments) using a certain service are accommodated to a single network slice generated for the service. Also, they do not consider real-time and on-demand reconfiguration of network slices. Therefore, we believe that the per-service network slicing in 5G network cannot accommodate the above-mentioned requirements in future beyond 5G networks.

In this chapter, we propose a novel network architecture, User-Oriented Network slicing Architecture (UONA). The proposed architecture has the following two major characteristics. One is that we maintain network slices on a per-user basis, not a per-service basis. Such high-resolution of network slices would satisfy various requirements from users. The other is decoupling the process of generating network slices into two subprocess, providing subslices by subslice providers and constructing end-to-end user network slices by network slice brokers. To the best of our knowledge, this is the first proposal of personalization of network slices with spatio-temporal, real-time, and on-demand reconfigurations. This chapter describes the fundamental design of UONA with research challenges to realize it for future beyond 5G networks. Furthermore, we present some evaluation results to demonstrate the effectiveness of per-user configuration of network slice, under the simple scenario of constructing a virtual cell for each user to decrease the overhead of handover procedure.

The rest of this chapter is organized as follows. In Section 4.2 we give the detailed description of UONA. Section 4.3 summarizes the research challenges for realizing UONA. We then give some evaluation results to preliminarily confirm the effectiveness of creating network slices on a per-user basis. The related work is presented in Section 4.5. Finally, in Section 4.6, we give the conclusion of our chapter and future work.

4.2 User-oriented Network slicing Architecture (UONA)

4.2.1 Motivation

One of the possible ways to realize network slicing on a per-user basis is to simply extending the existing per-service network slicing in 5G networks to per-user slicing. However, such

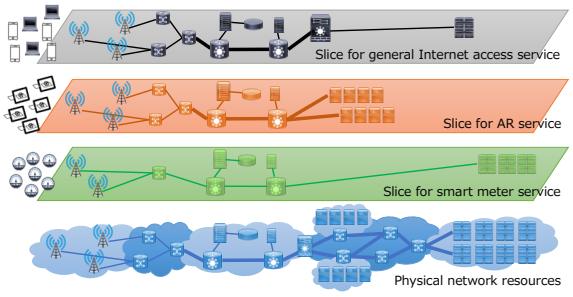


Figure 4.1: Service-based network slicing in 5G network

simple extension would increase the slicing resolution and scalability issues on the number of accommodated users arise. Resource isolation among such many slices becomes also a severe problem [56].

In general, appropriate spatio-temporal resolution of network slicing would be dependent on various factors such as the required quality and amount of resources, user mobility characteristics, application usage patterns. We believe that appropriate slice resolution would be different at the different part of the end-to-end user network. For example, for the applications which require the fixed network bandwidth, per-user slicing is necessary to guarantee the required bandwidth. On the other hand, the traditional mobile packet core networks should be shared among multiple users to decrease the overhead. Network slices for best effort services should be shared among users to enjoy the statistical multiplexing effect [57]. Appropriate resolution of RAN slicing for handover procedure would be dependent on the mobility characteristics of user terminals [58].

Furthermore, large processing overhead cannot be avoided for reconfiguring the entire end-to-end slice according to the temporal demand changes. Distributed reconfiguration of each network component would be a possible solution. However, per-service network slicing in 5G network cannot accommodate such requirement.

4.2.2 User-based network slicing

Figure 4.1 depicts the concept of service-based network slicing in 5G networks. Each network slice is maintained to accommodate a specialized service such as smart meter

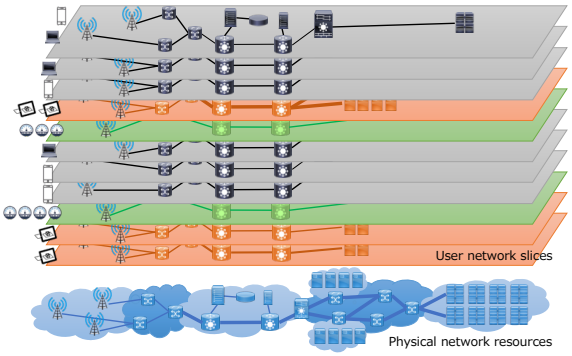


Figure 4.2: Service-based network slicing in 5G network

service, AR service, and so on. Users (terminals) for each service are accommodated to the corresponding slice. Since multiple users are accommodated to the same network slice, it is difficult to handle requirements and characteristics of each user.

On the other hand, in UONA, to provide appropriate network slices to satisfy users' diversified, personal, and spatio-temporally changing demands, network slices are generated on a per-user basis, not a per-service basis. Figure 4.2 depicts the concept of UONA, where network slices are constructed in per-user(s) basis. In general, each network slice in UONA reqires smaller amount of network and server resources compared with service-based slices in 5G network, since each slice accommodate smaller number of users. On the other hand, since each network slice is allocated to each user, it can be easily configured to satisfy diversifying users' demands.

Figure 4.3 explains the fundamental factors and their relationships in UONA. In the figure, a *subslice* is defined as a virtualized resource constructed from a part of end-to-end physical network resources. A *user network slice* is an end-to-end slice composed of subslices, to be provided to each user. In Figure 4.3, we assume that the end-to-end user network is divided into RAN, packet core, edge cloud, and core cloud, while the division of the network can be conducted arbitrarily. Also, Figure 4.3 explains only the data plane functions, but the same architecture can be constructed for control and management plane functions.

Also, to improve the variety of network slices, we decouple the process of creating end-to-end network slices into two subprocesses. One is generating subslices from physical

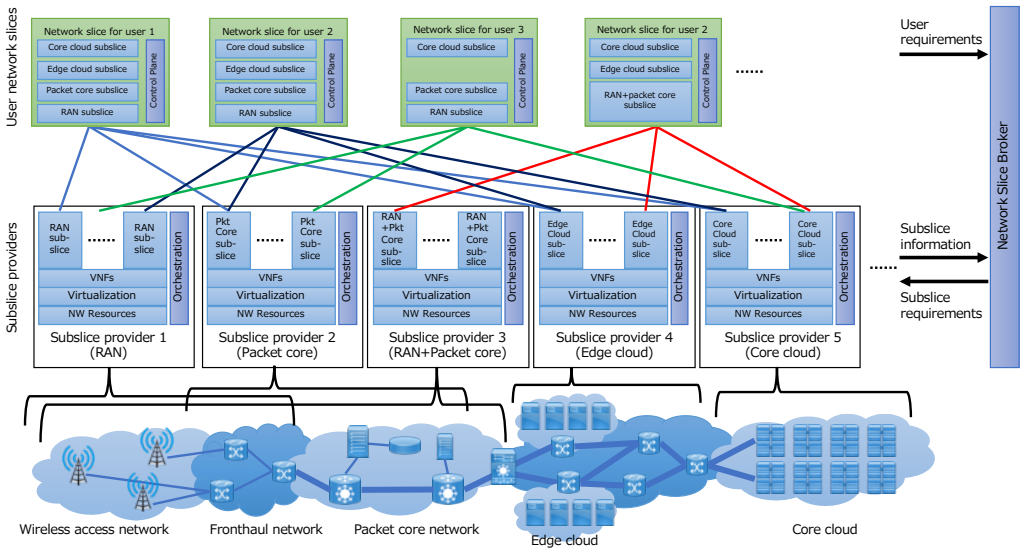


Figure 4.3: UONA architecture

network resources by *subslice providers*. The other is creating user network slices by *network slice brokers* from subslices according to users' requirements. We describe the detailed architecture in the following subsections.

4.2.3 Per-user network slice

In UONA, as shown in Figure 4.3, the network slice is generated and maintained on a per-user basis so that each user has its own, personalized network slice. However, the subslices composing the user network slice may be shared with multiple user network slices. For example, in Figure 4.3, the network slice for user 1 has RAN subslice exclusively for user 1, but the packet core, edge cloud, and core cloud subslices are shared with the network slice for user 2. The network slice for user 3 does not have a edge cloud subslice since user 3 does not require the resources at the edge cloud. The network slice for user 4 has a subslice that is composed from physical resources of RAN and packet core network.

Note that this per-user slicing can co-exist with per-service network slicing in 5G networks, where per-user slices are provided only for users requiring their own, personalized networks.

