論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Study on Network Resource Assignment for Efficient Communication Accommodation
著者(和文)	田辺和輝
Author(English)	Kazuki Tanabe
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11136号, 授与年月日:2019年3月26日, 学位の種別:課程博士, 審査員:山岡 克式,植松 友彦,尾形 わかは,府川 和彦,山田 功,北口 善明
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11136号, Conferred date:2019/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

A Study on Network Resource Assignment for Efficient Communication Accommodation

by

Kazuki Tanabe

A DISSERTATION

Submitted to
Tokyo Institute of Technology
in partial fulfillment of the requirements
for the degree of

DOCTOR OF ENGINEERING

2019

Acknowledgements

I am now very satisfied and appreciated for completing my Ph.D. thesis, with a remembrance of numerous unforgettable events during my Ph.D. studies. This study was supported by a lot of people whom I have met in my research life, and was never completed without their help, advices and cooperation.

First, I would like to express my sincere gratitude to my academic advisor, Prof. Katsunori Yamaoka. I have been the member of the Yamaoka Lab since I was an undergraduate student in fourth grade, and he has guided me with fruitful discussions and advices throughout my studies in the Lab. Especially, I could never forget his support after my failure in the entrance examination of the graduate school in 2012. After the notification of the examination result, Prof. Yamaoka reached his hand to me, and discussed my future life with me overnight. If he did not accept me as a researcher student next fiscal year, everything in the rest of my life would have changed. After I entered the graduate school and began my Ph.D. studies, he has told me not only the method for pursuing my research, but also how to improve my leadership and become a true leader in the global society. Without his selfless and continuous support, the efforts of my study did not come to friution and I could not complete my Ph.D. studies.

I would also like to thank Prof. Yoshiaki Kitaguchi, who has been an associate professor at Global Scientific Information and Computing Center, Tokyo Institute of Technology since 2017. After he joined Tokyo Tech and began the collaboration as the Yamaoka and Kitaguchi Labs, he gave me a lot of advices based on his abundant experience in network management. Moreover, during the hospitalization of Prof. Yamaoka, he had always worked hard for creating better environment for the members of the Yamaoka and Kitaguchi Labs, and was also in charge of the chief examiner of my Ph.D. thesis.

I also sincerely thank Prof. Sumiko Miyata at Shibaura Institute of Technology. Prof. Miyata has specialized in the Call Admission Control (CAC), and created an opportunity for me to begin a new research topic on a CAC in my undergraduate studies. When I faced difficulty and got lost in my study, she always gave me suggestive and novel advices. She also gave me a lot of advices when I was deciding whether to continue my studies in the doctoral course, as an ex-Ph.D. student in the Yamaoka Lab.

I am also truly grateful to Prof. Ken-ichi Baba at Kogakuin University. Prof. Baba has specialized in the CAC and photonic networking, and he always gave me constructive

opinions to improve my studies. He also gave me some chances to begin research projects with undergraduate students of Kogakuin University. I could have a precious and practical experience as a supervisor in these related research topics. Chapter 3 in this thesis is based on the research project with Prof. Miyata and Prof. Baba.

Moreover, I would like to express my deepest appreciation to Mr. Tsunemasa Hayashi and Mr. Hiroki Nakayama at BOSCO Technologies, Inc., for giving me an opportunity of research collaboration and having fruitful discussions. Mr. Hayashi accepted me as a part-time worker when I was in the master course. In the part-time job, I could have a lot of experience to solve the problems in the real network environment. He also gave me an opportunity of research collaboration with the Yamaoka Lab and his advices based on his experience in both research and business refined my research much more realistic and persuasive. Mr. Nakayama has been an expert engineer at BOSCO Technologies, and kindly tell me the way of programming and development in the part-time job, and also told me about technical trends in the field of network management. Chapter 4 in this thesis is based on the research project with BOSCO Technologies.

I would also like to thank all the colleagues at the Yamaoka and Kitaguchi Labs, Mr. Masaya Endo, Mr. Takuya Kosugiyama, Mr. Takumi Kada and Mr. Ryota Murakami, who worked on my research project. I also truly thank Mr. Kenta Kawai at Kogakuin University, who also joined my research project and tackled some related studies. Their efforts helped my studies to be much more concrete.

In addition, I would like to give my thanks to all the members of the Yamaoka and Kitaguchi Labs. I am sorry that I cannot express my feelings to everybody, since I met countless people there. Let me give special thanks to those with whom I spent a great deal of time with. Dr. Makoto Misumi at Fukuoka University, who completed his Ph.D. studies when I was an undergraduate student, helped me with his plentiful knowledge on networking protocols. Mr. Yuuhei Hayashi and Mr. Hironori Katagiri, who are my senior students, gave me a lot of advices to complete my bachelor thesis. Mr. Akihiro Terashima, who had been one of my classmates since the freshman year, helped and cheered each other up as a colleague and a friend in my undergraduate studies. Mr. Rikuho Sakamoto, Mr. Sho Noda, Mr. Takuya Okamoto and Mr. Taichi Miya, who are my junior students, supported me with the managing and maintaining the Lab network environment. Mr. Krittin Intharawijitr, who is a doctoral student at the Yamaoka and Kitaguchi Labs, helped me as an only one colleague in doctoral course, with his research experience and excellent English skills. Mr. Intharawijitr and I also spent a lot of enjoyable time with junior colleagues, Mr. Takeshi Akaoka, Mr. Takumi Matsuura and Mr. Issei Nakasone, and could overcome the heavy tasks in the peak season. I also would like to thank Ms. Mari Nakata, Ms. Rie Yamasaki and Ms. Kaoru Matsuzaki, secretaries at the Yamaoka

and Kitaguchi Labs. Throughout my studies in the Yamaoka and Kitaguchi Labs, I have received a lot of help and support in daily life from them. Especially, Ms. Nakata had supported me with my student life and discussed my private issues with meaningful advices. She also cooked delicious cuisines for the Lab members and I was usually helped by her when I was in financially difficult situations.

I am grateful to the financial support from the Japan Society for Promotion of Science (JSPS). I was helped by their support for my living expenses and research funds during the whole period of my Ph.D. studies.

These people are just a little part of those who have helped my Ph.D. studies, and I feel terribly sorry for not being able to introduce and express my gratitude to everybody here. I am feeling that I am one of the happiest person in the world for being surrounded by such marvelous people and having numerous irreplaceable experiences with them.

Of course, I can never forget to thank my mother, Hideko Tanabe, and my little sister Risa Tanabe. They have always supported me even though they are far apart in Hokkaido. Also, this dissertation is dedicated to my father Haruo Tanabe in heaven.

Finally, I would like to express my sincere thanks to my partner, Ayami Meguro. She had been also a member of the Yamaoka Lab since 2014 until 2016. Without an encounter with her and her continuous support for my decision to go on to the doctoral course, my life would have changed at all and my Ph.D. studies cannot be established.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Nowadays, owing to drastic spread of mobile devices e.g. smartphones, tablets and laptops, the traffic amount on telecommunication networks is rapidly increasing. These mobile devices frequently access to the Internet and communicate a large amount of traffic data, on both cellular networks and wireless LAN (WLAN) networks. Cisco has reported [1] that the monthly global mobile traffic on the Internet will count up to 49 [EB] by 2021, and it will represent 20 percent of total IP traffic (Fig. 1.1). To accommodate such a large amount of data traffic and satisfy the demands of users' applications, telecommunication network operators are required to enhance their network equipments and control the data traffic properly.

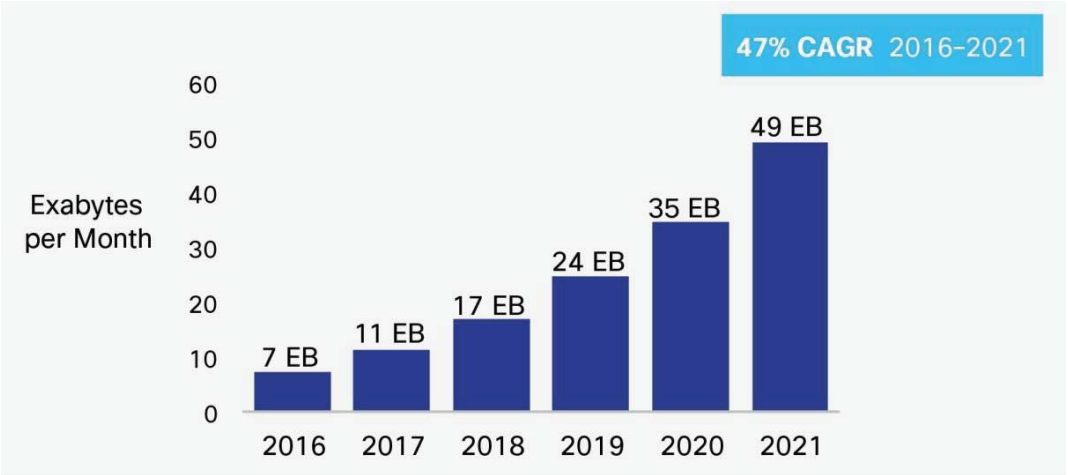


Fig. 1.1: Forecasts of Mobile Data Traffic by 2021 [1]

In the field of telecommunication networks, various communication technologies have been continuously studied and have evolved to meet such demands. For example, the technical standard of 5G cellular networks has been almost determined lately. In 5G networks, the maximum bandwidth reaches the order of 10 [Gbps], and mobile devices can send and receive much larger size of data in a shorter time. This high-speed mobile communication enables real time communications with richer data contents e.g. live streaming video, online game application etc.

In addition, some new types of mobile communications are rapidly growing in recent. The Internet of Things (IoT) [2] is a concept in which various kinds of sensor devices are connected to the Internet and can be accessible via wireless networks. As shown in Fig. 1.2, IoT can be operated and make paradigm shifts in a wide variety of target areas. For example, users can access to and control the devices at home by their smartphones. In the field of logistics, the information of materials e.g. location, destination, temperature, humidity, etc. can be tracked via the Internet. By using remote control devices via the Internet, a medical surgery can be operated for the patient in a distant place.

Fig. 1.2: Application Range of the Internet of Things [2]

Meanwhile, the machine-to-machine (M2M) communication is a concept of automated direct communications between devices. Although the target of the M2M communication is both wired and wireless connection, it is mainly operated by wireless communications between sensor devices and is sometimes considered as a subset of IoT [3].

Recently, connected cars have been attracted a lot of attention as an application of the IoT/M2M communications. Since connected cars are required to realize the automated driving, they are actively studied and developed by both academic and industrial society. Communications between connected cars and other devices e.g. connected cars, mobile base station, infrastructure of the Intelligent Transport Systems (ITS), are defined as the V2X communications, and are known for their strict requirements for communication latency. For example, V2X communications requires the maximum end-to-end delay of 100 [msec] in 4G networks [4], and 10 to 25 [msec] in 5G networks [5].

In 5G networks, three communication types are defined as the usage scenarios: enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC), and Ultra-Reliable and Low Latency Communications (URLLC). URLLC is especially defined for delay-sensitive communications, and has the strictest requirements for both latency and packet loss among these three usage scenarios. For example, the data plane end-to-end delay should be less than 0.5 [msec] for both uplink and downlink in URLLC [6]. In order to follow these usage scenarios and guarantee the Quality of Service (QoS) of each communication, telecommunication network operators are required to allocate the network resources much more appropriately than in conventional networks, according to current traffic situation.

According to the history of technological advancements in telecommunication networks mentioned above, both the number of communication devices and the amount of each communication traffic are continuously increasing. In order to accommodate a large amount of communication traffic properly, management of network resources e.g. bandwidth, sessions, VNFs, etc. is becoming much more important for telecommunication network operators.

In the following sections, major innovations related to network resources in the field of telecommunications are introduced.

1.2 Migration of Telecommunication Network into IP Network

Before the Internet became widespread among the public, the telecommunication system had been operated on the Public Switched Telephone Network (PSTN) and the telephone call requests had been processed by the telephone exchange. In the telephone exchange, each trunk circuit is assigned to each accommodated telephone call by the Common Channel Signaling System No. 7 (SS7) protocol. In the SS7 protocol stack, the communication resource for signaling control (Control plane, C-plane) is separated from the communication resource for actual voice signals (User plane, U-plane) and the communication QoS is guaranteed during the whole connection. However, due to deterioration of the legacy equipments including telephone exchange and high maintenance costs of these equipments, telecommunication networks had been required to be migrated into a new network architecture.

In the 2000's decade, the Next Generation Network (NGN) [7] have been studied and developed to alternate the conventional PSTN. The ITU-T defines that the NGN is a packet-based network able to provide telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies [8]. The mission of the NGN is to integrate the circuit-based telecommunication network into packet-based IP network. After the standardization of the NGN, telecommunication networks are migrated into the NGN in a lot of countries, and the operation of the PSTN is supposed to be terminated by 2025 in Japan [9].

In the NGN, the network architecture for multimedia communications is defined as the IP Multimedia Subsystem (IMS), and signaling of telephone call requests are controlled by Session Initiation Protocol (SIP), which has been used for IP telephony. When a new telephone call request arrives, the SIP server registers the call session and connects between two User Equipments (UEs). After the call session is initiated, the user voice data is encoded into packets and voice data packets are exchanged by the Real-time Transfer Protocol (RTP) session. Since RTP uses UDP for voice data transmission, and delayed or lost data is not retransmitted, the NGN has some priority classes and these voice data packets have higher priority than other IP packets. However, even though the communication bandwidth is theoretically reserved for each accommodated call, the data transmission is operated by packet exchange and the communication link is shared with other telephone calls and other types of Internet traffic. To avoid traffic congestion and accommodate the telephone traffic properly in such heterogeneous network environment, a call admission control (CAC) method is required in the NGN.

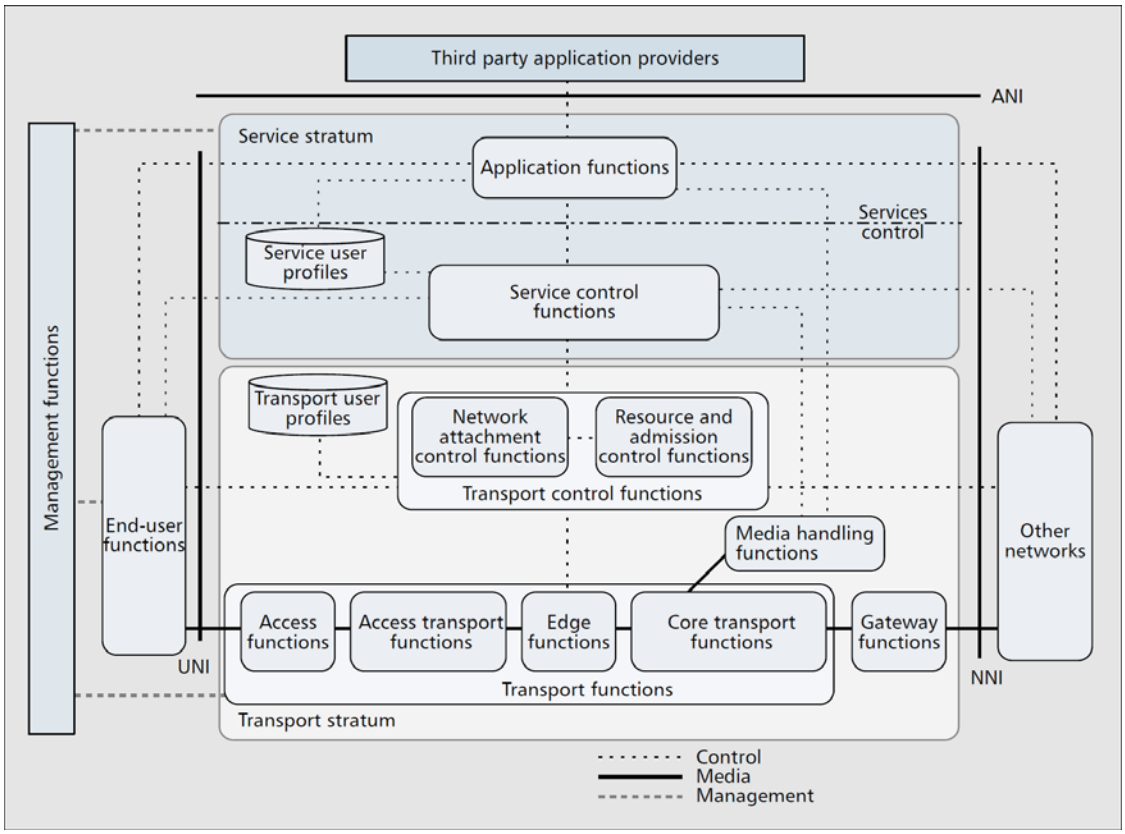


Fig. 1.3: NGN Architecture Overview [7]

1.3 Softwarization of Network Resource

Furthermore, the advancement in both hardware and software has promoted the computational power of communication devices, and has changed the whole architecture of telecommunication networks. On one hand, Software Defined Networking (SDN) is a state-of-the-art technology operated in not only enterprise data center networks but also the core networks of telecommunications operators [11]. In the SDN architecture, the network plane slicing of the C/U-plane is inherited as the control plane (C-plane) and the data plane (D-plane). The C-plane controls the routes of data packets by forwarding rules, which is determined and managed by the application system called as the management plane (M-plane). The D-plane simply processes and transmits data packets, according to the packet transfer rules determined on the C-plane. According to the separation of these two planes and the programmable packet control, SDN enables a flexible management in the IP network, and also in IP-based NGNs.