4.2.4 Subslice provider

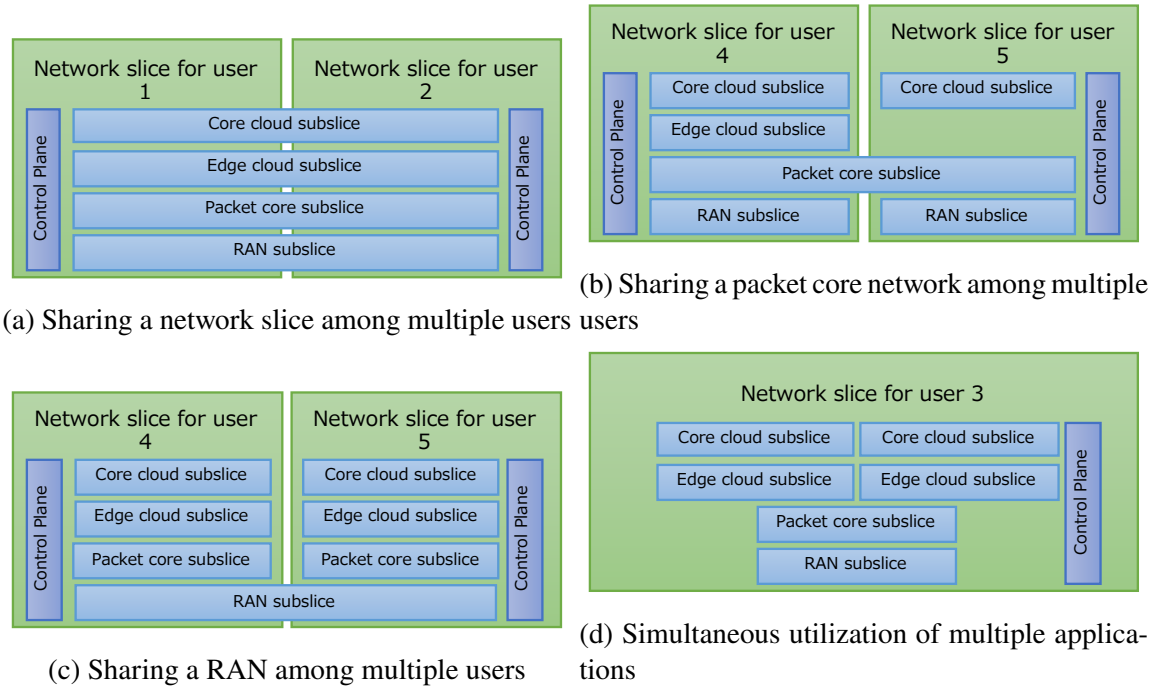
Subslice provider has physical netwrok resources to generate subslices through virtualization technologies. In Figure 4.3, Subslice providers 1, 2, 3, 4, and 5 has physical resources of RAN, packet core, both of RAN and packet core, edge cloud, and core cloud, respectively, and provide the corresponding subslices. Some subslice provider may has a physical resource of multiple network components, as Subslice provider 3, which may employ xHaul architecture [59] to effectively utilize the phisical resources. It is assumed that each subslice provider has its ow orchestration mechanism to create subslices.

The subslice providers would generate subslices according to the request from network slice broker. On the other hand, they should reconfigure their own subslices according to various factors such as the resource utilization and the mobility characteristics of users accommodated by their own subslices. Since the quality of subslices would determine their (monetary) values to be selected by slice brokers, the competitive environment among slice brokers can be created. A new subslice provider can join the UONA by providing subslices for a new network component, or providing subslices for a new combination of existing network components.

4.2.5 Network slice broker

Network slice broker constructs end-to-end network slices for users based on the requirements from the users and the availability of subslices from subslice providers. It interprets users' requirements into performance requirements of subslices and select subslices of each network component that satisfy the requirements. The slice broker can request subslice providers to provide new subslices when it cannot find appropriate subslices.

When creating user network slice, subslices can be shared by multiple user network slices, and multiple subslice can be assigned to a user network slice. In Figure 4.4 we depict the examples of subslice sharing. In Figures 4.4(a), two user network slices share the all subslices. This means that users 1 and 2 are accommodated to the identical network slice. In Figures 4.4(b) and 4.4(c), some subslices are shared by multiple user network slice. In Figure 4.4(d), one network slice has multiple subslices for a single network component to simultaneously use multiple applications hosted at different cloud environment. This



(a) Sharing a network slice among multiple users

(b) Sharing a packet core network among multiple users

(c) Sharing a RAN among multiple users

(d) Simultaneous utilization of multiple applications

Figure 4.4: Subslice sharing among user network slices

flexibility in subslice utilization can relax the scalability limitations on per-user network slicing in UONA.

4.2.6 Continuous reconfigurations of user network slices

One important function in UONA is continuous reconfigurations of user network slices based on various factors such as the network and application load changes and user mobility. The concept of subslices helps the reconfiguration to be easily realized. That is, we do not need reconfigure the whole network slice, but the partial reconfiguration in unit of subslice can be conducted.

Typical example is depicted in Figure 4.5. In the beginning (Figure 4.5(a)), a user network slice is constructed by four subslices. A RAN subslice includes one base station for connecting a UE. Edge and core cloud subslices have three and one servers, respectively. These cloud subslices and the RAN subslice are interconnected by a packet core subslice. In Figure 4.5(b), the UE moves and the RAN subslice is reconfigured so that to include

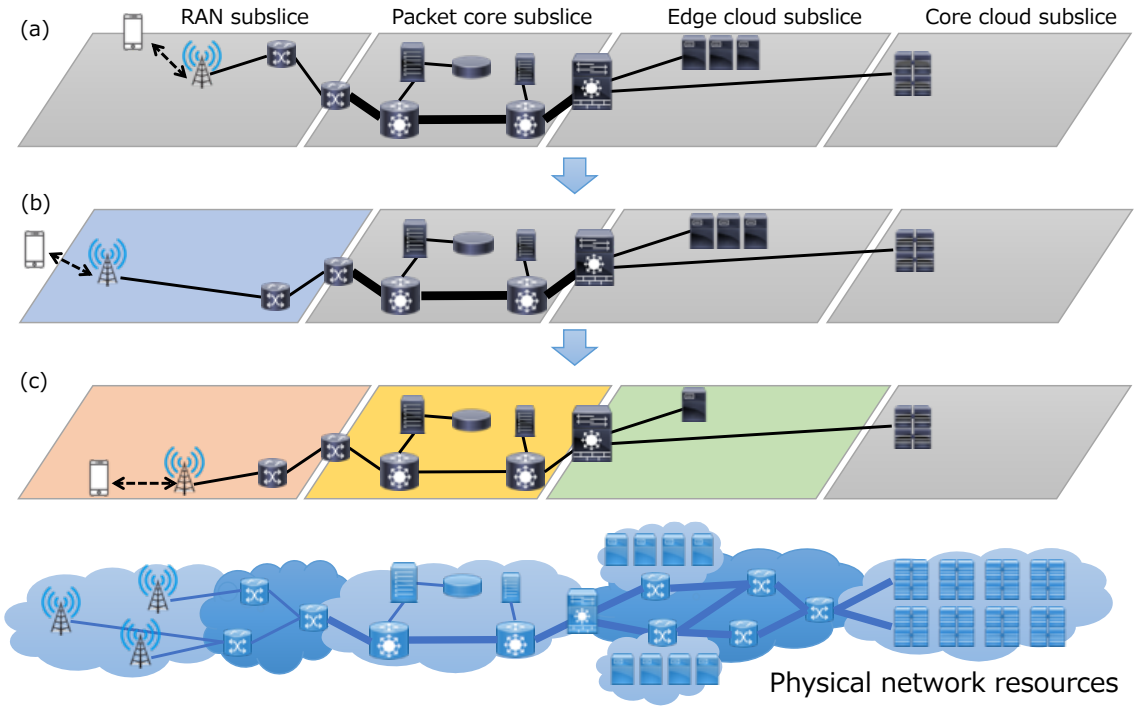


Figure 4.5: Continuous slice reconfiguration

the base station near the UE. In Figure 4.5(c) the UE again moves and the RAN subslice is reconfigured. Furthermore, since the network traffic between the UE and edge cloud servers decreases, the edge cloud subslice is reconfigured by decreasing the number of servers. The packet core subslice is also reconfigured so that the data-plane network bandwidth is reduced.

Two ways can be considered as user network slice reconfiguration. One is *subslice adjustment*, that reconfigures the currently-used subslice by the corresponding subslice provider. In this case, the composition of the user network slice, meaning that which subslices are used for composing the user network slice, remain unchanged. The other is *subslice replacement*, where a new subslice is prepared by the subslice provider, and the network slice broker switches the subslice from an old one to the new one. The latter method is similar to the slice selection method in [60], but in UONA only a part of the end-to-end slice is reselected.

We assume that subslice adjustment is better to be used for small changes in subslices, such as parameter tuning. Even when the subslice is shared by multiple user network slices, the reconfiguration of the subslice can be conducted easily, compared with each user network slice has individual subslice.

On the other hand, subslice replacement is necessary to reconfigure the subslice largely, such as the vertical handover among different types of wireless networks. In subslice replacement, the continuity of the network service is important issue for avoiding the performance degradation on the replacement. One possible way is to apply the Session and Service Continuity (SSC) technology considered in [61].

4.2.7 Network slicing ecosystem

The concept of subslice provider and slice broker would be a part of a new ecosystem of network slicing. The subslice providers should provide subslices for network component composing an end-to-end user network slice. They do not need to consider the end-to-end performance. This would bring the variety and specialized subslices for each network component. A new subslice provider can join UONA easily by prodiving subslices for a part of the network.

On the other hand, the network slice brokers concentrate the combination of subslices to make a user network slice for each user according to the user's requirements. Various brokers would have different schemes to construct user network slices to motivate users to exploit the slice brokers.

4.3 Research challenges

In this section we summarize the research challenges for realizing UONA.

4.3.1 Multi-resolution subslicing

For maintaining per-user network slice, some subslices can be created on a per-user basis as in Figure 4.2. Such finer resolution subslicing require more sophisticated virtualization technologies. To the best of our knowledge, for computing resources, per CPU-core

virtualization is a promising method for effective resource isolation. However, for per-user subslicing, sharing one CPU-core by multiple user subslices is required. Therefore, resource slicing and isolation technologies in such fine resolution should be realized for UONA.

4.3.2 Signaling protocol

The realizing the ecosystem of network slicing described in Subsection 4.2.7, the signaling protocol should be established and standardized among multiple stake holders, corresponding to network slice brokers, subslice providers, and end users in UONA. Note that subslice providers can utilze any framework and protocol for creating subslices, because that is hidden from the network slice brokers.

4.3.3 Optimizing user network slices

When constructing user network slices, network slice brokers should take various factors into account, such as spatio-temporal availability of subslices and user requirements. In the orchestration and resource allocation of NFV environment, the problems are often modelled as optimization problems and solve them by introducing heuristic algorithm or by using solver software platform [62]. The similar problem modeling can be applied to our problem, while the computational overhead would become a important issue for continuous reconfiguration of user network slices.

4.4 Preliminary evaluation

In this section, we present some numerical evaluation results to demonstrate the effectiveness of per-user optimization of subslice configuration and realtime reconfiguration based on user mobility characteristics. For that purpose, we borrow the *virtual cell (VC)* construction problem discussed in [63]. The authors in [63] considered that a VC is formed for each UE from multiple physical cells, and the context information of the UE is shared by the physical access points in the VC. The control-plane signaling overhead on handover between physical cells in the VC becomes small as compared with the normal handover.

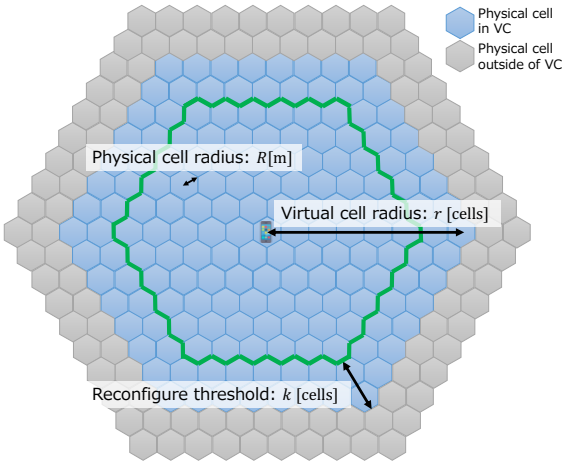


Figure 4.6: Evaluation model

Furthermore, the VC is continuously reconfigured based on the estimation of the moving direction and velocity of the UE.

4.4.1 System model

Based on [63], we model the system as depicted in Figure 4.6. We assume hexagonal cells and a virtual cell is also formed by physical cells in a hexagonal area whose center is a physical cell at which a UE is located. In Figure 4.6, the blue area is the virtual cell of the UE at the center. When the UE reaches near the edge of the virtual cell (green line in Figure 4.6), the virtual cell is reconfigured so that the UE becomes the center of the virtual cell.

We consider the signaling overhead for handover procedure and reconfigure of the virtual cell. We call the handover between physical cells in the VC as *soft handover*, and the handover from the physical cell in the VC to the physical cell out of the VC (gray cells in Figure 4.6) as *hard handover*. Because the context information of the UE is shared among the physical cells in the VC, the signaling overhead of soft handover is smaller than that of hard handover. On the other hand, when reconfiguring the VC, the context information of the UE should be removed from the physical cells leaving from the VC, and it should be added to the newly joining physical cells. Therefore, by moving the VC with the mobility

of the UE, the handover overhead remains small, while the overhead for reconfiguring the VC cannot be ignored.

4.4.2 Performance analysis

The radius of the hexagonal physical cell is denoted as R. We also denote the radius of the VC, in unit of the number of physical cells, as r. Note that r=0 when the VC is formed by only one physical cell. Then, the number of physical cells in the VC, denoted as s, is $3r^2 + 3r + 1$ [64]. When the average velocity of a UE is $E[V]$, the rate at which the UE leaving the physical cell, denoted by μ_{p}, is calculated as follows.

$$\mu_{\mathrm{p}} = \frac{4}{\sqrt{3}} \frac{E[V]}{R}$$

Also, the rate at which the UE leaving the virtual cell whose radius is r, denoted by $\mu_{\mathrm{v}}(r)$ is as follows.

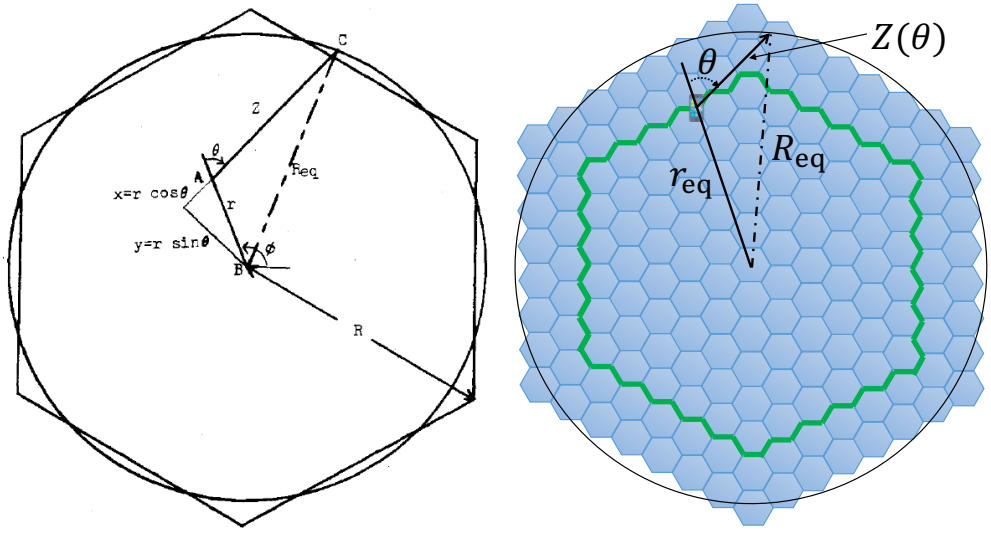
$$\mu_{\mathrm{v}}(r) = \frac{2r+1}{3r^2+6r+3} \mu_{\mathrm{p}}$$

Note that these equations are based on the fluid flow model in [64].

When the reconfiguration threshold is k, as depicted in Figure 4.6, meaning that the reconfiguration of the VC is activated when the UE reaches the physical cell located at k cells from the edge of the VC, the rate at which the reconfiguration of the VC is activated becomes $\mu_{\mathrm{rs}}(r,k) = \mu_{\mathrm{v}}(r-k)$.

The overhead of handover procedures and VC reconfiguration is evaluated by the number of signaling messages exchanged. The numbers of signaling messages for soft and hard handovers are denoted as m_{s} and m_{h}, respectively. The number of signaling messages for VC reconfiguration is dependent on the radius of VC, r, and the reconfiguration threshold, k. Therefore, it is denoted by $m_{\mathrm{rs}}(r,k)$.

Even when the VC is reconfigured based on the mobility of the UE, hard handover occurs when the UE reaches the border of the VC before completing the reconfiguration of the VC. We assume that in such a case, after finishing the hard handover procedure, the configuration of a new VC is reactivated because the previous reconfiguration fails.



(a) The model in [1] (b) Our model

Figure 4.7: Interpretation of the analysis in [1]

We denote the occurrence rate of such reconfiguration failure by μ_{f}, and the number of signaling messages to configure a new VC by $m_{\mathrm{rh}}(s)$.

For deriving μ_{f}, we exploit the analysis model in [1], where the authors derived the average time duration from when a call to a UE is originated in a certain cell to when the UE reaches the border of the cell. We interpret a cell in [1] into a virtual cell in our analysis, and the location at which a call is originated is considered as the location where the reconfiguration of VC is activated. The interpretation of the analysis model is depicted in Figure 4.7.