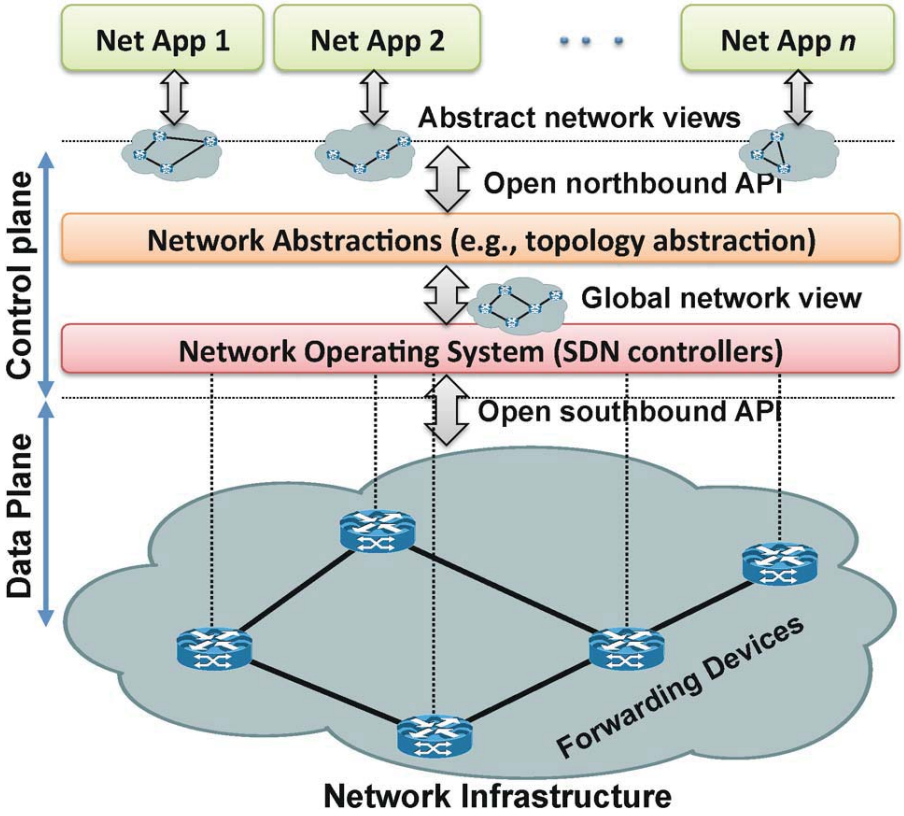


Fig. 1.4: SDN Architecture and Its Fundamental Abstractions [11]

On the other hand, Network Functions Virtualization (NFV) is widely studied and developed in both academic and industrial society. The concept of NFV is proposed by ETSI [12], and Fig. 1.5 shows the NFV architectural framework. In the NFV, each network resource is deployed as a virtual machine defined as the Virtual Network Function (VNF). Each VNF is deployed on the NFV Infrastructure (NFVI), which is a set of virtualized infrastructure such as computing resource, storage and network, and is hosted on some hardware resources. The VNFs and the NFVI are managed by a system defined as the NFV Management and Network Orchestration (MANO).

There are two advantages of the NFV: low cost network deployment and flexible network management. Because the NFV enables various network entities to operate on a virtualized environment, both the capital expenditure (CAPEX) and the operation expenditure (OPEX) can be reduced, compared to the legacy hardware equipments e.g. routers, switches, firewalls etc. In addition, the computational resources of each VNF can be easily controlled according to the traffic demand, as long as the hardware resources of the NFVI e.g. CPU cores, RAM capacity, storage capacity, etc. are sufficient.

In recent telecommunication networks, the functions of various network entities are

operated by the NFV, and deployed as the VNFs in general-purpose servers. Compared to the conventional networking system based on physical hardware, the communication resource of each virtualized network entity should be appropriately assigned, to satisfy the required QoS criteria determined for each communication type.

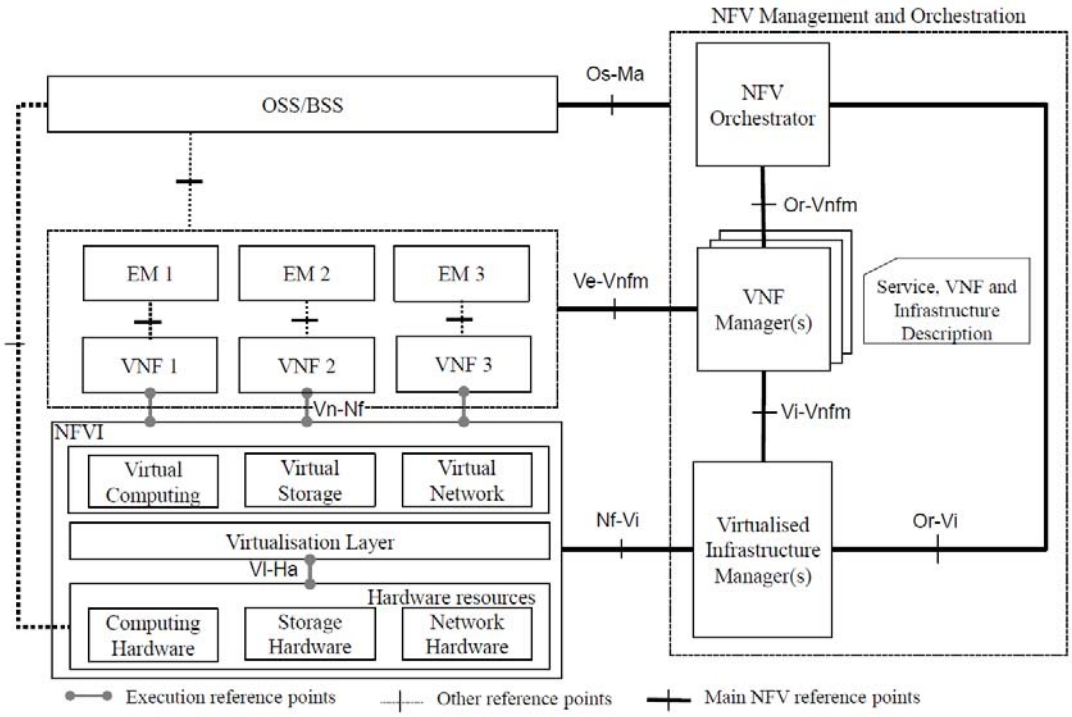


Fig. 1.5: NFV reference architectural framework [12]

Chapter 2

Purpose and Overview of Study

2.1 Issues

As mentioned in Chapter 1, the technologies in the field of telecommunication have been advanced to satisfy the users' traffic demands. Since the core networks of both wired and mobile communications are based on IP network, proper assignment of communication resources becomes much more important, in comparison with the conventional network such as PSTN or Asynchronous Transfer Mode (ATM) network. If an excessive number of communication sessions are accommodated in such IP-based environment without control, heavy congestion on both the C-plane and the D-plane degrades the QoS of the entire network.

For example, in the IP-based telephone networks such as NGN, call session management is done by the SIP protocol and users' voice data is exchanged as the VoIP packets. Accommodating too many sessions may increase the packet loss rate and end-to-end delay of each VoIP session. Thus, the network operator is required to determine the maximum number of VoIP sessions (defined as number of lines in this study), with which the QoS of each accommodated session is guaranteed.

On the other hand, in the virtualized mobile core network, the capacity of accommodated sessions can be flexibly modified by the VM resource management of the C-plane entities. However, excessive deployment of session accommodation capacity also increases the amount of user data packets. If the incoming data traffic exceeds the packet processing capability of the D-plane entities, the QoS requirement of mobile communications, such as delay sensitive M2M communications, real time live streaming video traffic and high data rate traffic of smartphone game applications, cannot be guaranteed. Since the total hardware resource of the host server is limited, and shared by both the C-plane and D-plane, the mobile network operator is required to determine both the maximum number of session accommodation and data packet processing capability, so that the QoS of every accommodated session is satisfied.

Therefore, for guaranteeing the communication QoS in the congested network environment, it is required to appropriately assign the network resource, in accordance with current traffic demands. However, there still have been some problems around telecommunication networks. In this chapter, some major issues around telecommunication networks are introduced, and the purpose of this study is discussed. Moreover, the overview of this study is explained.

2.1.1 Congestion on Telephone Networks during Emergencies

During emergencies e.g. natural disasters, conflicts, terrorisms, etc., telecommunication networks become heavily congested by a large number of telephone call requests. For example, the fixed-line telephone traffic increased to 4 to 9 times the usual traffic amount, and the mobile telephone traffic also increased to 50 to 60 times the usual traffic amount, after the Great East Japan Earthquake in 2011 [13]. Figure 2.1 shows the increase in cellular phone calls in northern Japan area, right after the Great East Japan Earthquake [14]. Moreover, some network infrastructures become damaged and the network operators are required to operate their telecommunication networks with limited infrastructures. After the Great East Japan Earthquake, approximately 1.9 million subscriber lines were damaged and about 29,000 mobile base stations were damaged [13]. Figure 2.2 shows the change in the number of damaged telecommunication facilities [15].

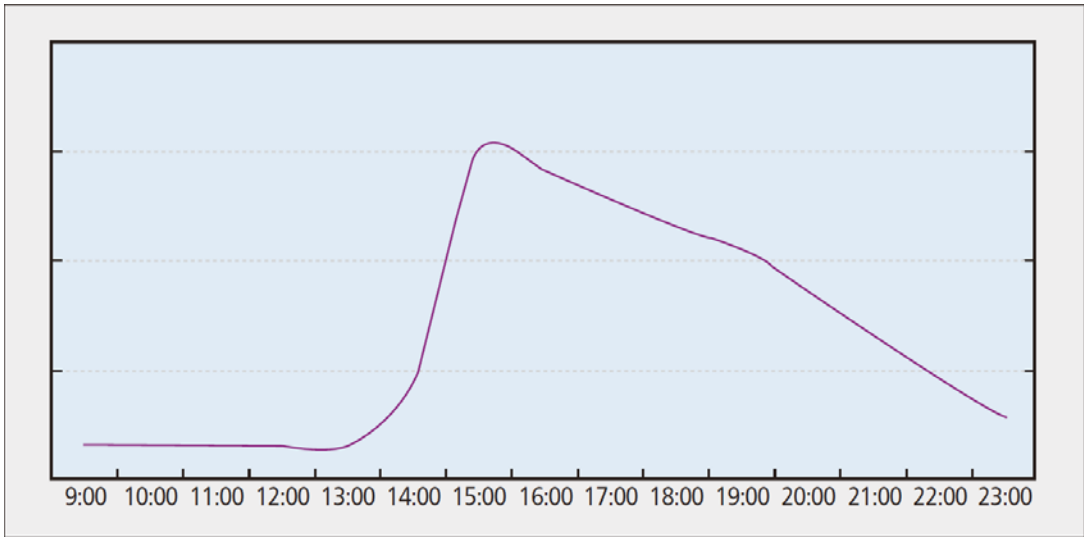


Fig. 2.1: Number of Cellular Phone Calls on March 11, 2011 in Northern Japan [14]

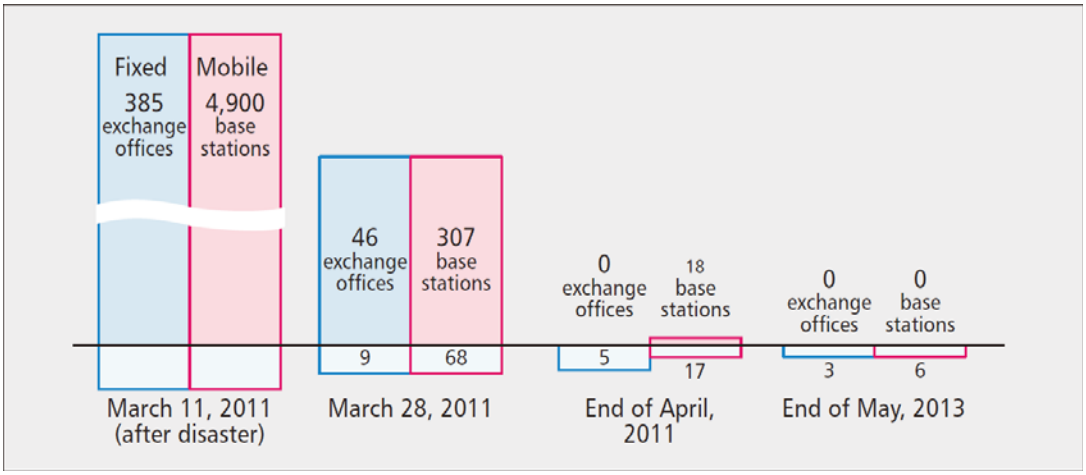


Fig. 2.2: Number of Damaged Telecommunication Facilities after the Great East Japan Earthquake [15]

To facilitate rescue, report and restoration during emergencies, the voice communications of important users e.g. police, fire station, media, etc. should be accommodated, even in such a congested and limited network environment. The ITU-T defines Emergency Telecommunication Service (ETS) [16] for protecting important communications during emergencies, and the users of ETS (emergency call users) are determined to have higher priority to other general calls. In some countries, telecommunication network operators determine some specific telephone users as the emergency call users e.g. the Government Emergency Telecommunications Service (GETS) in the United States. In the core network such as telephone exchange in the PSTN and SIP server in the NGN, several communication resources are reserved for emergency calls in advance. When the telephone exchange or SIP server is extremely congested, network operators also restrict call requests of general calls and guarantee the communication resources for emergency calls. However, immoderate call restriction may grow the uneasiness of general call users and result to increase in general call reattempts. Even though the communication period is shortly limited, general call users can relieve their fear and uneasiness by hearing the real voice of their families or friends. Therefore, general calls should also be accommodated, as long as the required number of emergency calls are certainly accommodated.