In Figure 4.7(b), $Z(\theta)$ is the moving distance of the UE until it reaches the border of the virtual cell, when the moving direction is represented by θ. This can be calculated as follows.

$$Z(\theta) \;=\; \sqrt{R_{\mathrm{eq}}^2 - (r_{\mathrm{eq}} \sin\theta)^2} - r_{\mathrm{eq}} \cos\theta \;\; \left(-\frac{\pi}{2} \le \theta \le \frac{\pi}{2}\right)$$

where $R_{\mathrm{eq}} = 3\sqrt{\frac{3\sqrt{3}}{2\pi}} rR$ is the radius of the circle approximating of the hexagonal area of the VC, $r_{\mathrm{eq}} = 3\sqrt{\frac{3\sqrt{3}}{2\pi}}(r-k)R$ is the distance from the center of the VC to the location

at which the reconfiguration of VC is activated, and θ means the moving direction of the UE with respect to a vector from the center of the cell to where the VC reconfiguration is activated. See Figure 4.7(b) for details.

We denote the cumulative distribution function of $Z(\theta)$ as $F(Z)$ $\left(R_{\mathrm{eq}} - r_{\mathrm{eq}} \leq Z \leq \sqrt{R_{\mathrm{eq}}^2 - r_{\mathrm{eq}}^2}, 0 \leq F(Z) \leq 1\right)$. By representing the time required for reconfiguration as T_{r}, μ_{f} can be obtained as follows.

$$\mu_{\mathrm{f}} = \begin{cases} 0 & vT_{\mathrm{r}} \leq R_{\mathrm{eq}} - r_{\mathrm{eq}} \\ 1 - F(vT_{\mathrm{r}}) & R_{\mathrm{eq}} - r_{\mathrm{eq}} \leq vT_{\mathrm{r}} \leq \sqrt{R_{\mathrm{eq}}^2 - r_{\mathrm{eq}}^2} \\ 1 & \sqrt{R_{\mathrm{eq}}^2 - r_{\mathrm{eq}}^2} \leq vT_{\mathrm{r}} \end{cases}$$

Note that we assume that the UE moves linearly with the velocity of v, as in [1]. For detailed analysis method, please refer the Appendix in [1].

From the above discussion, the average signaling overhead can be obtained as follows.

$$O_{\mathrm{UONA}} = \mu_{\mathrm{p}} m_{\mathrm{s}} + \mu_{\mathrm{rs}}(r,k) m_{\mathrm{rs}}(r,k) + \mu_{\mathrm{f}}(m_{\mathrm{h}} + m_{\mathrm{rh}}(s)) \tag{4.1}$$

For user-perceived performance, the frequency at which hard handovers occur is also evaluated, since hard handovers take more time than soft handovers. The hard handover frequency of our method, H_{UONA}, can be obtained as follows.

$$H_{\mathrm{UONA}} = \mu_{\mathrm{rs}}(r,k)\mu_{\mathrm{f}} \tag{4.2}$$

4.4.3 Other methods for performance comparison

RCBD: Reconfiguration on border of virtual cell

We can consider a simpler reconfiguration method, where a VC is reconfigured only when a UE reaches the border of the VC. Note that this method corresponds to our UONA method with k = 0, meaning that the green line in Figure 4.6 is located on the border of the virtual cell. The signaling overhead and hard handover frequency of this method is obtained as

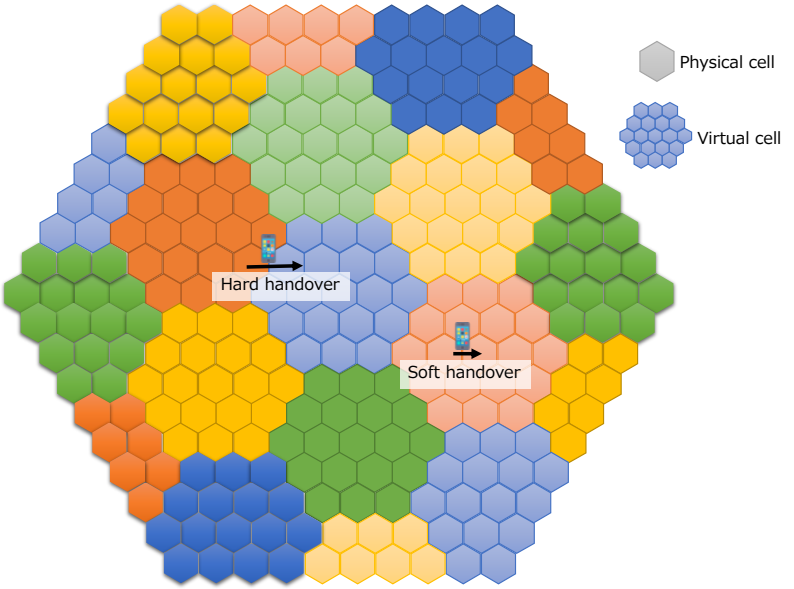


Figure 4.8: Handover based on 5G service-based slicing and slice selection

follows.

$$O_{\mathrm{RCBD}} = \mu_{\mathrm{p}} m_{\mathrm{s}} + \mu_{\mathrm{v}}(r)(m_{\mathrm{h}} + m_{\mathrm{rh}}(s)) \tag{4.3}$$

$$H_{\mathrm{RCBD}} = \mu_{\mathrm{v}}(r) \tag{4.4}$$

Note that this equation can be obtained by substituting k = 0 and μ_{f} = 1 in Equation (4.1).

5gSS: Slice selection with 5G's service-based network slicing

We further consider a method based on the concepts of service-based slice construction and slice selection [60]. In this case, as depicted in Figure 4.8 the virtual cell is fixedly constructed to cover all cells in the area. Each virtual cell corresponds to a network slice that is constructed in advance. Handovers across physical cells in a virtual cell are treated as soft handover, and handovers across physical cells belonging to different virtual cells are treated as hard handover. Therefore, the average signaling overhead and hard handover

frequency in this case can be derived as follows.

$$O_{\mathrm{5gSS}} = \mu_{\mathrm{p}} m_{\mathrm{s}} + \mu_{\mathrm{v}}(r) m_{\mathrm{h}} \tag{4.5}$$

$$H_{\mathrm{5gSS}} = \mu_{\mathrm{v}}(r) \tag{4.6}$$

NoVC: No use of virtual cells

Finally, when we do not utilize the virtual cell, hard handover occurs at every movement of a UE between physical cells. The average signaling overhead and hard handover frequency in this case is calculated as follows.

$$O_{\mathrm{NoVC}} = \mu_{\mathrm{p}} m_{\mathrm{h}} \tag{4.7}$$

$$H_{\mathrm{NoVC}} = \mu_{\mathrm{p}} \tag{4.8}$$

By numerically comparing Equation (4.1) with Equations (4.3), (4.5), and (4.7), the effect of per-user configuration and continuous reconfiguration of virtual cell is evaluated. We expect that by appropriately configuring r and k according to the user mobility characteristics, O_{UONA} in Equation (4.1) becomes smaller compared with other methods. Smaller value of O_{UONA} than O_{NoVC} means that per-user optimization is effective. When O_{UONA} is smaller than O_{RCBD}, we can say that the continuous reconfiguration of virtual cell with the user mobility is meaningful. Furthermore, by comparing O_{UONA} with O_{5gSS}, we can assess the effectiveness of the user-based network slicing in UONA against the service-based network slicing in 5G networks.

For user-perceived performance, we compare the hard handover frequency of UONA, H_{UONA}, with other methods (H_{RCBD}, H_{5gSS}, and H_{NoVC}), by using Equations (4.2), (4.4), (4.6), and (4.8).

4.4.4 Numerical evaluation results and discussions

Parameter settings

The numbers of signaling messages for soft and hard handovers, m_{s} and m_{h}, are set to 8 and 15, respectively, based on the signaling flow presented in [65]. The number of signaling

Table 4.1: Parameter settings

Parameter	Description	Values
R [m]	Physical cell radius	5, 50, 100
v [km/h]	Velocity of UE	4, 40, 100, 400
T_r [sec]	Time for VC reconfiguration	0.001, 0.01, 1

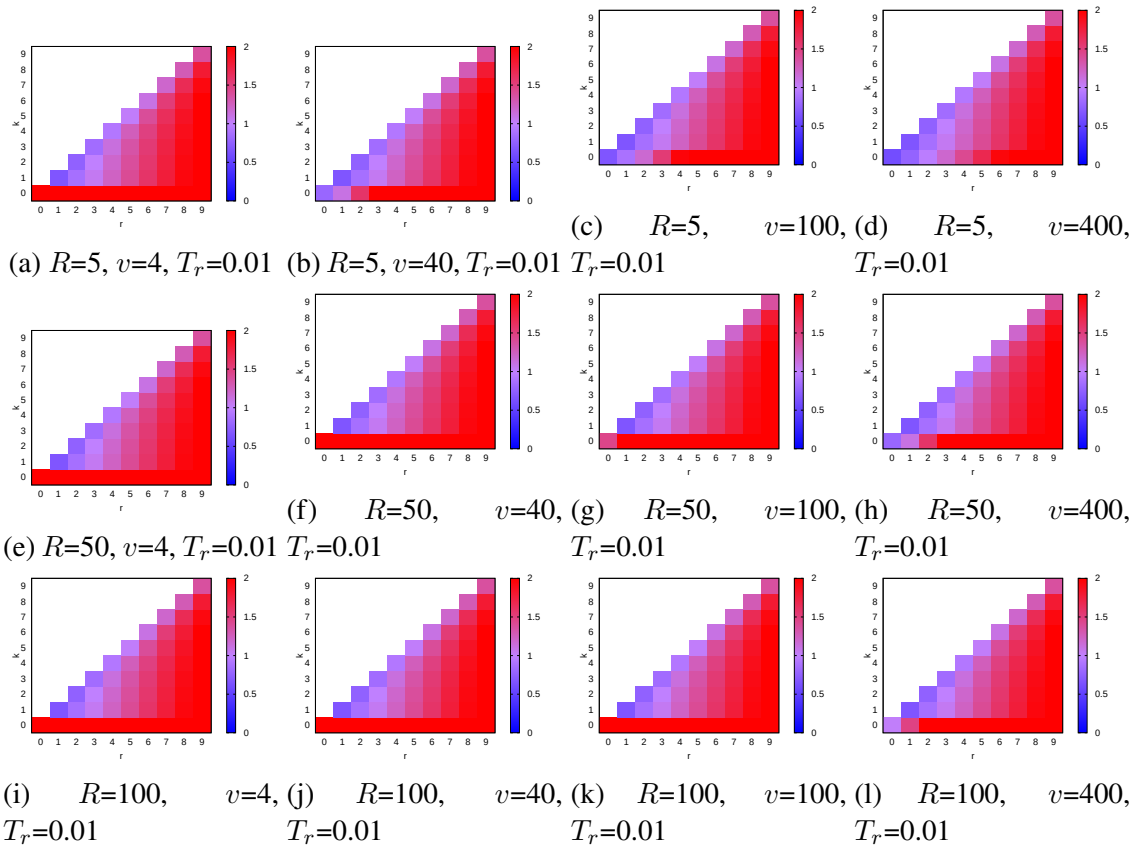
messages for configuration of a new VC, $m_{\mathrm{rh}}(s)$, is set to $2 \cdot s$, assuming that the context information of a UE is removed from physical cells in the old VC, and it is sent to all physical cells in the VC. The number of signaling messages for reconfiguration of VC, $m_{\mathrm{rs}}(r, k)$, is set to $2((r+1)-k)(2(r+1)-1)$, assuming that the context information of a UE is removed from physical cells leaving the VC, and it is sent to physical cells newly joining the VC.

For the environmental parameters and the velocity of the UE is summarized in Table 4.1. In total, the number of parameter combinations is $3 \times 4 \times 3 = 36$. For each combination, we assess the settings of virtual cell radius, r, and reconfiguration threshold, k, where $0 \leq r, k \leq 9$ $(k \leq r)$.

Signaling overhead

Figure 4.9 plots $(O_{\mathrm{UONA}}/O_{\mathrm{NoVC}})$, that is, the ratio of the signaling overhead of UONA to that of NoVC, to assess the effect of introducing virtual cells and its continuous reconfiguration. In the graphs, x-axis and y-axis mean the settings of r and k, respectively. Blue and red regions mean $(O_{\mathrm{UONA}}/O_{\mathrm{NoVC}}) < 1$ and $(O_{\mathrm{UONA}}/O_{\mathrm{NoVC}}) > 1$ with given combination of (r, k), respectively. White regions represent the impossible combinations of (r, k), that is, $r < k$. In the figure we change R and v, while T_r is fixed to 0.01 [sec]. Note that the results with the different values of T_r are omitted due to space limitation, but the obtained results are similar to Figure 4.9.

From this figure, we can observe that blue region becomes wider when R is smaller, or when v is larger. This is because smaller R and larger v means that a UE moves faster on smaller cells, that brings more frequent handovers. In that situations, the effect of UONA is large, meaning that introducing the virtual cell and its per-user reconfiguration can significantly decrease hard handover freqhency with large signaling overhead.



(a) R=5, v=4, T_r=0.01 (b) R=5, v=40, T_r=0.01 (c) R=5, v=100, T_r=0.01 (d) R=5, v=400, T_r=0.01

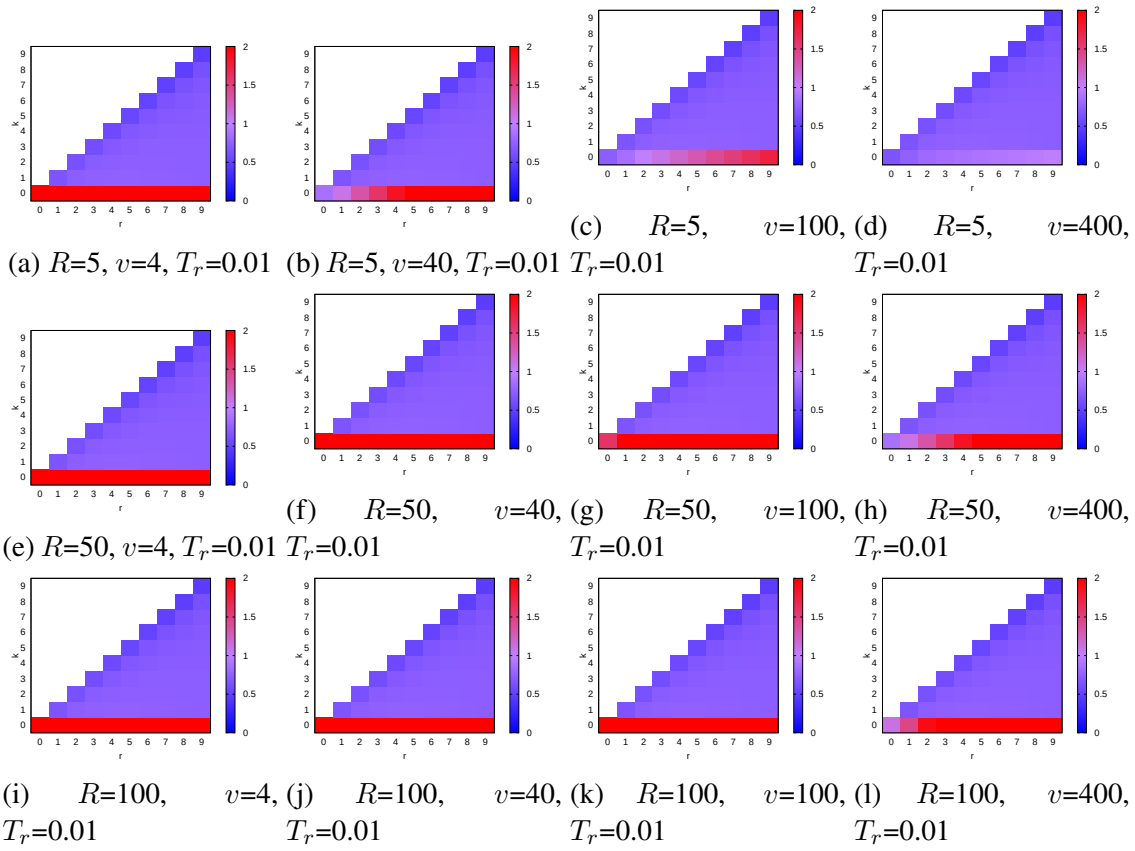
(e) R=50, v=4, T_r=0.01 (f) R=50, v=40, T_r=0.01 (g) R=50, v=100, T_r=0.01 (h) R=50, v=400, T_r=0.01

(i) R=100, v=4, T_r=0.01 (j) R=100, v=40, T_r=0.01 (k) R=100, v=100, T_r=0.01 (l) R=100, v=400, T_r=0.01

Figure 4.9: Effect of per-user optimization of virtual cell configurations $(O_{\mathrm{UONA}}/O_{\mathrm{NoVC}})$

Figure 4.10 plots $(O_{\mathrm{UONA}}/O_{\mathrm{RCBD}})$, that is, the ratio of performance of UONA to that when virtual cell is introduced but its reconfiguration with user mobility is not applied. We can confirm the similar characteristics to Figure 4.9, meaning that smaller R and larger v enlarge the blue regions. Furthermore, the blue region in Figure 4.10 is quite wider than that in Figure 4.9. These results mean that the continuous reconfiguration of virtual cell with user mobility in UONA has large impact on decreasing the signaling overhead.

In Figure 4.11, $(O_{\mathrm{UONA}}/O_{\mathrm{5gSS}})$ is plotted to compare the performance of UONA with service-based slicing of 5G network. Compared with Figures 4.9 and 4.10, UONA outperforms 5gSS in narrower parameter regions of (r, k). However, for all cases we can find the optimal parameters so that the signaling overhead of UONA is smaller than 5gSS. This clearly confirms the effect of per-user slicing in UONA.