2.1.2 Congestion on Virtualized Mobile Core Networks

As mentioned in Sect. 1.1, the traffic amount of mobile data communications is continuously increasing, and the usage of the mobile communications is becoming much wider. Among them, the IoT and M2M are new concepts of communications and the demands

of these communications are especially rising. M2M communications will be used for various use cases e.g. autonomous car or connected car, Augmented Reality (AR). These M2M communications are known for strict end-to-end delay constraint, and their data traffic may arrive in bursts because the network of M2M communications consists of a large number of mobile devices. In the future cyber communication society where a lot of equipments and systems e.g. public infrastructure, transportation, living facilities, are connected to the network as IoT/M2M devices, a slight increase in the end-to-end delay may result in a danger situation of our human's life. Nevertheless, the conventional mobile core networks which are composed of physical hardware equipments, are sometimes difficult to deal with the drastic change in such traffic demands.

Recently, as an application the NFV, some part of the core network of 4G mobile cellular network, called as the Evolved Packet Core (EPC) [17] shown in Fig. 2.3, have been virtualized and operated as the Virtualized EPC (vEPC). The conventional EPC is composed of some types of hardware equipments called as entities, and each entity has a different function. In the vEPC network, these EPC entities are deployed as virtual machines (VMs) on the general-purpose server called as the vEPC server. As mentioned to the SDN architecture in Sect. 1.3, the network resources in vEPC networks are also separated into the C-plane and the D-plane. On the C-plane, the two key functions, signaling packet processing and session accommodation, are operated by Mobility Management Entity (MME). On the D-plane, user data packets are processed and transmitted inside the career network by Serving Gateway (S-GW), while the packets are processed and transmitted to the external IP networks by Packet Data Network Gateway (P-GW). The packet processing function of these two D-plane entities are usually combined as the S/P-GW in vEPC networks. The virtualization of these mobile core entities and the network slicing enable mobile network operators to reduce the CAPEX and OPEX of their career network, and to flexibly manage the processing resource of each vEPC entity.

Similarly, the core networks of the forthcoming 5G mobile cellular network, called as the 5G Core (5GC) [18] shown in Fig. 2.4, will also be composed of virtualized equipments. In the 5GC, the data packet processing function of S/P-GW is inherited to the User Plane Function (UPF). On the other hand, the function of MME, signaling packet processing and management of accommodated sessions, is divided into two entities, Access and Mobility management Function (AMF) and Session Management Function (SMF), respectively. As it has been important to manage ongoing sessions properly on MME in the vEPC network, the session management becomes much more important in the virtualized 5GC network.

However, there still have been some problems in the conventional virtualized mobile core network. Although the flexible resource assignment is an advantage of the virtualized

mobile core network, the VM resource shortage of entities on one plane will result in heavy congestion on both planes, since most of these virtualized entities are deployed on the same vEPC server, and required to share the fixed hardware resource e.g. CPU cores, threads, RAM capacity, storage capacity etc. For guaranteeing the service level agreement (SLA) which is a commitment between network operators and users, the mobile network operators are required to avoid such congestion caused by network resource shortage on each plane.

In order to accommodate delay sensitive M2M communications, a certain amount of VM resource is required for the MME. If the MME becomes congested by a large number of signaling packets from M2M communications, the increased signaling packet processing delay also affects the end-to-end delay on the D-plane and the strict allowable delay of M2M communications cannot be guaranteed. On the other hand, the VM resource shortage of the S/P-GW may lead to heavy congestion of user data packets on the D-plane. As mentioned in Sect. 1.1, the number of mobile communication devices are continuously increasing, and each communication bandwidth varies from narrowband communications of IoT/M2M devices to broadband communications of smartphones and tablets. In addition, demands of real time communications with rich contents such as livestreaming video and online game application, are especially increasing. For guaranteeing QoS of such real time communications on the D-plane, a certain packet processing capability of the S/P-GW is also required.

Therefore, to appropriately accommodate incoming communication sessions and process a large number of user data packets in a cost effective virtualized mobile core network, the VM resources should be properly assigned to each entity, according to the traffic situation.

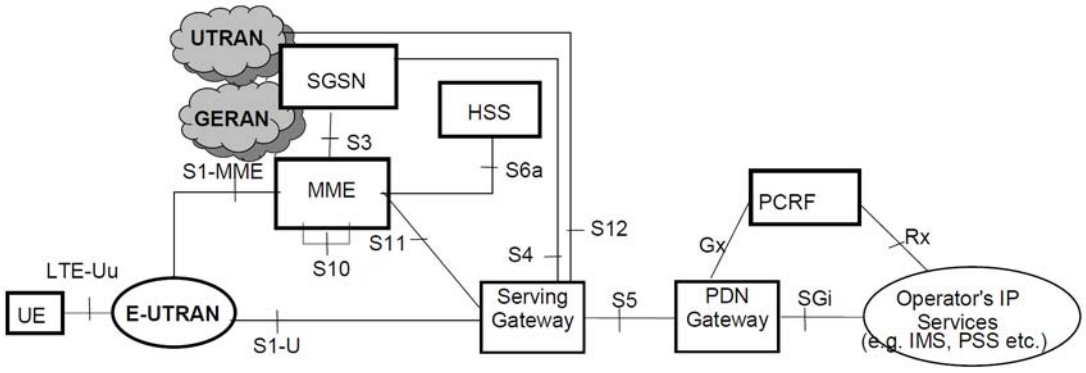


Fig. 2.3: Main Architecture of EPC [17]

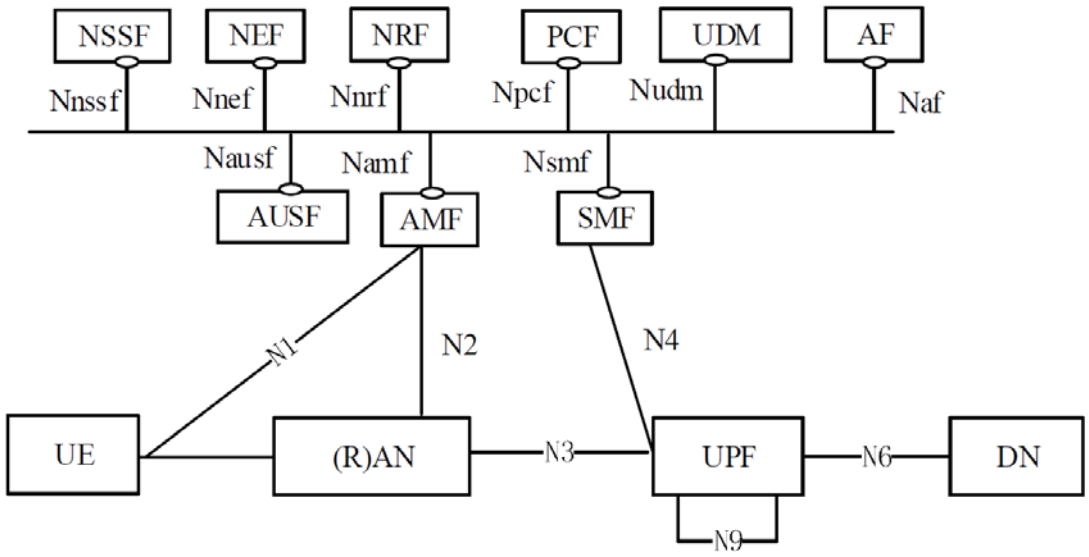


Fig. 2.4: Main Architecture of 5GC [18]

2.2 Purpose of Study

In the light of the issues in the previous section, excess accommodation in IP-based telecommunication networks causes QoS degradation for not only newly accommodated sessions, but also most of existing sessions. To appropriately guarantee the QoS of each communication session, determination of the maximum number of accommodatable sessions is the most important factor for both voice and data communications.

In the field of telecommunication networks, various call admission controls (CACs) have been proposed and operated to improve traffic accommodation under network congestion. The basic strategy of CAC is to prioritize some types of communication traffic and to reserve required communication resources e.g. lines, bandwidths, frequency bands, etc. and to guarantee the QoS of accommodated communication sessions for a certain period. In general CAC methods, the targets of control e.g. sessions, calls, packets etc. are judged to be accommodated or not, at the moment of their arrivals. If there are enough capacity for accommodation or the target has higher priority to terminate other communication resources which is being accommodated, the target is accepted and then accommodated into the communication system. This study defines that the acceptance and accommodation of incoming sessions are equivalent, and the term "accommodate" is used throughout this paper.

However, these issues have not been solved by conventional CAC methods. For example, to accommodate emergency calls in congested telephone networks during emergencies, several number of telephone lines are exclusively assigned for emergency calls. Although such trunk reservation control is effective for instantaneous congestion sudden after emergencies, reserving too many lines for a long period may cause a situation where a large number of general calls are blocked even though reserved lines are available. To accommodate more general calls and relieve general call users' stress and anxiety under emergencies, a new CAC method, which guarantees both required number of emergency calls and maximum communication time for accommodated general calls, is required.

Meanwhile, resource contention of the C-plane and the D-plane in virtualized mobile core network is also an unsolved issue. The increase in broadband communications by smartphone applications requires high data packet processing rate on the D-plane, while the signaling packet congestion of M2M communications on the C-plane will cause some fatal incidents in a future cyber networking environment. To accommodate incoming sessions and guarantee the QoS requirements of different communication types, the VM resource of both C/D-plane entities should be properly assigned, according to the current traffic condition.

In this thesis, two CAC methods are proposed for solving the issues in Sect. 2.1. The proposed methods aim at improving QoS and QoE of users, by solving the congestion

of both voice and data communication traffic. The goal of this study is to realize an efficient network resource management in a cost effective telecommunication network environment. For derivation of the proper communication resource assignment in the proposed methods, telecommunication network environment under congestion is theoretically modeled by using queueing theory. The proposed methods guarantee the efficient accommodation of communication traffic, and increase the number of accommodated sessions.

2.3 Research Overview

This thesis proposes two CAC methods to appropriately assign network resources for certain accommodation of congested communication traffic, by using queueing theory. The overview of each CAC method is described in this section.

Chapter 3 proposes the CAC for telecommunication system e.g. SIP server and telephone exchange. In Chapter 3, it is assumed that the holding time of general call users should be positively limited and much more general call users can communicate with their family, friends, etc., even in congested telecommunication network.

Moreover, Chapter 4 targets at mobile cellular networks and proposes a CAC method of VM resource assignment in a virtualized mobile core network. The proposed network model is theoretically based on the virtualized EPC architecture in 4G network, but also targets at the virtualized 5GC network architecture, and thus the proposed method can be applied to the 5G network. The purpose of the proposed CAC method in Chapter 4 is to accommodate both delay sensitive communications of M2M devices and broadband communications of smartphones and tablets.

2.3.1 Threshold Relaxation and Holding Time Limitation Method for Accommodating More General Calls under Emergency Trunk Reservation (Chapter 3)

As mentioned in Subsect. 2.1.1, telecommunication networks become congested due to large numbers of call requests during emergencies e.g. earthquake, hurricane, terrorism etc. Also, some infrastructure breaks down, so undamaged communication resources must be utilized more efficiently. Therefore, several line resources in telephone exchanges and SIP servers are generally reserved for emergency calls whose users communicate crucial information. The number of lines reserved for emergency calls is determined by a threshold, on a trunk reservation control method. To accommodate both required emergency calls and more general calls, the traffic intensity of arriving emergency calls should be estimated in advance, and a threshold should be configured so that the number of reserved lines becomes lower than the estimation. Moreover, this study proposes that the holding time for general calls should be positively limited. By guaranteeing the holding time sufficient for communicating essential information, holding time limitation reduces long-period calls so more general calls are accommodated.

Chapter 3 proposes a new CAC method to utilize undamaged communication resources more efficiently during emergencies. The proposed method accommodates more general calls by collaboratively relaxing the threshold of trunk reservation and limiting holding time of general calls. This method is targeted at not only the telephone exchange but also various systems on networks, e.g. base stations of the wireless network or SIP servers. With the proposed method, the threshold is configured in consideration of the ratio of traffic intensities estimated in advance.

In Chapter 3, the call session management equipments such as telephone exchanges and SIP servers are modeled as a queueing loss system and call-blocking rates of both emergency and general calls are calculated by using computer simulation. The comparison with the conventional holding time limitation method showed that the proposed method accommodates the required number of emergency calls by appropriately relaxing the threshold, while suppressing the increase in call-blocking of general calls.

2.3.2 vEPC Optimal Resource Assignment Method for Accommodating M2M Communications (Chapter 4)

Recently, the concept of a vEPC (Virtualized Evolved Packet Core) has been introduced as a framework for Network Functions Virtualization (NFV). In the vEPC technology, entities of the EPC (Evolved Packet Core), core networks of the 4G network, is deployed as a virtual machine in a IA server. Virtualizing each entity of the EPC enables much more flexible and cost-effective management of the mobile core network, in comparison with the conventional EPC network with physical equipments.

Moreover, the 5G mobile network environment has been studied and developed, and the demand of Machine-to-Machine (M2M) communications is also drastically increasing. M2M communications in 5G networks require much faster response than are possible in 4G networks. However, if both the control plane (C-plane) and the data plane (D-plane) functions of the EPC are migrated into a single vEPC server, M2M devices and other user equipments (UEs) share the same resources. To accommodate delay-sensitive M2M sessions in vEPC networks, not only signaling performance on the C-plane but also packet processing performance on the D-plane must be optimized.

Chapter 4 proposes a method for optimizing resource assignment of C-plane and D-plane Virtualized Network Functions (VNFs) in a vEPC server, called the vEPC-ORA method. We distinguish the communications of M2M devices and smartphones and model the vEPC server by using queueing theory. Numerical evaluations of optimal resource assignment shows that the proposed method minimizes the blocking rates of M2M sessions and smartphone sessions. We also confirmed that the mean packet processing time is kept within the allowable delay for each communication type, as long as the vEPC server has enough VM resources. Moreover, we study a resource granularity effect on the optimal resource assignment. Numerical evaluations under a fixed number of hardware resources of MME and S/P-GW is done for various resource granularities of the vEPC server. The evaluation results of numerical evaluations showed that the vEPC-ORA method derives the optimal resource assignment in practical calculation times.

2.4 Organization of this thesis

The rest of this thesis is organized as follows. Chapter 3 proposes the CAC method to appropriately configure the trunk reservation threshold in emergency situations. In Chapter 3, the telecommunication network in emergency situations and the behavior of both emergency calls and general calls are modeled by queueing method, and appropriate threshold is derived from estimated traffic intensity of both calls. Computer simulation of proposed method in various traffic situations shows the threshold configured by the proposed method accommodates both required emergency calls and more general calls, in comparison with the conventional method.

Chapter 4 describes the CAC method to assign the VM resource vEPC server. Numerical evaluations of the vEPC-ORA method shows the optimal VM resource assignment accommodates sessions of both the M2M communications and other smartphone communications, and processes their user data packets within their allowable delay. Finally, Chapter 5 concludes this thesis.

Chapter 3

Threshold Relaxation and Holding Time Limitation Method for Accommodating More General Calls under Emergency Trunk Reservation

3.1 Introduction

During emergencies such as earthquakes, hurricanes, or terrorist attacks, telecommunication networks become congested with calls requesting help, checking on family and friends, and so on. For example, the number of call attempts on cellular phones increased to about 10 times the usual amount in the disaster area just after both the Wenchuan Earthquake in 2008 [19] and the Great East Japan Earthquake in 2011 [20]. This congestion causes various problems. Moreover, some infrastructure such as telephone exchanges, base stations of mobile network, SIP servers and communication links break down during disasters. Thus, undamaged communication resources must be utilized more efficiently [21].