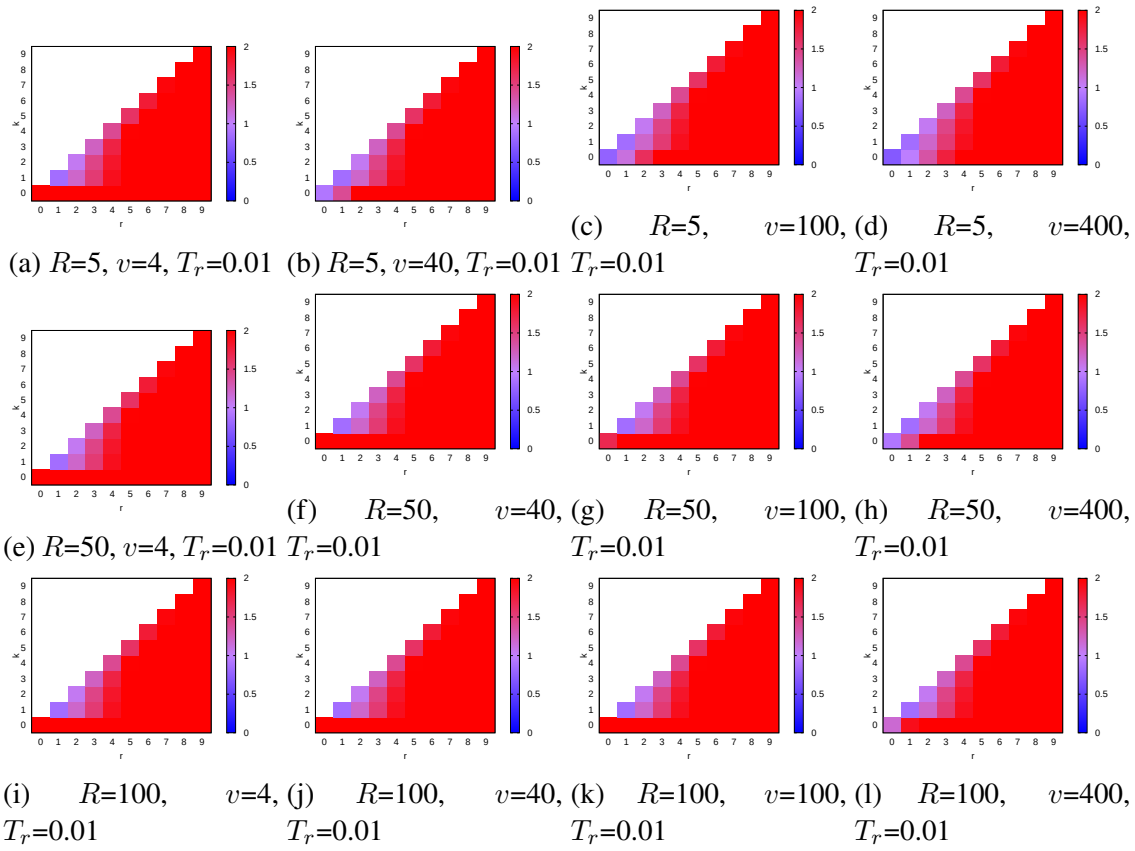
(a) R=5, v=4, T_r=0.01 (b) R=5, v=40, T_r=0.01 (c) R=5, v=100, T_r=0.01 (d) R=5, v=400, T_r=0.01

(e) R=50, v=4, T_r=0.01 (f) R=50, v=40, T_r=0.01 (g) R=50, v=100, T_r=0.01 (h) R=50, v=400, T_r=0.01

(i) R=100, v=4, T_r=0.01 (j) R=100, v=40, T_r=0.01 (k) R=100, v=100, T_r=0.01 (l) R=100, v=400, T_r=0.01

Figure 4.10: Effect of continuous reconfiguration of virtual cell configurations $(O_{\mathrm{UONA}}/O_{\mathrm{RCBD}})$

In Figure 4.12, we plot the average signaling overhead of UONA with other three methods. Note that the results of UONA are values when we choose (r, k) so that the signaling overhead becomes smallest. The x axis means the combination of parameters of R, v, and T_r. Note that y axis is logarithmic. We can see from this figure that UONA has smallest signaling overhead among four methods for all parameter combinations. This corresponds to the fact that for all of 36 parameter combinations, the settings of r and k are found so that O_{UONA} becomes smaller than O_{RCBD}, O_{NoVC}, and O_{5gSS}. In overall, by setting r to be smaller than around five, and k to be near to r, O_{UONA} becomes the smallest among the four methods.



(a) R=5, v=4, T_r=0.01 (b) R=5, v=40, T_r=0.01 (c) R=5, v=100, T_r=0.01 (d) R=5, v=400, T_r=0.01

(e) R=50, v=4, T_r=0.01 (f) R=50, v=40, T_r=0.01 (g) R=50, v=100, T_r=0.01 (h) R=50, v=400, T_r=0.01

(i) R=100, v=4, T_r=0.01 (j) R=100, v=40, T_r=0.01 (k) R=100, v=100, T_r=0.01 (l) R=100, v=400, T_r=0.01

Figure 4.11: Effect of per-user slicing ($O_{\mathrm{UONA}}/O_{\mathrm{5gSS}}$)

Hard handover frequency

In Figure 4.13, we present the evaluation results of hard handover frequency of UONA with other three methods in the same style as Figure 4.12. In overall, when the physical cell radius (R) is large and the velocity of UE (v) is small, hard handover rarely occurs in all methods. This is because the frequency at which UE crosses the border of physical cell due to the small velocity of UE and large physical cell. Among the four methods, NoVC has the largest handover frequency since it does not use virtual cell and a hard handover occurs every time a UE crosses the border of physical cell. The result of RCBD is identical to those of 5gSS in all combinations of R, v, and T_r because virtual cells are not reconfigured according to the UE's movement in both methods, as shown in Equations (4.4) and (4.6).

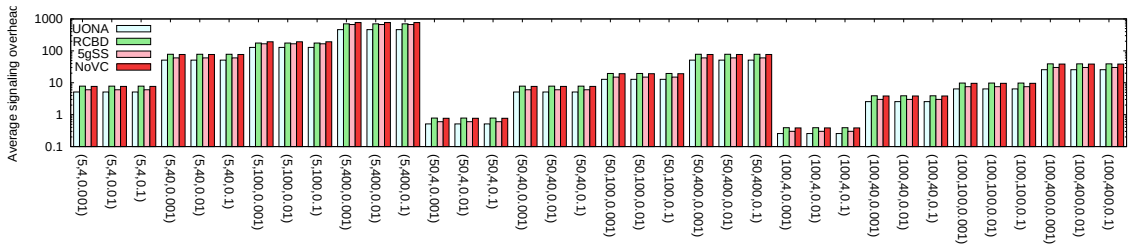


Figure 4.12: Average signaling overhead

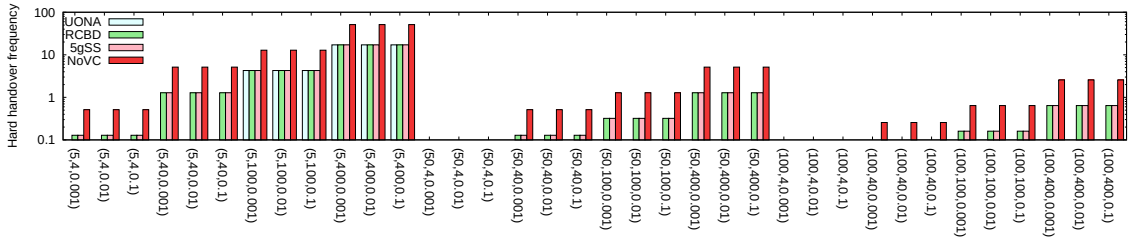


Figure 4.13: Hard handover frequency

On the other hand, UONA has zero frequencies of hard handover in many cases. This is the effect of the combnation of per-user network slicing and its continuous reconfiguration to follow the movement of UE. When the physical cell radius is small and the velocity of UE is large, hard handovers occur even in UONA. However, its frequency is identical to that of RCBD and 5gSS.

4.5 Related work

There are many researches on the end-to-end network slicing for 5G networks [52, 66]. Here we briefly summarize some of them and make clear their relationships to our work.

In [67], the authors show the network architecture of end-to-end network slicing in 5G networks and summarize the related researches, mainly focusing on the issues for slicing of RAN and mobile packet core networks and present some use cases. [68, 69] discuss the orchestration for network slicing in 5G networks. Some researches [70, 71] focus on

the orchestration architecture under multi-domain or multi-operator environment. In all the above researches, it is assumed that the network slice is created and maintained per service/application basis, meaning that each network slice is generated for accommodating multiple UEs under a certain service/application. Also, most of them do not explicitly consider the on-demand and realtime reconfiguration of network slices based on user mobility and temporal network load changes.

For on-demand and realtime network slicing, the authors in [72] proposed the architecture of "Bring Your Own Network." They consider the mobility and distribution of users to create the network slices. However, this work assumed that the network slice is maintained in per-service basis, and per-user network slice is not considered. Also, the main objective of the architecture is to decrease the control-plane signaling overhead, not to improve user-perceived performance. This method can be regarded as one of slice selection methods [60], where a appropriate network slice is selected for each user according to the characteristics of the user and used applications. UONA is different from such slice selection methods where network slices are maintained in a per-user basis and slice reconfiguration is conducted in a unit of subslice.

The authors in [73–75] propose the resource allocation algorithms for end-to-end network slicing in 5G networks, where they do not specify the resolution of network slicing. In our network architecture, we assume that each network component slices are separately maintained by multiple operators under the physical resources they own. As slicing/virtualization methods in each network component in 5G networks, there are many existing works, such as [76–78] for RAN slicing, and [79, 80] for slicing for mobile core networks. These methods can be applied to slice composition by each operator.

The concept of *slice broker* can be found in some existing works. In [81], the authors proposed the capacity broker in RAN slicing. [82, 83] proposed the different idea of slice broker, which selects an end-to-end slice according to user demand. In [84], brokering system is proposed to provide appropriate network configurations to cloud customers in SDN-based networks. These works are different from our architecture, where the slice broker collects subslices from subslice providers to create end-to-end user network slices according to the users' requirements.

4.6 Conclusions

In this chapter, we proposed the novel network architecture, UONA, which manages end-to-end network slices on a per-user basis and realtime and continuous reconfigurations of the slices according to the user mobility and load changes. We also introduced the concept of subslice provider and network slice broker, where end-to-end network slices are constructed from subslices generated from physical resources. We described the fundamental design of the proposed architecture and discuss its advantages against the per-service slicing in 5G networks. Also, numerical evaluation results were presented to demonstrate the effectiveness of UONA. We confirmed the UONA's effectiveness of per-user optimization of virtual cell settings to decrease the control-plane signaling overhead and improve user-perceived performance.

In future work, we plan to make more detailed the network slice brokering model. We also extend the numerical evaluations to include the whole end-to-end network including RAN, EPC, and edge/core cloud environments.

Chapter 5

Conclusion and Future Work

In this thesis, we studied mobile communication system architecture for accommodating IoT device and mobile users.

In Chapter 2, we focused and studied on the mobile core network accommodating IoT devices. 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. The result of this chapter can be applied to the optimization of packet core subslices in UONA proposed in Chapter 4.

In Chapter 3, we presented the performance analysis results of the performance of mobile cellular networks for accommodating C-IoT communications. We conducted the mathematical analysis for deriving the upper limit on number of accommodated UEs in the network, and the time for completing the data transmission for each UE. We assessed the effect of the immediate release of radio resources, and comparison between LTE and NB-IoT. The result of this chapter can be applied to the optimization of RAN subslices in UONA proposed in Chapter 4.

In Chapter 4, we proposed the novel network architecture, UONA, which manages end-to-end network slices on a per-user basis and realtime and continuous reconfigurations of the slices according to the user mobility and load changes. We presented the overall design of UONA and explain its advantages, as well as the research challenges to realize it. We also demonstrated the effectiveness of UONA by presenting the numerical evaluation results. We confirmed that UONA outperforms the service-based slicing in 5G networks

and other methods by selecting control parameters in per-user basis. In UONA, the result of Chapter 2 and Chapter 3 can be used to optimize the network capacity in packet core subslices and RAN subslices, respectively.

There are several challenging tasks as future works. In this study, the effectiveness of UONA was verified by focusing on the lower layers, however, for the deployment of UONA, it is necessary to optimize the system including the upper layers and to solve the function placement problem.

Bibliography

[1] D. Hong and S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, pp. 77–92, Aug. 1986.

[2] C. W. Paper, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2018-2023*, Feb. 2020.

[3] M. Bahrami, "Cloud Computing for Emerging Mobile Cloud Apps," in *Proceedings of 2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, pp. 4–5, Mar. 2015.

[4] C. Christopher, *An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications: Second Edition*. Wiley, May 2014.

[5] A. Sassan, *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*. Academic Press, Nov. 2013.

[6] B. Finley and A. Vesselkov, "Cellular IoT Traffic Characterization and Evolution," in *Proceedings of 2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pp. 622–627, Apr. 2019.

[7] 3GPP TR 36.300 V13.4.0, *Evolved Universal Terrestrial Radio Access (EUTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*, June 2016.

[8] "GSMA | NTT DoCoMo Rolls out LTE-M Across Japan | Internet of Things," available at https://www.gsma.com/iot/mobile-iot-lpwa/ntt-docomo-rolls-out-lte-m-across-japan/.

[9] "AT&T, KDDI, KPN, NTT DOCOMO, Orange, Telefonica, Telstra, TELUS and Verizon Back Deployment of LTE-M for Internet of Things," available at http://news.kddi.com/kddi/corporate/english/newsrelease/2017/02/27/2340.html.

[10] 3GPP TR36.752 V14.0.0, *Study of Narrow-Band Internet of Things (NB-IoT) Radio Frequency (RF) requirement to co-existence with Code Division Multiple Access (CDMA)*, Jan. 2017.

[11] ITU-R, *Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[12] 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, June 2015.

[13] 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.

[14] 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.

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

[16] J. Schlienz and D. Raddino, "Narrowband Internet of Things Whitepaper," *ROHDE SCHWARZ White Paper*, pp. 1–42, 2016.

[17] 3GPP TS 45.820 V13.1.0, *Cellular System Support for Ultra-low Complexity and Low Throughput Internet of Things (CIoT)*, Dec. 2015.

[18] 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*, pp. 3602–3606, June 2013.

[19] X. An, F. Pianese, I. Widjaja, and U. G. Acer, "DMME: Virtualizing LTE mobility management," in *2011 IEEE 36th Conference on Local Computer Networks*, pp. 528–536, Oct. 2011.

[20] ——, "DMME: A Distributed LTE Mobility Management Entity," *Bell Labs Technical Journal*, vol. 17, no. 2, pp. 97–120, Sept. 2012.

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

[22] 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, July 2014.

[23] 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.

[24] 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*, pp. 389–396. IEEE, May 2014.

[25] 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*, pp. 74–80. IEEE, Nov. 2016.

[26] 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)*, pp. 1–6, May 2017.

[27] 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*, pp. 33–38. ACM New York, NY, USA, Aug. 2014.

[28] 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)*, pp. 8–14b, Nov. 2013.

[29] 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*, pp. 43–48, Oct. 2016. [Online]. Available: http://doi.acm.org/10.1145/2980055.2980062

[30] 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)*, pp. 1–6, Dec. 2015.

[31] 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.

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

[33] A. ElNashar and M. El-saidny, *Practical Guide to LTE-A, VoLTE and IoT: Paving the way towards 5G*. Wiley, 2018.

[34] O. Liberg, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things: Technologies, Standards, and Performance*. Elsevier Science, 2017. [Online]. Available: https://books.google.co.jp/books?id=n9wkDwAAQBAJ

[35] S. Abe, G. Hasegawa, and M. Murata, "Performance Analysis of Periodic Cellular-IoT Communication with Immediate Release of Radio Resources," in *Proceedings of CQR2020*, pp. 1—-6, 2020.

[36] B. Martínez, F. Adelantado, A. Bartoli, and X. Vilajosana, "Exploring the Performance Boundaries of NB-IoT," *CoRR*, vol. abs/1810.00847, 2018.

[37] B. Hsieh, Y. Chao, R. Cheng, and N. Nikaein, "Design of a UE-specific Uplink Scheduler for Narrowband Internet-of-Things (NB-IoT) Systems," in *Proceedings of 2018 3rd International Conference on Intelligent Green Building and Smart Grid (IGBSG)*, pp. 1–5, Apr. 2018.

[38] M. Lauridsen, I. Z. Kovacs, P. Mogensen, M. Sorensen, and S. Holst, "Coverage and Capacity Analysis of LTE-M and NB-IoT in a Rural Area," in *Proceedings of 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, Sept. 2016.

[39] S. Persia and L. Rea, "Next Generation M2M Cellular Networks: LTE-MTC and NB-IoT Capacity Analysis for Smart Grids Applications," in *Proceedings of 2016 AEIT International Annual Conference (AEIT)*, pp. 1–6, Oct. 2016.

[40] M. Hasan, E. Hossain, and D. Niyato, "Random Access for Machine-to-machine Communication in LTE-advanced Networks: Issues and Approaches," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 86–93, June 2013.

[41] G. C. Madueno, J. J. Nielsen, D. M. Kim, N. K. Pratas, C. Stefanovic, and P. Popovski, "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 675–688, Mar. 2016.