The ITU-T published recommendations for an Emergency Telecommunications Service (ETS) [16] against this congestion during emergencies. In some countries, several lines in telephone exchanges are generally reserved for emergency calls (i.e., calls made by such users as the police, emergency responders, road crews, and the news media) as a part of ETS. However, there has been little research on ETS. In addition, most research on congestion under emergency conditions has only targeted wireless networks, and there has been little research on congestion of general networks including wired telecommunication networks, base stations of wireless networks, and SIP servers on the Internet and NGN. Moreover, there are various scales of emergencies and there has been little research on methods that consider the scale of congestion. For example, a disaster such as the Great

East Japan Earthquake causes network congestion in a large area, while a train crash or terrorism causes congestion in a small area and the local network is heavily congested by a large number of call attempts.

On wired telecommunication networks, the number of lines to reserve for emergency calls can be determined by using a trunk reservation control method [22] [23]. With such a method, as shown in Fig. 3.1, an incoming emergency call is accommodated unless all the lines are being used, while an incoming general call is accommodated only if the number of lines being used is below the threshold. When this number is equal to or above the threshold, an incoming general call is rejected and then lost. The need to ensure that emergency calls are accommodated can lead to the threshold being configured more strictly than necessary. However, when incoming general calls greatly outnumber emergency calls, many general calls are rejected even though lines reserved for emergency calls are still free. To prevent this happening, the traffic intensity of arriving emergency calls should be estimated in advance, and a threshold should be configured so that the number of reserved lines becomes lower than the estimation, as long as the required number of emergency calls is accommodated without fail (this is defined as "relaxing the threshold" in this chapter).

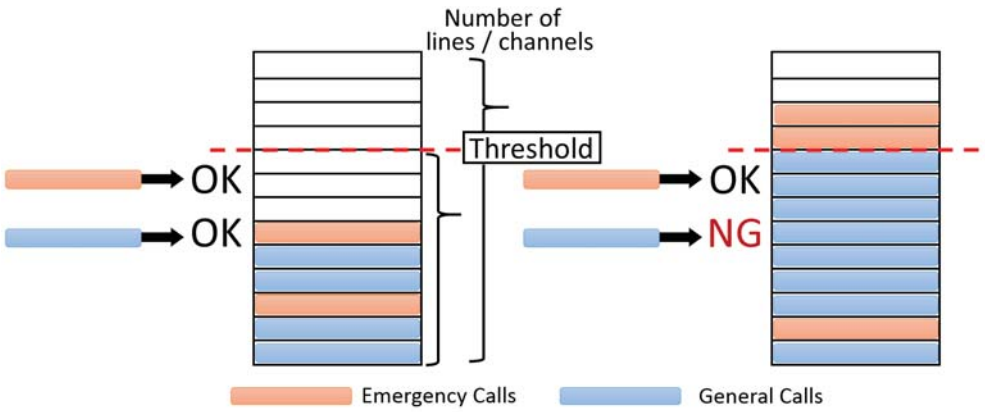


Fig. 3.1: Telephone exchange under trunk reservation control

Nowadays, both wired and mobile telecommunication networks are being migrated from the telephone exchange to an IP network [7] [10] [24]. In this IP network, although many users communicate with short transmission periods such as HTTP or FTP, some users occupy bandwidth by long realtime transmission periods such as VoIP. If the number of VoIP sessions increases, bandwidth is occupied during these sessions or the number of accommodated sessions exceeds the performance limit of SIP servers. This causes many

calls to be blocked. Furthermore, although VoIP sessions can be accommodated up to the limit of SIP servers, strained bandwidth on the transmission link causes more packet loss and transmission delay, even on accommodated sessions.

Another factor affecting congestion during emergencies is that some general telephone users may occupy bandwidth for long periods while chatting about trivial matters. Their long-period calls reduce the number of accommodated general calls. During emergencies, general telephone users should reduce their holding time as much as possible. So that more general calls are accommodated during emergencies, the holding time for general calls should be limited. By guaranteeing the holding time sufficient for communicating essential information, holding time limitation reduces long-period calls and enables more general calls to be accommodated. However, there has been little research on such approaches.

This chapter proposes a new CAC method to utilize undamaged communication resources more efficiently during emergencies. The proposed method accommodates more general calls by collaborative relaxing the threshold of trunk reservation control and limiting holding time of general calls. In this chapter, it is assumed that the holding time limit for minimum communication is given, and the threshold is relaxed by using estimated traffic parameters, including this holding time limit. This method is targeted at not only the telephone exchange but also various systems on networks, e.g. base stations of the wireless network or SIP servers. These systems are theoretically regarded as a telephone exchange on a wired telecommunication network. The proposed method distinguishes between emergency calls and general calls, and model a telephone exchange during an emergency statically as an $M_1, M_2/M, D/s/s, th$ loss system in order to investigate the appropriate threshold. The objective is to allow more general calls to be accommodated than in previous studies, while ensuring that the required number of emergency calls is still accommodated.

This chapter is organized as follows. Related studies are discussed and the purpose of this chapter is described in Sect. 2. Section 3 describes the key strategy of the proposed method. Section 4 describes the model setting of this study and then the proposed method of threshold relaxation and holding time limitation is explained. Section 5 evaluates the effect of the proposed method on call-blocking rates using computer simulation. Section 6 concludes this chapter.

3.2 Related Study

Some conventional methods have been proposed to reduce congestion of wireless network during emergencies and these methods can be categorized into “holding time limitation”

[25]–[26] and "call preemption" [27]–[29].

Okada proposed a method for limiting the holding time of cellular phone calls during emergencies [25]. In this method, the holding time limit depends on the current traffic conditions. For example, if the number of call requests increases to N times the normal amount, the time limit is varied so that the mean holding time is reduced to $1/N$. Computer simulation showed that the proposed method increased the number of accommodated calls above that in the situation where the holding time limit is fixed. In addition, Okada proposed a method for protecting emergency cellular phone calls by limiting only the holding time of general calls [26]. Emergency cellular phone calls are accommodated without fail by limiting only the holding time of general calls. Computer simulation showed that this method reduces the blocking rate considerably and then reduces the enforced call termination rate at handover without reducing the holding time of emergency calls. However, it does not reserve specific channels for protecting emergency calls. If the telecommunication network is congested with general calls, emergency calls may not be accommodated.

Zhou and Beard proposed methods for a cellular emergency network [27][28][29]. They assumed that each session used the same amount of wireless resources. When an incoming emergency session fails to find free channels, and if the number of active emergency sessions is less than the preemption threshold, the incoming emergency session will preempt a randomly picked ongoing general session. The preempted general session will be put into the handoff/preempted session queue. For an arriving public handoff session, it will also be buffered in the handoff/preempted session queue when no capacity is immediately available. If the number of active emergency sessions is higher than the threshold, there is no preemption. Numerical analysis showed that this method can guarantee a certain amount (75%) of channel resources for public (general) sessions. However, this method causes public sessions (general calls) to be terminated suddenly, which greatly stresses terminated callers and makes general callers call again, which causes more call-blocking. Moreover, mean holding times of both emergency and general sessions are assumed to follow the same exponential distribution.

3.3 Strategy

This section expresses the key idea of accommodating more general calls while accommodating required emergency calls.

First, the traffic intensities of both emergency calls and general calls are estimated considering traffic situation (e.g. number of telephone lines, arrival rates of both calls, mean holding time of emergency calls and holding time limit of general calls). In the

aftermath of such emergencies as earthquake or civil disturbances, the number of calls increases drastically and varies frequently. To accommodate required emergency calls, the demand of both types of calls must be estimated. For example, the arrival rates and the mean holding time are measured in a fixed time period, and traffic intensities are calculated from these values and updated as the traffic situation varies. Since some methods to estimate appropriate traffic intensities have already been proposed [30] [31], a traffic estimation method is not the target of this study. This chapter focuses that the traffic situation is stable in some periods so that the threshold can be configured appropriately in accordance with the estimated traffic intensities in advance.

Second, the *Guarantee Call-blocking Rate* strategy of threshold relaxation and general call holding time limitation is proposed [32]. The traffic intensity of arriving emergency calls is estimated, and the threshold is relaxed so that α times the estimation is reserved for emergency calls (α is defined as threshold relaxation rate, $0 < \alpha \leq 1$). This strategy also configures a general call holding time limit. The computer simulation in Subsect. 3.4.4 shows that much fewer reserved lines are needed for guaranteeing emergency calls when the *Guarantee Call-blocking Rate* strategy is operated.

Figure 3.2 shows why having a few reserved lines enables emergency calls to be accommodated. When the holding time limit of general calls is determined to be relatively short by using the *Guarantee Call-blocking Rate* strategy, the general calls leave the telecommunication network within a short period. Since the interval between the arrivals of emergency calls is longer than that between the leavings of general calls on average, a few reserved lines become allocatable when the next emergency call arrives at the telephone exchange.

Finally, the threshold is relaxed by using these estimated traffic intensities. Trunk reservation is needed to protect the required emergency calls from heavy congestion. However, reserving too many lines may block some general calls even though these reserved lines are still free. Since emergency calls can be accommodated into not only reserved lines but also other non-reserved lines, the number of reserved lines should be much lower than the estimated traffic intensity of emergency calls. Therefore, the threshold is relaxed in consideration of the ratio of estimated traffic intensities.

Figure 3.3 shows the reason for this strategy [33]. If the arrivals of emergency calls and general calls follow different Poisson distributions, the traffic intensities of both type of calls can be estimated from each arrival rate and mean holding time. When the number of lines being used is equal to or above the threshold, only an incoming emergency call is accommodated in reserved lines, and an incoming general call cannot be accommodated if all the non-reserved lines are occupied. When any non-reserved lines become free, which type of incoming call is accommodated (emergency or general) depends on the ratio of

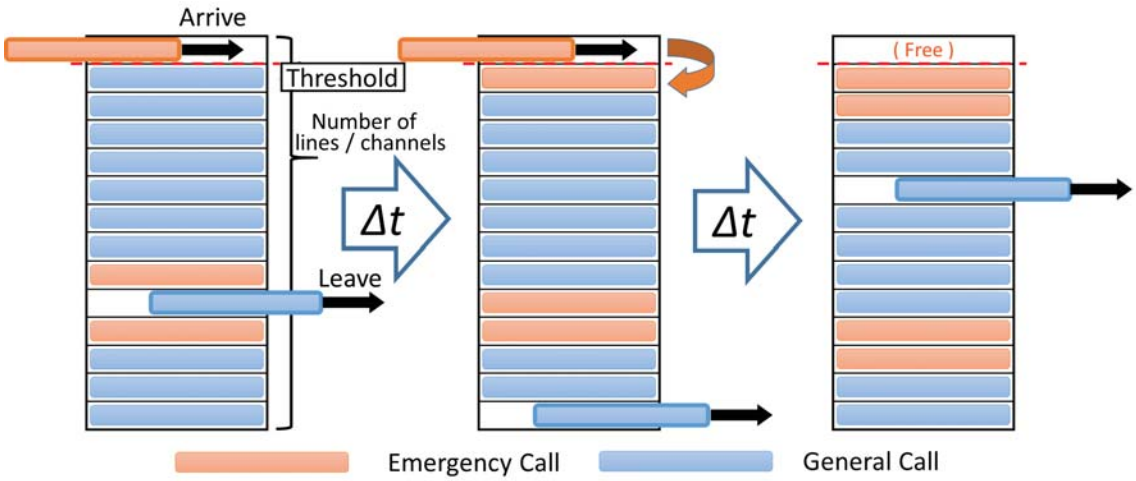


Fig. 3.2: Telephone exchange with a few reserved lines under heavy congestion

the arrival rates. Moreover, since the general call holding time is limited, general calls leave the telephone exchange within a shorter period than do emergency calls. Therefore, it is assumed that the ratio of traffic intensities in non-reserved lines is the same as that of estimated traffic intensities.

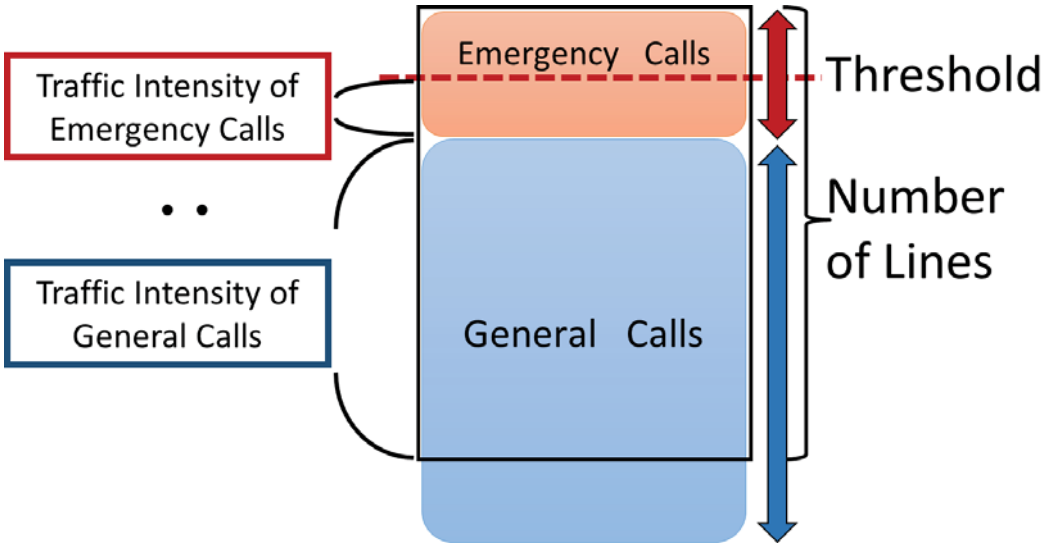


Fig. 3.3: Telephone exchange in a steady state

3.4 Threshold Relaxation Method

3.4.1 Holding Time Distribution of General Calls

In this chapter, various systems such as base stations of cellular network and SIP servers are modeled as a telephone exchange system. If the holding time limitation is operated and especially when the time limit is relatively short, holding time of general calls is expected not to follow an exponential distribution. When telecommunication networks are heavily congested and especially when the holding time limit h_g is relatively short, general callers become anxious about the next opportunity of talking. If their call request is accommodated, they will tend to talk until the time limit is reached, and the holding time distribution is supposed to be as shown in the left graph of Fig. 3.4.

Thus, it is assumed that all general callers talk for a fixed holding time and holding time of general calls follows a deterministic distribution, as shown in the right graph of Fig. 3.4.

Meanwhile, in the related studies i.e. Okada's [26], it is assumed that users' holding time follows an exponential distribution and it is truncated at the holding time limit h_g. Since this model is appropriate when the time limit is quite long, this truncated exponential distribution will be applied to this study for future works.

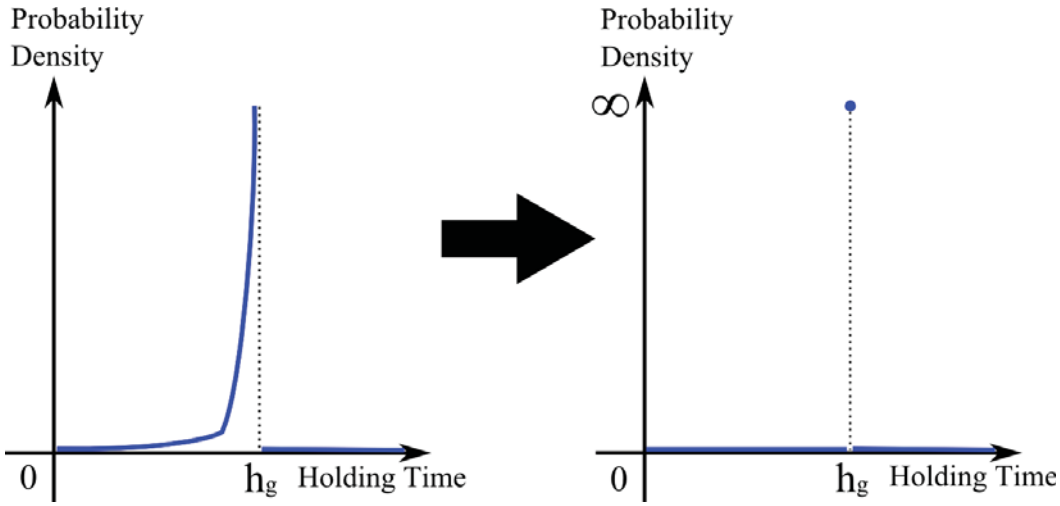


Fig. 3.4: Approximation of general call holding time distribution

3.4.2 Model Settings

Based on the assumption in Sect. 3.4.1, this telephone exchange during an emergency is modeled by using queueing theory, i.e., as an $M_1, M_2/M, D/s/s, th$ loss system (Fig. 3.5).

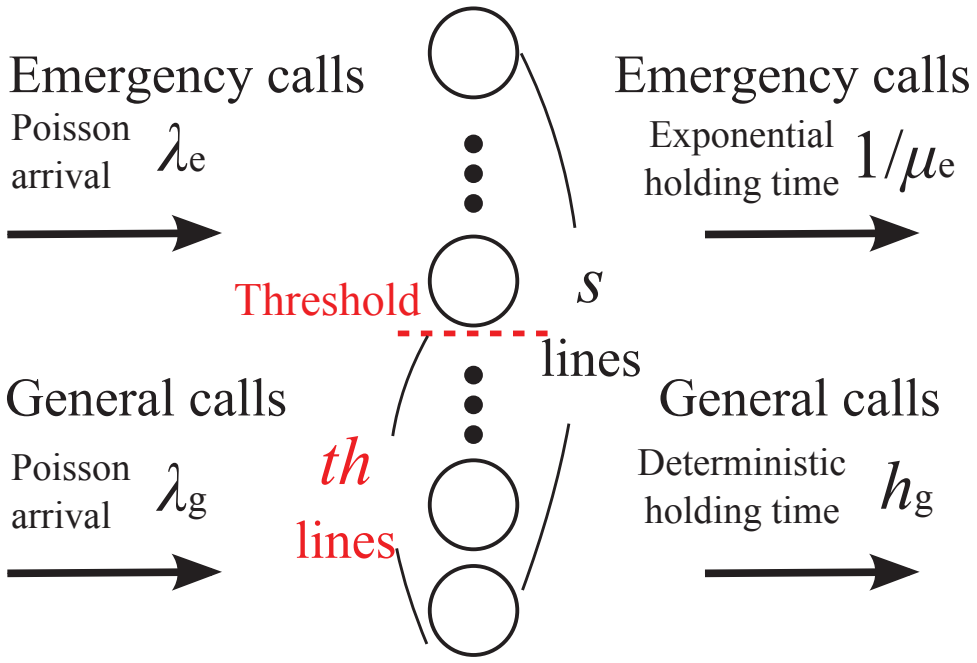


Fig. 3.5: Model settings

The flow chart of trunk reservation is shown in Fig. 3.6 and model settings are as follows:

1. Let s be the maximum number of telephone lines and th be the threshold for accommodating an incoming general call.
2. An incoming emergency call is accommodated unless all lines are being used. An incoming general call is accommodated only when the number of lines being used n is below th. When n is equal to or above th, an incoming general call is rejected.
3. The arrivals of emergency calls and general calls follow Poisson distributions with averages of λ_e and λ_g, respectively.
4. The holding times of emergency calls and general calls follow an exponential distribution with an average of $1/\mu_e$ and a deterministic distribution with constant value

h_g, respectively.

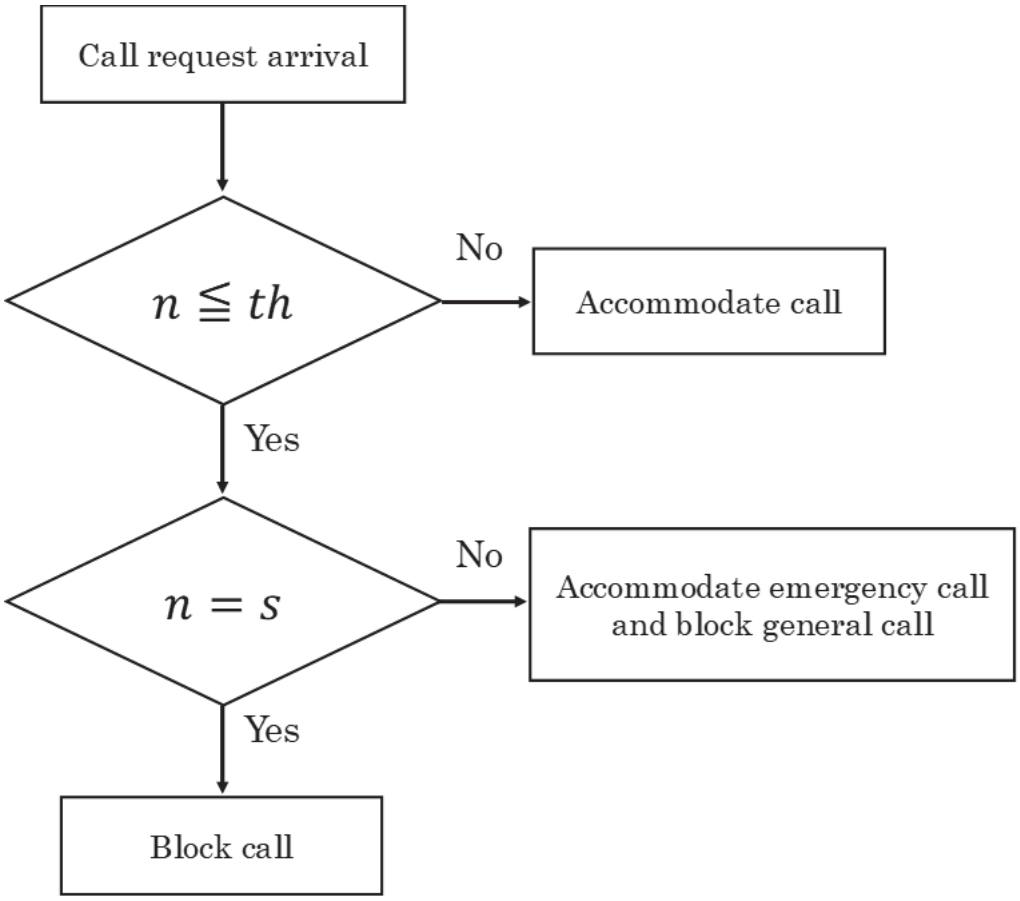


Fig. 3.6: Flow chart of trunk reservation

3.4.3 Traffic Intensity

Before the proposed method is explained, the relationship between traffic intensity and the call-blocking rate is shown in Figs. 3.7 and 3.8 [34]. Figures 3.7 and 3.8 are the results of computer simulation of trunk reservation and holding time limitation, and there are eight values of h_g and ten values of th for $s = 5000$ [lines], $\lambda_e = 300$ [calls/min], $\lambda_g = 9700$ [calls/min], and $1/\mu_e = 60$ [sec]. These two figures consist of the same calculation results, which are classified with th in Fig. 3.7 and with h_g in Fig. 3.8. The

traffic intensity of accommodated general calls a_g^* is calculated as

$$a_g^* = \frac{(\mathrm{NAGC} - \mathrm{NBGC}) \times h_g}{T} \ [\mathrm{erl}] \tag{3.1}$$

where NAGC is the number of arrived general calls, NBGC is the number of blocked general calls, and T is the simulation time. In this simulation, $T = 360,000$ [sec] (100 one-hour trials).

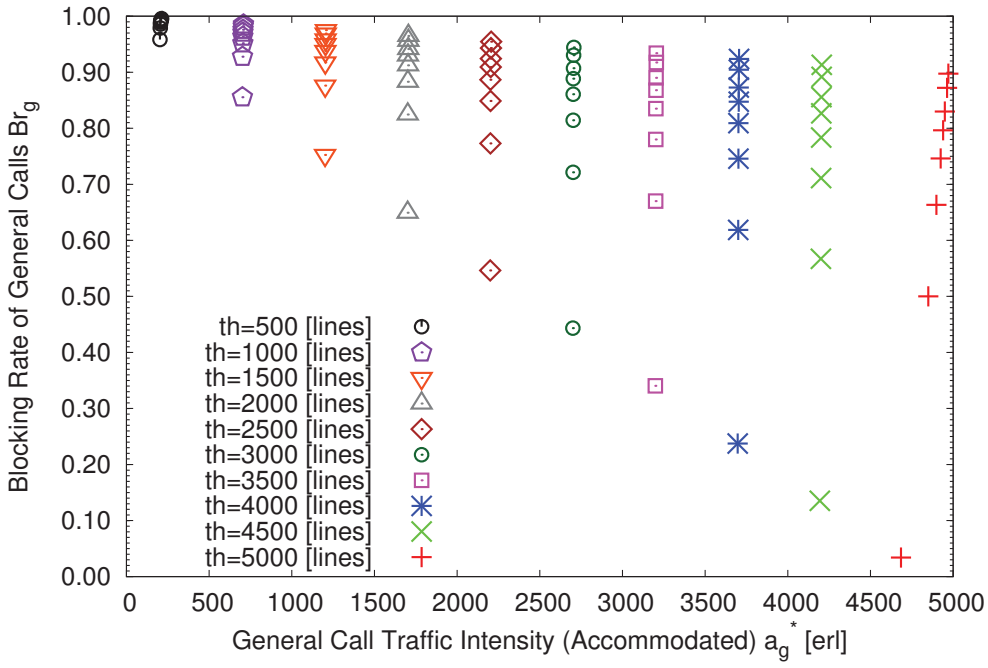


Fig. 3.7: Traffic intensity of accommodated general calls vs. general call-blocking rate (1)

Figure 3.7 shows that a_g^* was almost the same when th was the same, regardless of the value of h_g. Figure 3.8 shows that Br_g increased with h_g. These results are shown because non-reserved lines are always used owing to a telephone call rush. Because a_g^* increased substantially with h_g when trunk reservation was not applied ($th = s = 5000$), an upper limit on the traffic intensity of general calls can be determined by configuring a threshold. However, no methods for threshold configuration or holding time limitation have been proposed in [34].

According to this relationship, this chapter proposes a method to relax threshold th by using estimated traffic intensity of both emergency and general calls.

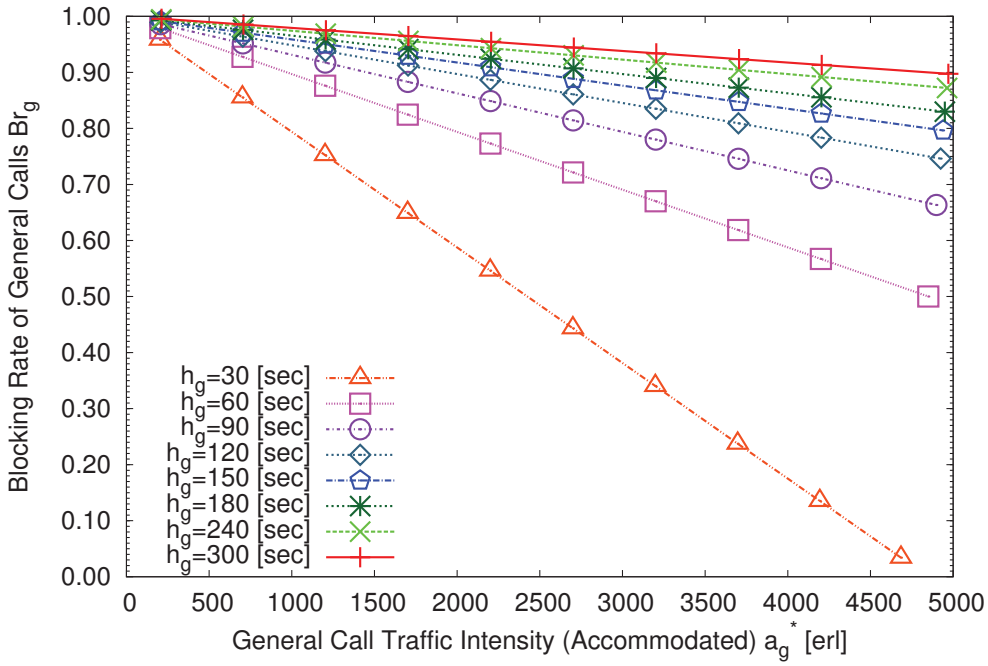


Fig. 3.8: Traffic intensity of accommodated general calls vs. general call-blocking rate (2)

3.4.4 Preliminary Experiment: *Guarantee Call-blocking Rate* Strategy

This subsection proposes a strategy of threshold relaxation and holding time limit configuration (*Guarantee Call-blocking Rate* Strategy) [32]. This strategy configures the value of threshold th and holding time limit of general calls h_g by the number of lines s, arrival rates λ_e and λ_g, and mean holding time of emergency calls $1/\mu_e$. The algorithm of the method is as follows:

(a) Estimate the traffic intensity of emergency calls a_e by using Eq. (3.2).

$$a_e = \lambda_e \times 1/\mu_e = \frac{\lambda_e}{\mu_e} \tag{3.2}$$

(b) Configure threshold relaxation rate α $(0 < \alpha \leq 1)$.

(c) Configure threshold th so that $\lceil \alpha \times a_e \rceil$ of telephone lines are reserved for emergency calls.

$$th = \lceil s - \alpha a_e \rceil \tag{3.3}$$

(d) Considering that the traffic intensity of general calls a_g would be calculated by using Eq. (3.4), configure h_g so that $a_g = th$.

$$a_g = \lambda_g h_g \tag{3.4}$$

$$h_g \left(= \frac{a_g}{\lambda_g} \right) = \frac{th}{\lambda_g} = \frac{\lceil s - \alpha a_e \rceil}{\lambda_g} \tag{3.5}$$

With this strategy, the holding time of general calls is limited in accordance with the traffic intensity of general calls a_g.

In this subsection, the effect of *Guarantee Call-blocking Rate* Strategy on call-blocking rates is evaluated for both emergency and general calls. The call-blocking rates of both emergency and general calls are calculated for various traffic situations by computer simulation with the $M_1, M_2/M, D/s/s, th$ loss model.

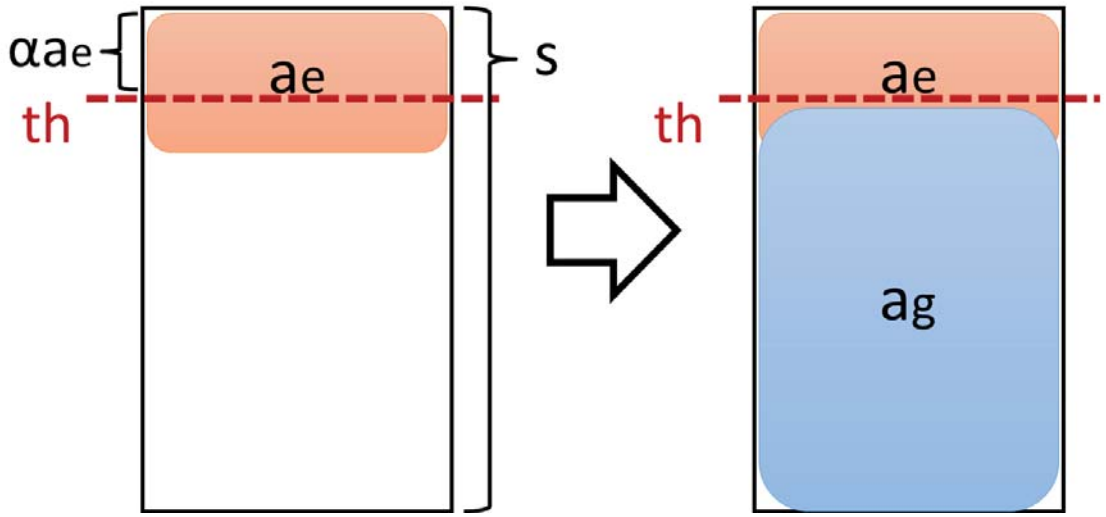


Fig. 3.9: *Guarantee Call-blocking Rate* strategy

Figure 3.10 shows the relationship between the number of telephone lines s and the emergency call-blocking rate Br_e for $\lambda_e = 300$ [calls/min], $\lambda_g = 9700$ [calls/min], and $1/\mu_e = 60$ [sec]. Figure 3.11 shows the relationship between s and the general call-blocking rate Br_g in the same traffic situations as Fig. 3.10. The value of h_g is determined by using the *Guarantee Call-blocking Rate* strategy, with six values of α. The arrival ratio ($\lambda_e : \lambda_g = 3 : 97$) is the same as that used by Okada [26]. The total simulation time is 360,000 [sec] (100 one-hour trials). The results of *Guarantee Call-blocking Rate* Strategy

are compared with those obtained without using methods or trunk reservation, in which only the holding time limitation is operated.