[42] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Muñoz, and J. M. López-Soler, "Optimized LTE Data Transmission Procedures for IoT: Device Side Energy Consumption Analysis," *CoRR*, vol. abs/1704.04929, 2017.

[43] 3GPP TS 36.213 V15.7, *Physical layer procedures (Release 15)*, Sept. 2019.

[44] M. Sauter, *From GSM to LTE-Advanced Pro and 5G: An Introduction to Mobile Networks and Mobile Broadband*. Wiley, 2017.

[45] J. Y. Cheah and J. M. Smith, "Generalized M/G/C/C State Dependent Queueing Models and Pedestrian Traffic Flows," *Queueing Systems*, vol. 15, pp. 365–386, 1994.

[46] S. Abe, G. Hasegawa, and M. Murata, "Effects of C/U Plane Separation and Bearer Aggregation in Mobile Core Network," *IEEE Trans. Network and Service Management*, vol. 15, no. 2, pp. 611–624, 2018.

[47] B. Avi-Itzhak and S. Halfin, "Expected Response Times in a Non-Symmetric Time Sharing Queue with a Limited Number of Service Positions," in *Proceedings of ITC-12*, pp. 1485—-1493, 1988.

[48] 3GPP TS 38.101-1 V15.10.0, *5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 15)*, Sept. 2017.

[49] Y. Chen, L. Cheng, and L. Wang, "Prioritized resource reservation for reducing random access delay in 5G URLLC," in *Proceedings of 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, 2017.

[50] J. Thota and A. Aijaz, "On Performance Evaluation of Random Access Enhancements for 5G URLLC," in *Proceedings of 2019 IEEE Wireless Communications and Networking Conference (WCNC)*, p. 1–7. IEEE Press, 2019.

[51] S. Rommer, P. Hedman, M. Olsson, L. Frid, S. Sultana, and C. Mulligan, *5G Core Networks: Powering Digitalization*. Elsevier Science, 2019. [Online]. Available: https://books.google.co.jp/books?id=82C-DwAAQBAJ

[52] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Communications Magazine*, vol. 55, pp. 94–100, May 2017.

[53] M. D. Ree, G. Mantas, A. Radwan, S. Mumtaz, J. Rodriguez, and I. E. Otung, "Key management for beyond 5G mobile small cells: A survey," *IEEE Access*, pp. 59 200–59 236, May 2019.

[54] Z.-Y. Wu, M. Ismail, J. Kong, E. Serpedin, and J. Wang, *Efficient Integration of 5G and Beyond Heterogeneous Networks*. Springer, 2020.

[55] Q. M. Qadir, T. A. Rashid, N. K. Al-Salihi, B. Ismael, A. A. Kist, and Z. Zhang, "Low power wide area networks: A survey of enabling technologies, applications and interoperability needs," *IEEE Access*, pp. 77 454–77 473, Nov. 2018.

[56] X. Yang, Y. Liu, I. C. Wong, Y. Wang, and L. Cuthbert, "Effective isolation in dynamic network slicing," in *Proceedings of IEEE WCNC 2019*, Apr. 2019.

[57] J. Cao, W. S. Cleveland, and D. X. Sun, "Bandwidth estimation for best-effort Internet traffic," *Statistical Science*, vol. 19, pp. 518–543, 2004.

[58] X. Dong, L. Zhao, H. Zhao, and C. Pan, "RAN slicing-based handover scheme in HetNets," in *Proceedings of IEEE DSP 2018*, Nov. 2018.

[59] J. Gutiérrez, N. Maletic, D. Camps-Mur, E. García, I. Berberana, M. A. a nd Anna Tzanakaki, V. Kalokidou, P. Flegkas, D. Syrivelis, T. Korakis, P. Legg, D. Markovic, G. Limperopoulos, J. Bartelt, J. K. Chaudhary, M. Grieger, N. Vucic, J. Zou, and E. Grass, "5G-XHaul: A Converged Optical and Wireless Solution for 5G Transport Networks," *Transactions on Emerging Telecomminications Technologies*, July 2016.

[60] J. J. D. Rivera, T. A. Khan, A. Mehmood, and W.-C. Song, "Network slice selection function for data plane slicing in a mobile network," in *Proceedings of APNOMS 2019*, Sept. 2019.

[61] 3GPP TS 32.899 V15.1.0, "Telecommunication management; Charging management; Study on charging aspects of 5G system architecture phase 1," Mar. 2018.

[62] J. G. Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, pp. 518–532, Aug. 2016.

[63] N. Meng, H. Zhang, and H. Lu, "Virtual cell-based mobility enhancement and performance evaluation in ultra-dense networks," in *Proceedings of IEEE WCNC 2016*, Apr. 2016.

[64] V. Casares-Giner, V. Pla, and P. Escalle-García, "Mobility models for mobility management," in *Network Performance Engineering*, D. D. Kouvatsos, Ed. Berlin, Heidelberg: Springer, 2016, pp. 716–745.

[65] N. Meng, H. Zhang, and B. Lin, "User-centric mobility management based on virtual cell in ultra-dense networks," in *Proceedings of ICCC 2016*, July 2016.

[66] S. Zhang, "An overview of network slicing for 5G," *IEEE Wireless Communications*, vol. 26, pp. 111–117, Apr. 2019.

[67] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-end network slicing for 5G mobile networks," *Journal of Information Processing*, vol. 25, pp. 153–163, Feb. 2017.

[68] F. Meneses, M. Fernandes, D. Corujo, and R. L. Aguiar, "SliMANO: An expandable framework for the management and orchestration of end-to-end network slices," in *Proceedings of IEEE CloudNet 2019*, Nov. 2019.

[69] R. Montero, F. Agraz, A. Pagès, and S. Spadaro, "End-to-end 5g service deployment and orchestration in optical networks with qoe guarantees," in *Proceedings of ICTON 2018*, July 2018.

[70] T. Lin, S. Marinova, and A. Leon-Garcia, "Towards an end-to-end network slicing framework in multi-region infrastructures," in *Proceedings of IEEE NetSoft 2020*, June 2020.

[71] Q. W. et al., "SliceNet: End-to-end cognitive network slicing and slice management framework in virtualised multi-domain, multi-tenant 5G networks," in *Proceedings of IEEE BMSB 2018*, June 2018.

[72] Y. Shoji, L. Zhong, M. Ito, Y. Kitatsuji, K. Nakauchi, and H. Yokota, "Bring your own network – a network management technique to mitigate the impact of signaling traffic on network resource utilization –," in *Proceedings of CCNC 2014*, July 2014.

[73] D. Harutyunyan, R. Fedrizzi, N. Shahriar, R. Boutaba, and R. Riggio, "Orchestrating end-to-end slices in 5G networks," in *Proceedings of IEEE CNSM 2019*, Oct. 2019.

[74] H. Halabian, "Distributed resource allocation optimization in 5G virtualized networks," *IEEE Journal of Selected Areas on Communications*, vol. 37, pp. 627–642, Mar. 2019.

[75] X. Li, R. Ni, J. Chen, Y. Lyu, Z. Rong, and R. Du, "End-to-end network slicing in radio access network, transport network and core network domains," *IEEE Access*, pp. 29 525–29 537, Feb. 2020.

[76] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Communications Magazine*, vol. 55, pp. 102–108, June 2017.

[77] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Communications Magazine*, vol. 56, pp. 184–192, Jan. 2018.

[78] M. R. Raza, C. Natalino, P. Öhlen, L. Wosinska, and P. Monti, "Reinforcement learning for slicing in a 5G flexible RAN," *Journal of Lightwave Technology*, vol. 37, pp. 5161–5169, Oct. 2019.

[79] M. Rayani, D. Naboulsi, R. Glitho, and H. Elbiaze, "Slicing virtualized EPC-based 5G core network for content delivery," in *Proceedings of ISCC 2018*, June 2018.

[80] A. Mohammadkhan and K. Ramakrishnan, "Re-architecting the packet core and control plane for future cellular networks," in *Proceedings of ICNP 2019*, Oct. 2019.

[81] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Communications Magazine*, vol. 54, pp. 32–39, July 2016.

[82] J. Backman, S. Yrjölä, K. Valtanen, and O. Mämmelä, "Blockchain network slice broker in 5G: Slice leasing in factory of the future use case," in *Proceedings of the Conference on Internet of Things - Business Models, Users, and Networks*, Nov. 2017.

[83] C. Gutterman, E. Grinshpun, S. Sharma, and G. Zussman, "RAN resource usage prediction for a 5G slice broker," in *Proceedings of ACM Mobihoc 2019*, July 2019.

[84] J.-L. Chen, Y.-W. Ma, P.-S. Chiu, and D.-W. Jiang, "SDNBroker: Heterogeneous cloud serving systems over software-defined networking," in *Proceedings of CITS 2014*, July 2014.