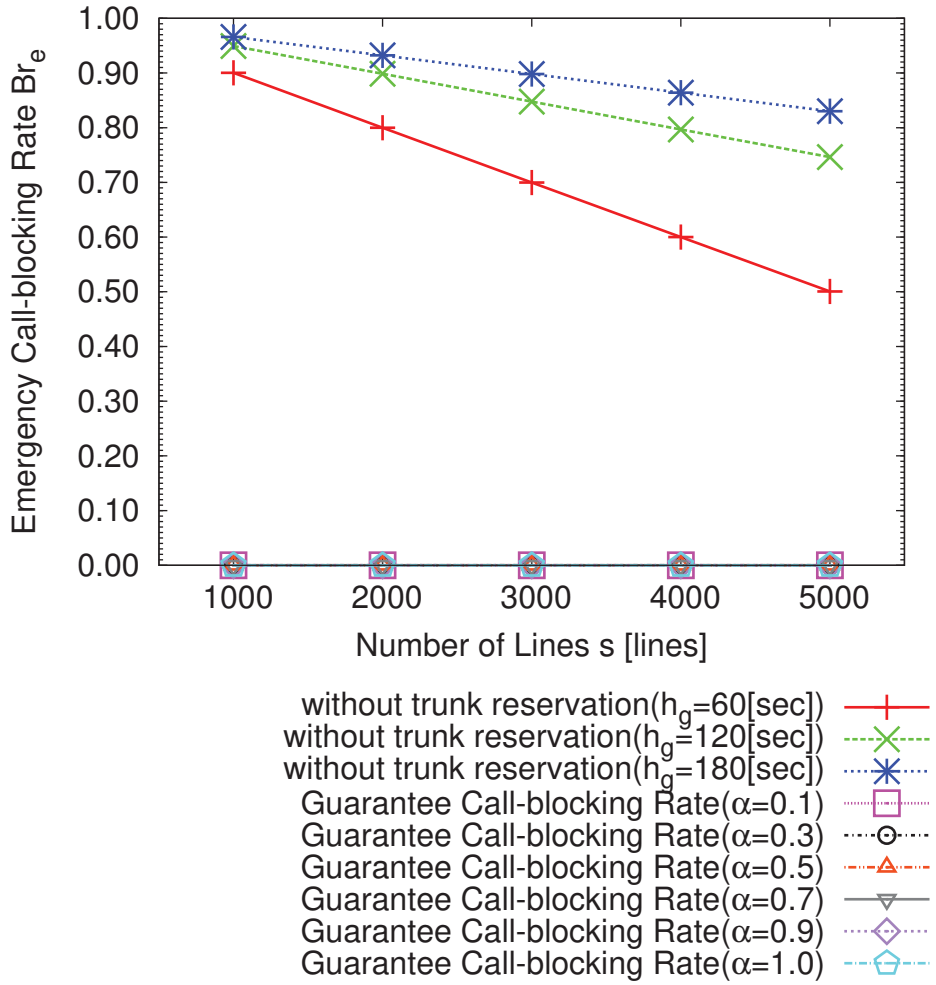


Fig. 3.10: Effect of *Guarantee Call-blocking Rate* strategy on emergency call-blocking rate

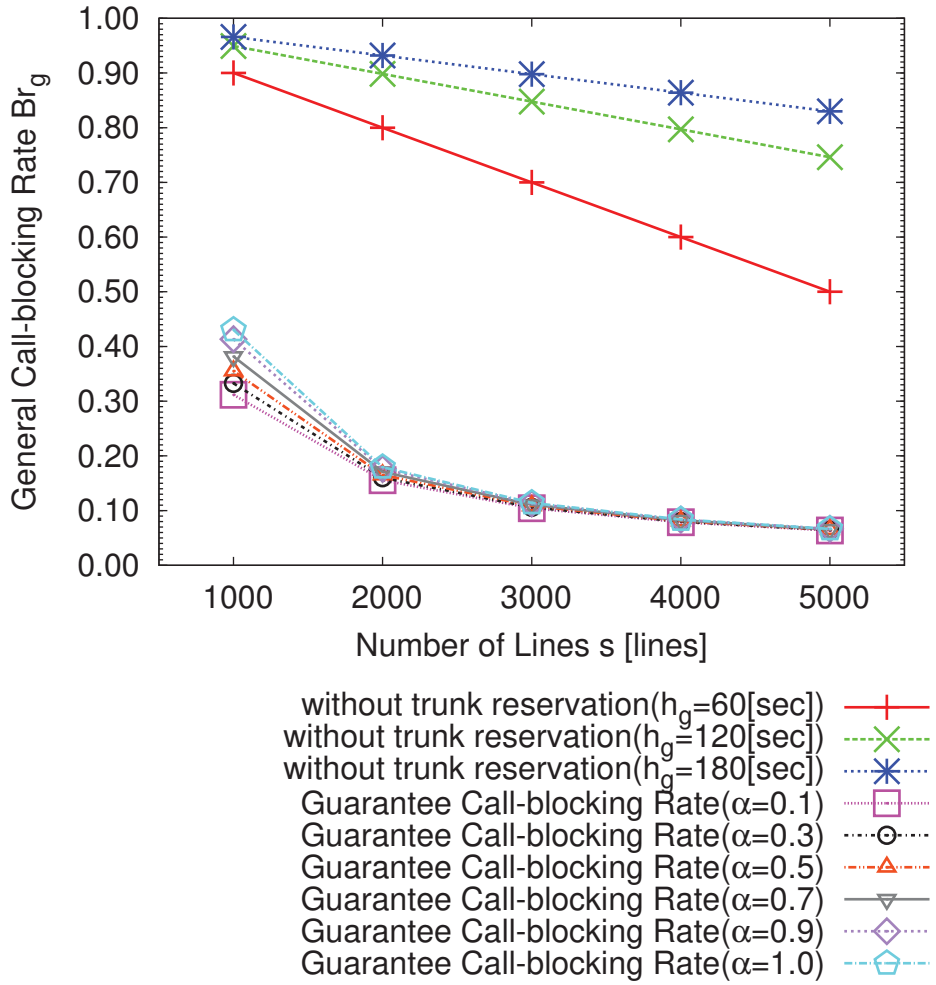


Fig. 3.11: Effect of *Guarantee Call-blocking Rate* strategy on general call-blocking rate

As shown in Fig. 3.10, *Guarantee Call-blocking Rate* Strategy can reduce Br_e to almost zero by threshold relaxation for each scale of the telephone exchange. Figure 3.11 shows that Br_g also can be reduced, compared with the results obtained without using trunk reservation. These two figures show that *Guarantee Call-blocking Rate* Strategy enables more general calls to be accommodated than by just limiting their holding time, while the required number of emergency calls is still accommodated.

The objective of the *Guarantee Call-blocking Rate* strategy is to share available telephone lines with more general callers and reduce the general call-blocking rate Br_g to almost zero. However, when a large number of general calls arrives, the value of h_g becomes too short, as shown in Fig. 3.12 (when $\lambda_e = 300$ [calls/min], $\lambda_g = 9700$ [calls/min], $1/\mu_e = 60$ [sec], and $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9, 1.0$).

Although general callers should give essential information quickly, this is almost im-

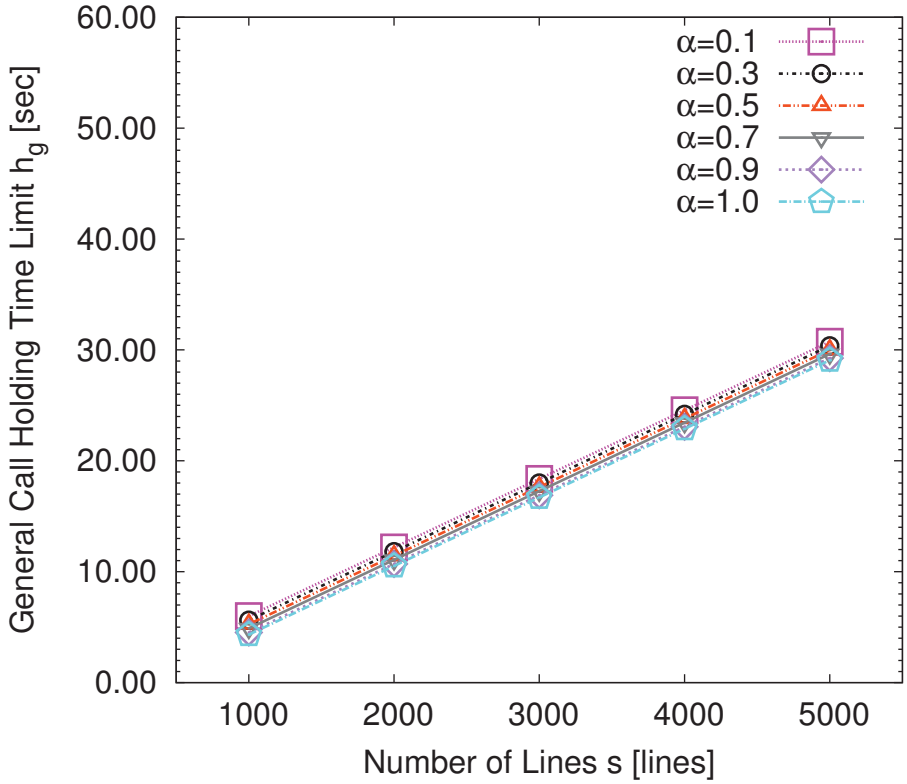


Fig. 3.12: The value of holding time limit h_g configured by *Guarantee Call-blocking Rate* strategy

possible when the time limit is below 10 seconds. Because the holding time limit should not vary frequently for relieving users during emergencies, the proposed method does not automatically configure the holding time limit by the threshold. Instead of the automatic holding time configuration, next subsection proposes a threshold relaxation method in the next subsection. The proposed method automatically configures the threshold by estimating traffic intensities of arriving calls, with an assumption that the holding time limit required for minimum communication is given.

3.4.5 Proposed Method

Under heavy congestion during emergencies, almost all telephone lines in a telephone exchange are expected to be in use. Based on the result of the preliminary experiment, a new method for configuring the values of th is proposed, so that the resources in a telephone exchange are fully used.

With the proposed method, the traffic intensity of both incoming emergency calls and general calls is estimated in advance and used for configuring a threshold. If the number

of reserved lines is directly configured by the estimated traffic intensity of emergency calls, almost all emergency calls will be accommodated. However, since telephone lines below the threshold are available for not only general calls but also emergency calls, there may be some free reserved lines. This is wasteful from the point of view of resource utility and causes more general calls to be blocked. Therefore, this method of relaxing the threshold and increasing the number of non-reserved lines is proposed.

(a) It is assumed that the value of general call holding time limit h_g, required for minimum communication in emergencies, is given in advance.

(b) Estimate the traffic intensity of emergency calls a_e by using the emergency call arrival rate λ_e and mean holding time of emergency calls $1/\mu_e$ in Eq. (3.2). The traffic intensity of general calls a_g is also estimated by using the general call arrival rate λ_g and the holding time limit h_g in Eq. (3.4).

(c) Configure threshold th as:

$$th = \begin{cases} 0 & \text{when } s < a_e \\ \left\lceil \dfrac{(s-a_e)(a_e+a_g)}{a_g} \right\rceil & \text{when } a_e \le s \le a_e + a_g \\ s & \text{when } s > a_e + a_g \end{cases} \tag{3.6}$$

As mentioned in Sect. 3.3, Fig. 3.3 shows the reason Eq. (3.6) is used for configuring a threshold. The value of th is thus determined by solving the following equation:

$$s - a_e = th \times \frac{a_g}{a_e + a_g} \tag{3.7}$$

Meanwhile, $s < a_e$ means that the traffic intensity of arriving emergency calls is more than the number of lines in telephone exchange. Although all the lines are reserved, not all emergency calls can be accommodated in this situation. Therefore, th is configured to be 0. $s > a_e + a_g$ means that the total traffic intensity is less than the number of lines and the system is not congested. Trunk reservation is no longer needed in this situation. Then, th is configured to be s.

In the proposed method, value of threshold relaxation rate α in the preliminary experiment can be calculated by Eq. (3.8).

$$\alpha = \frac{(s - th)\mu_e}{\lambda_e} \tag{3.8}$$

The proposed method is based on the author's previous study [33]. In the previous method, target general call-blocking rate Br_g^* is given, and the traffic intensity of general

calls a_g is estimated by Eq. (3.9) so that the $1 - Br_g^*$ of general calls is accommodated. Then, a_e of Eq. (3.2) and a_g is substituted for Eq. (3.9), and threshold th is configured.

$$a_g = (1 - Br_g^*)\lambda_g h_g \tag{3.9}$$

$$th = \min\left(\left\lceil \frac{(s - a_e)(a_e + a_g)}{a_g} \right\rceil, s\right) \tag{3.10}$$

However, emergency call-blocking rate Br_e increases as Br_g^* increases. This is because the estimated traffic intensity of general calls a_g decreases in Eq. (3.9), by abandoning some general calls. Threshold th is then configured so that less lines are reserved, compared with the th of $Br_g^* = 0$. Figure 3.13 shows the relationship between Br_g^* and call-blocking rates Br_e, Br_g for $s = 5000$ [lines], $\lambda_e = 300$ and $\lambda_g = 9700$ [calls/min], $1/\mu_e = 60$ [sec] and $h_g = 40$ [sec]. From Fig. 3.13, some emergency calls are blocked when $Br_g^* \geq 0.3$. This is because the threshold is configured as $th = s$ by Eq. (3.10).

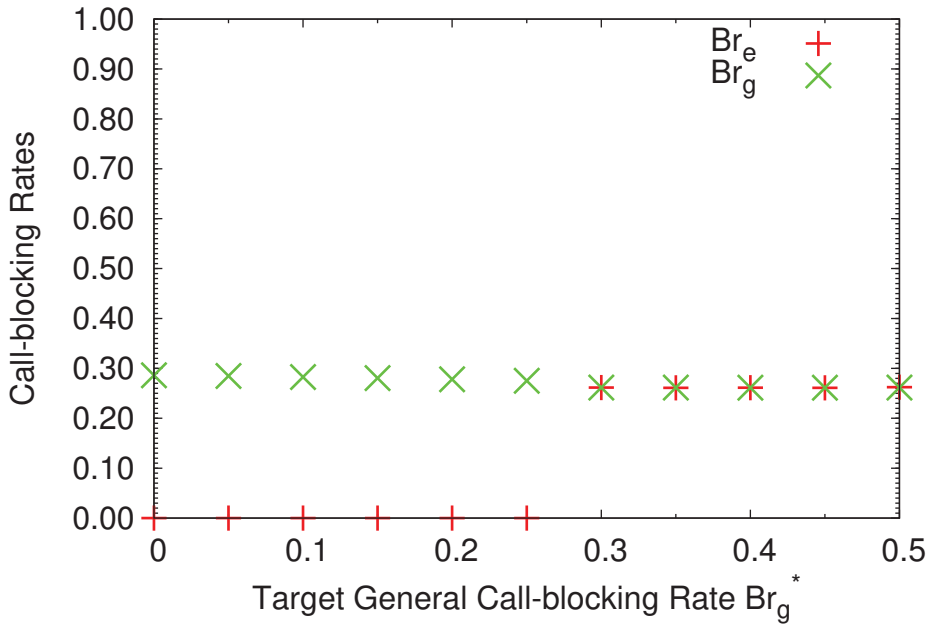


Fig. 3.13: Effect of target general call-blocking rate Br_g^* on call-blocking rates ([33])

Although the appropriate value of Br_g^* must be manually configured in accordance with the traffic condition in [33] , abandoning arriving general calls too much may lead to a lack of reserved lines. This chapter does not use Br_g^* in the proposed method.

3.5 Evaluation of the Proposed Method on Call-blocking Rates

3.5.1 Evaluation of Proposed Method: Small Systems Cases

As mentioned before, Okada proposed a holding time limitation method for protecting emergency calls in wireless cellular networks [26]. However, it does not reserve specific channels for emergency calls. When the telecommunication network is congested due to a large number of general calls, emergency calls may not be accommodated even if the holding time of general calls is strictly limited. This section shows the call-blocking reduction effect of the proposed method by applying the holding time limit configured by Okada's conventional method to the proposed method.

In Okada's work [26], the holding time limit of general calls is configured as follows. Arrivals of both emergency calls and general calls follow the same Poisson distribution and their holding time follows the same exponential distribution. Mean holding time in a normal situation is $1/\mu$ and the holding time limit of general calls is h_g. When the mean holding time of general calls under holding time limitation is h_{ave}, h_{ave} follows Eq. (3.11).

$$\begin{aligned} h_{ave} &= \int_0^{h_g} x\mu e^{-\mu x}dx + \int_{h_g}^{\infty} h_g\mu e^{-\mu x}dx \\ &= \frac{1}{\mu}(1 - e^{-\mu h_g}) \end{aligned} \tag{3.11}$$

From Eq. (3.11), these formulas are obtained:

$$e^{-\mu h_g} = 1 - \mu h_{ave} \tag{3.12}$$

$$h_g = -\frac{1}{\mu}\log(1 - \mu h_{ave}) \tag{3.13}$$

The holding time limit of general calls h_g is configured by using $1/\mu$ and h_{ave} in [26].

Figure 3.14 shows the relationship between holding time limit h_g and emergency call-blocking rate Br_e. In the evaluation of conventional method, only holding time limitation is operated. This subsection assumes a small system with $s = 160$. $1/\mu_e$ is given three values: 60, 120, and 180 [sec]. In the conventional method, the mean holding time under limitation h_{ave} must be shorter than $1/\mu_e$ in order to calculate Eq. (3.13). Therefore, h_{ave} is given two values (20 and 40 [sec]) when $1/\mu_e = 60$ [sec] and three values (20, 40, and 60 [sec]) when $1/\mu_e = 120, 180$ [sec]. The total simulation time is 360,000 [sec]. Figure 3.15 shows the relationship between h_g and general call-blocking rate Br_g in the same traffic situations. Arrival rates λ_e and λ_g are configured so that $\lambda_e : \lambda_g = 3 : 97$ and

total traffic intensity equals 448[erl], as an assumption of five times the normal traffic load [26]. For example, when $1/\mu_e$=120 [sec] and h_{ave}=20 [sec], it is obtained that $\lambda_e = 35.06$ [calls/min] and $\lambda_g = 1133.63$ [calls/min] by solving these simultaneous equations:

$$\begin{cases} \lambda_e \cdot \dfrac{120}{60} + \lambda_g \cdot \dfrac{20}{60} = 448 \\ \lambda_e : \lambda_g = 3 : 97 \end{cases} \tag{3.14}$$

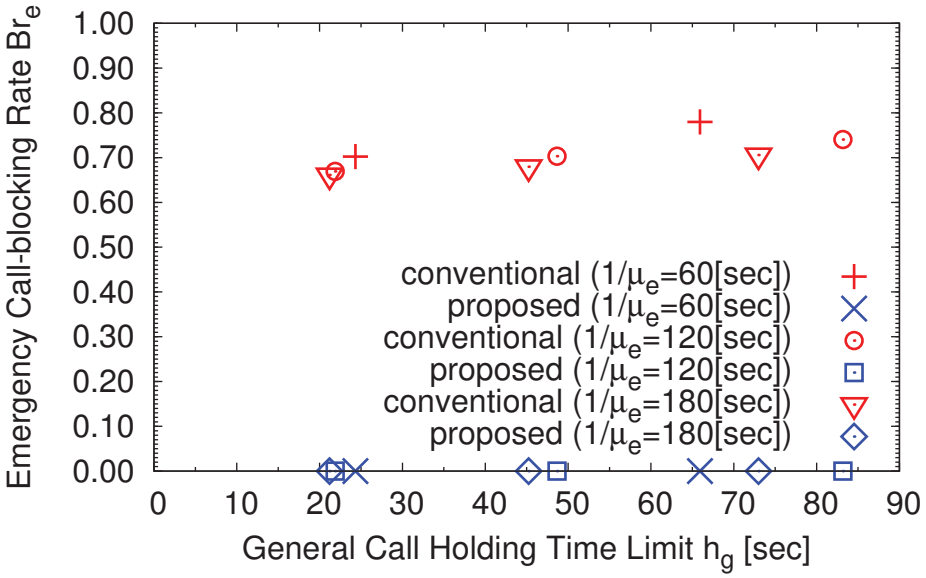


Fig. 3.14: Emergency call-blocking rate ($s = 160$)

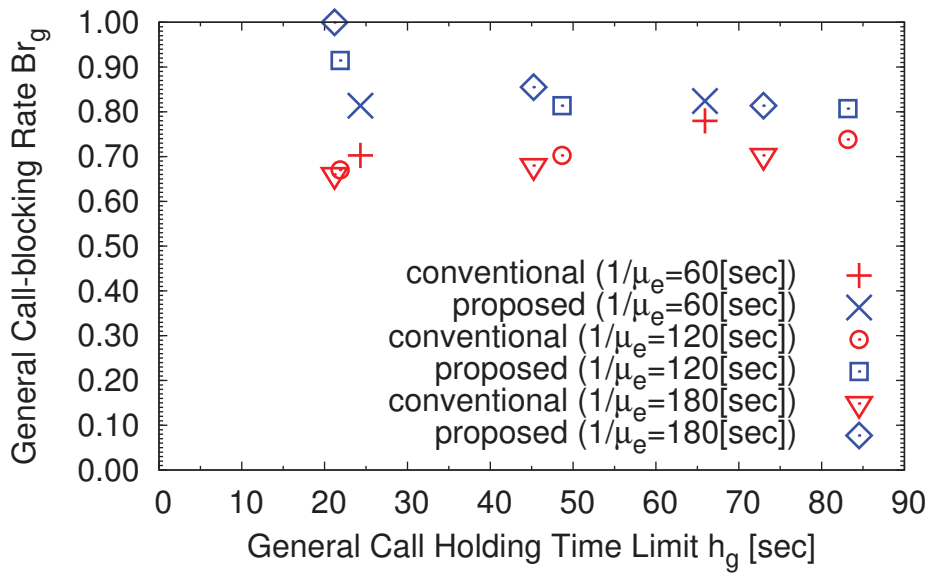


Fig. 3.15: General call-blocking rate ($s = 160$)

Figure 3.14 shows that proposed method reduces call-blocking of emergency calls to almost zero, while about half of emergency calls are blocked and the required number of emergency calls is not accommodated in the conventional method. On the other hand, Fig. 3.15 shows that proposed method blocks more general calls than conventional method by accommodating required emergency calls. Br_g increases, especially when $1/\mu_e$ is long. From these results, emergency call users also should terminate their calls quickly, as long as they can communicate essential information.

3.5.2 Evaluation of Proposed Method: Large Systems Cases

3.5.2.1 Parameter Settings

This subsection assumes large systems with thousands of channels. Call-blocking rates of both emergency and general calls are calculated for various traffic situations. Parameter settings are as listed in the following table:

Table 3.1: Parameter settings

s	1000, 2000, 3000, 4000, 5000 [lines]
$\lambda_e + \lambda_g$	10000 [calls/min]
$\lambda_e : \lambda_g$	1:99, 3:97, 5:95, 10:90, 15:85, 20:80, 25:75
$1/\mu_e$	60, 120, 180 [sec]
h_g	5, 10, ..., 85, 90, 100, ..., 170, 180 [sec]
Simulation time	360000 [sec]

3.5.2.2 Comparison from Conventional Method

First, the call-blocking reduction effect of the proposed method is explained, according to the parameter settings listed in 3.5.2.1. The holding time limit configured by Okada's conventional method is applied to the proposed method.

Figure 3.16 shows the relationship between holding time limit h_g and emergency call-blocking rate Br_e for $s = 5000$ [lines], $\lambda_e = 300$ and $\lambda_g = 9700$ [calls/min]. Figure 3.17 shows the relationship between h_g and general call-blocking rate Br_g under the same traffic situations. $1/\mu_e$ is given three values: 60, 120, and 180 [sec]. h_{ave} is given two values (20 and 40 [sec]) when $1/\mu_e = 60$ [sec] and five values (20, 40, 60, 80, 100 [sec]) when $1/\mu_e = 120, 180$ [sec]. The ratio of arrival rates $\lambda_e : \lambda_g = 3 : 97$ is the same as the ratio used in [26].

Figure 3.16 shows that Br_e is also reduced to almost zero by the proposed method, while Br_e increases as holding time limit h_g increases, in the conventional method. Moreover, Fig. 3.17 shows that the increase in Br_g of the proposed method is smaller than that of the conventional method. From these two figures, the proposed method accommodates the required number of emergency calls by appropriately relaxing the threshold, while suppressing the increase in call-blocking of general calls.

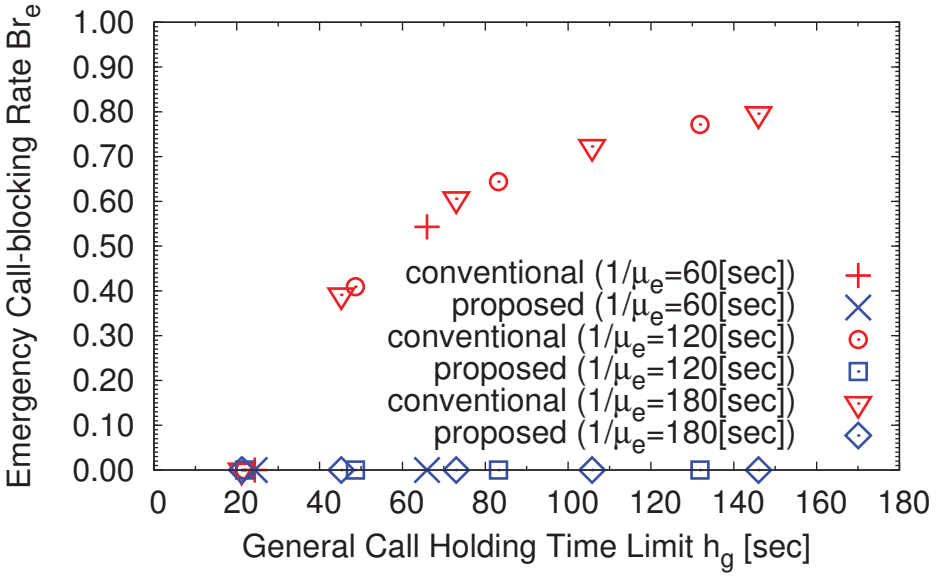


Fig. 3.16: Emergency call-blocking rate ($s = 5000$)

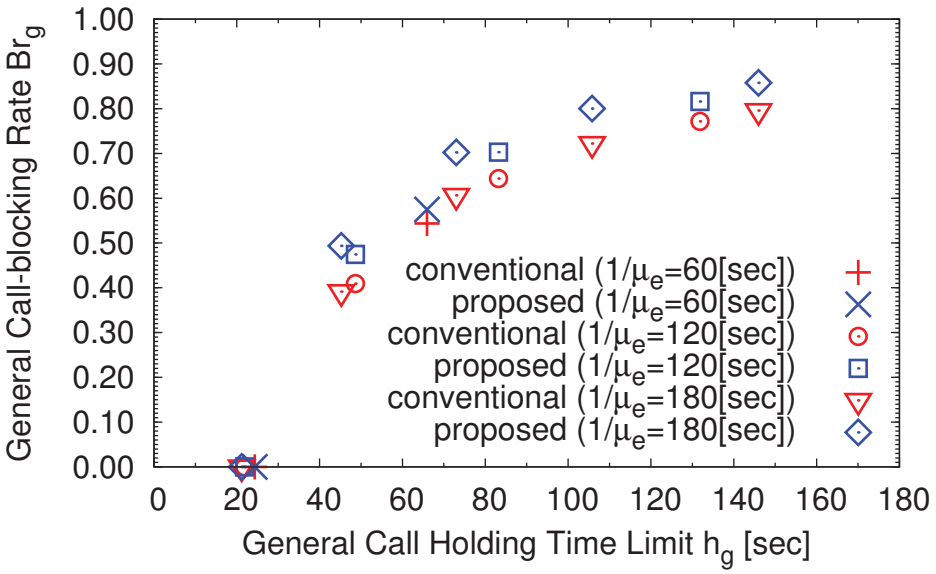


Fig. 3.17: General call-blocking rate ($s = 5000$)

Furthermore, the call-blocking reduction effect is shown in comparison with the results of redundant trunk reservation. If the threshold is configured so that the number of reserved lines is equal to the estimated traffic intensity of emergency calls, the emergency calls can be securely accommodated. However, such redundant reserved lines cause more

call-blockings of general calls and the threshold should be relaxed while the required number of emergency calls are accommodated.

Figure 3.18 shows the relationship between the call-blocking rates of the proposed method and the fixed threshold method for $s = 5000$ [lines], $\lambda_e = 300$, $\lambda_g = 9700$ [calls/min] and $1/\mu_e = 60$ [sec]. Since the traffic intensity of emergency calls is calculated as $a_e = 300$, the threshold is configured as $th = s - a_e = 4700$ [lines] in the fixed threshold method. The results of Fig. 3.18 show that Br_g of the proposed method is reduced from Br_g of the fixed threshold method, while Br_e is kept to almost zero in both methods. For example, Br_e is reduced by 4.7% at $h_g = 50$ [sec], 10.5% at $h_g = 40$ [sec] and 59.9% at $h_g = 30$ [sec]. Figure 3.18 shows that the threshold relaxation of the proposed method can accommodate more general calls, while the accommodation of emergency calls are guaranteed.

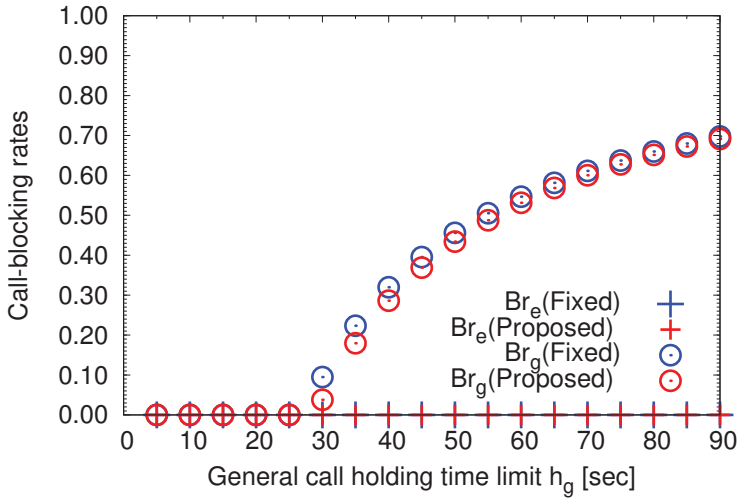


Fig. 3.18: Comparison between proposed method and fixed threshold method

3.5.2.3 Number of Lines

Second, Fig. 3.19 shows the relationship between h_g and Br_e for $\lambda_e = 300$, $\lambda_g = 9700$ [calls/min], and $1/\mu_e = 60$ [sec]. Figure 3.20 shows the relationship between h_g and Br_g under the same traffic situations. Plots are classified with the number of lines s, and there are five values.

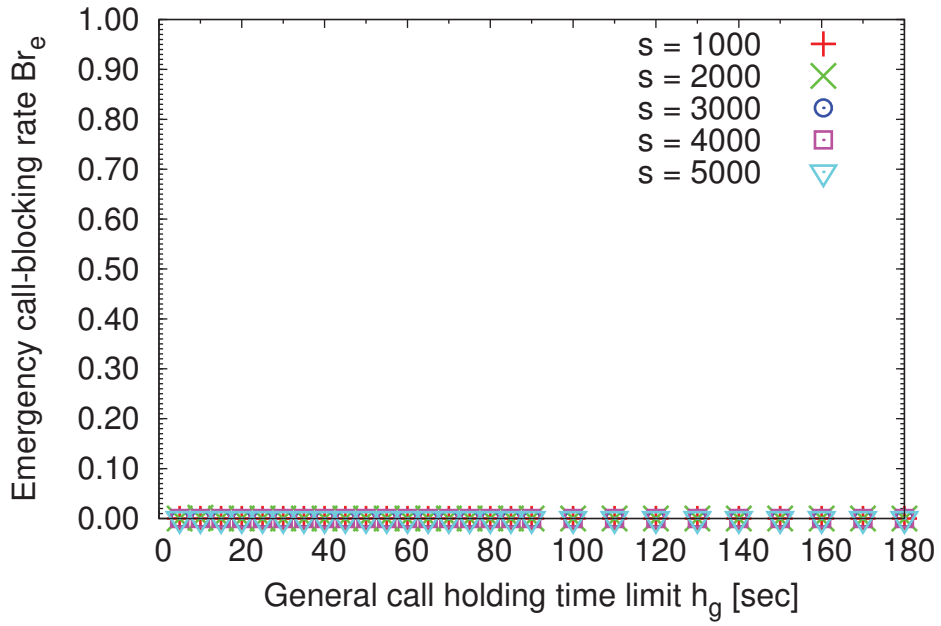


Fig. 3.19: Relationship between h_g and emergency call-blocking rate (five values of s)

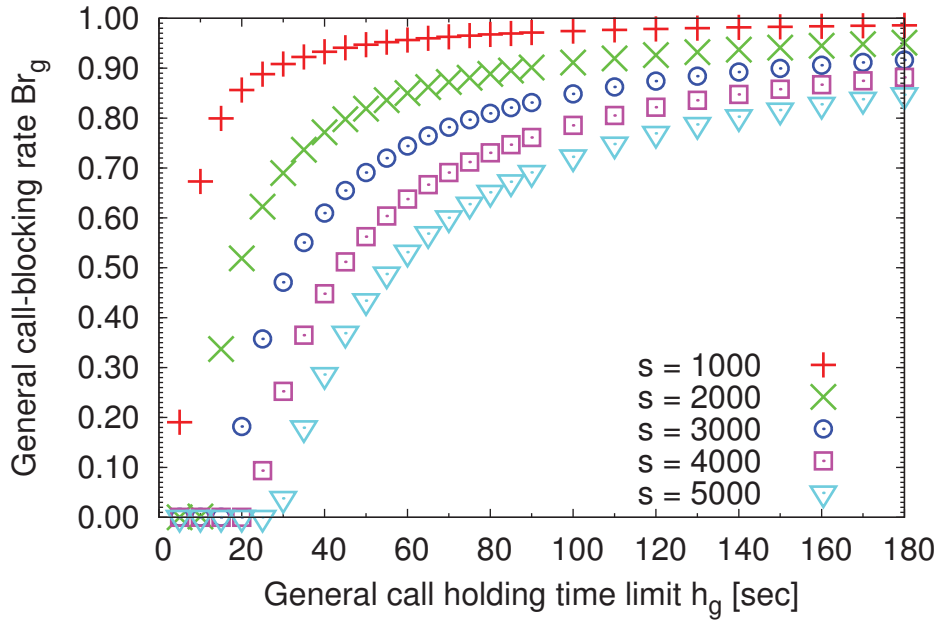


Fig. 3.20: Relationship between h_g and emergency call-blocking rate (five values of s)

As shown in Fig. 3.19, almost all emergency calls are accommodated for each number of lines s, under these traffic situations. The proposed method accommodates the required number of emergency calls unless traffic intensity of emergency calls a_e exceeds the capacity of the systems. Meanwhile, Br_g drastically decreases when h_g is well limited. Moreover, the reduction effect on general call-blocking rate becomes more significant as the number of lines s becomes small. In other words, holding time limitation is effective especially when the scale of the system is smaller than the traffic demand. The proposed method utilizes limited communication resources more efficiently.

3.5.2.4 Arrival Ratio

Finally, the relationship between the effect of the proposed method and the ratio of arrival rates is investigated. Figure 3.21 shows the relationship between h_g and Br_e for $s = 5000$ [lines], $\lambda_e + \lambda_g = 10000$ [calls/min] and $1/\mu_e = 60$ [sec]. Plots are classified with the arrival ratio $\lambda_e : \lambda_g$, and there are seven values. Figure 3.22 shows the relationship between h_g and Br_e under the same traffic situation.

As shown in Fig. 3.21, Br_e was reduced to almost zero for each arrival ratio $\lambda_e : \lambda_g$. Figure 3.22 shows that Br_g increased as a larger number of emergency calls arrived against the number of general calls. However, Br_g also reduced drastically due to the holding time limitation. Therefore, the holding time limitation is more effective when many emergency calls arrive, e.g. just after an earthquake or terrorist attack.

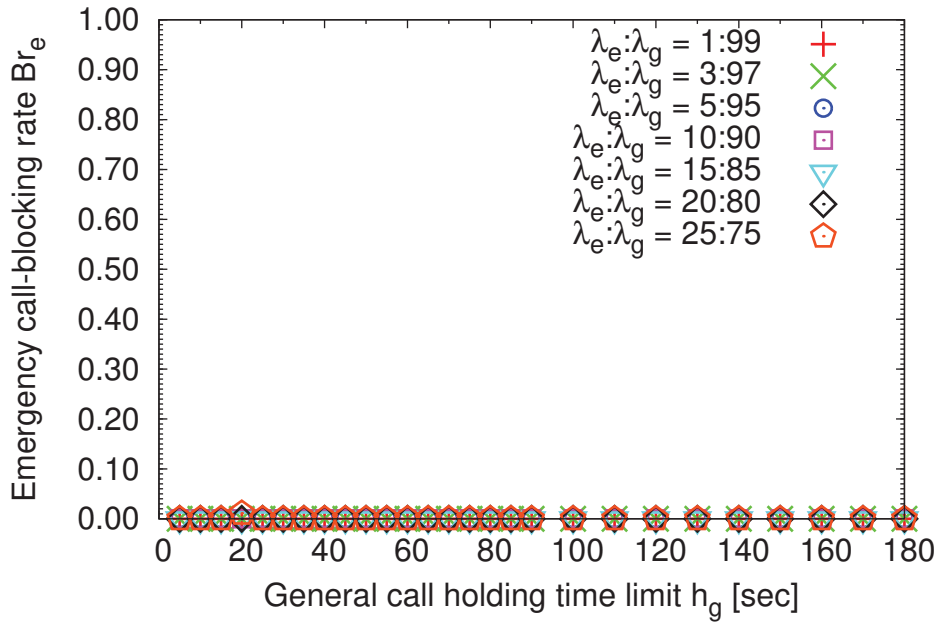


Fig. 3.21: Relationship between h_g and emergency call-blocking rate ($\lambda_e + \lambda_g = 10000$[calls/min])

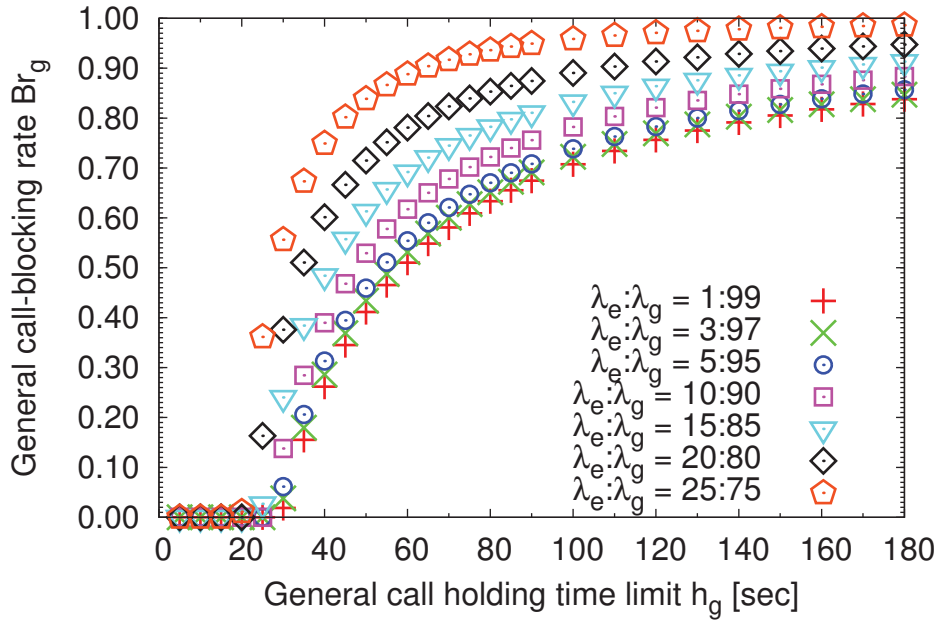


Fig. 3.22: Relationship between h_g and emergency call-blocking rate ($\lambda_e + \lambda_g = 10000$[calls/min])

3.5.3 Discussion

In this section, the call-blocking reduction effect of the proposed method is investigated, and compared with that of a conventional method [26]. Two scenarios are considered: small systems cases e.g. SIP servers in a rural area or small wireless base stations, and large systems cases e.g. SIP servers in the backbone network or large wireless base stations in a metropolitan area.

These two scenarios show that the proposed method reduces the call-blocking rate of emergency calls to almost zero in various scale of systems, by appropriate threshold relaxation. Although the call-blocking rate of general calls increases from the result of the conventional method without trunk reservation, the increase in call-blocking rate is suppressed by collaboration of trunk reservation and holding time limitation.

However, general call-blocking rate drastically increases in small systems cases, especially when the mean holding time of emergency calls $1/\mu_e$ is long. This increase is due to the difference between the optimum value of threshold th and the value configured by the proposed method. When emergency callers talk for a long period of time, too many lines may be reserved by Eq. (3.6). Therefore, emergency callers also should reduce their holding time, as long as an essential information can be communicated.

3.6 Conclusion

This chapter proposed a threshold relaxation and general call holding time limitation method for use during emergencies. In this chapter, various systems are theoretically modeled as a wired telephone exchange and a $M_1, M_2/M, D/s/s, th$ queueing loss system. The call-blocking reduction effect of the proposed method is compared with that of a conventional method. From the results of a computer simulation, the proposed method accommodates required emergency calls and suppresses the increase in call-blocking of general calls, by a collaboration of threshold relaxation and holding time limitation of general calls.

Chapter 4

vEPC Optimal Resource Assignment Method for Accommodating M2M Communications

4.1 Introduction

Mobile data traffic is constantly growing with the increasing number of User Equipments (UEs), e.g., smartphones and tablets. Cisco reports that mobile data traffic will reach 49 [EB] per month in 2021, which is about seven times that generated in 2016 [1]. The increase in UEs leads to congestion on both the control plane (C-plane) and the data plane (D-plane) of mobile core networks. Moreover, Machine-to-Machine (M2M) communications [35] using 4G networks are increasing rapidly. M2M communications are automated communications among machines and UEs, and deeply related to the Internet of Things (IoT) [36]. M2M communications are used in vehicular networks for autonomous cars [37] [38], sensor modules that measure temperature or humidity, such as smart cities [39] [40], and so on. Especially there have been various studies on M2M communications among autonomous cars, which are called the Vehicle-to-Vehicle (V2V) communications [41] [42] [43] [44] [45], and autonomous cars are considered as an example of not only M2M communications but also IoT, in most of these studies.

In M2M communications, UEs frequently access mobile core networks by exchanging signaling packets and sending data to database servers. These signaling packets cause heavy congestion on the C-plane, despite the small size of data packets.

Furthermore, 5G mobile network environment has been studied and developed these days [46]. M2M communications in 5G networks require much faster response than in 4G networks. For example, in V2V communications, an end-to-end delay should be less than 20 [msec] [42]. If this allowable delay is not satisfied under a congestion, not only the Service Level Agreement (SLA) of mobile network operators [47] may not be satisfied,

but also the congestion may cause dangerous situations such as traffic accidents. Mobile network operators face the challenge of improving their network infrastructure against this problem.

Nowadays, the concept of a vEPC (Virtualized Evolved Packet Core) has been introduced [48] [49] [50] [51]. The vEPC is a framework for Network Functions Virtualization (NFV). The function of each hardware component of the 3GPP Evolved Packet Core (EPC) is implemented as a Virtualized Network Function (VNF) and works as an application on Virtual Machines (VMs). Virtualization of the EPC has the advantages of reducing capital expenditures (CAPEX) and operating expenditures (OPEX), and some mobile network operators and vendors have developed vEPC-based LTE core networks [52] [53]. The vEPC also enables mobile network operators to deploy each function of the EPC more flexibly, in accordance with traffic demands.

However, there are some problems on vEPC networks. M2M communications usually have a severe allowable end-to-end delay [54]. In a conventional EPC network, some Mobility Management Entities (MMEs) are located close to the evolved NodeB (eNodeB), and connections to such MMEs are limited for M2M devices to satisfy the allowable delay of these M2M communications. If both the C-plane and the D-plane functions of the EPC are migrated into a single VM host, M2M devices and other UEs share the same MME resource. On the one hand, when a large number of smartphones and tablets send signaling requests to the MME, some M2M sessions whose signaling packets exceed their allowable delay may be blocked. On the other hand, assigning too many resources to the MME allows accommodation for more UEs, but resource shortages of the D-plane increase the packet processing delay and decrease the throughput of the UEs. To accommodate these M2M sessions, not only signaling performance on the C-plane but also packet processing performance on the D-plane are required.

Since the traffic requirements of M2M communications are different from other communications, MME resources should be separated for M2M communications and other communications. Nevertheless, the existing implementation of each component for a conventional vEPC, including the MME, is based on conventional EPC components and follows 3GPP specifications. First, it is needed to study the traffic characteristics of the basic vEPC model. A proposal for separating MME resources in accordance with device type is for future work.

This chapter proposes a method for optimizing resource assignment of C-plane and D-plane VNFs in a vEPC server, called as vEPC-ORA (Optimal Resource Assignment) method. The purpose of this study is to reduce the blocking rate of M2M sessions and to process data packets within an allowable delay.

This chapter is organized as follows. Related studies and the purpose of this study are

introduced in Sect. 4.2. Section 4.3 describes the model settings of a virtualized mobile core network. In Sect. 4.4, the proposed method is explained, and optimal resource assignment of the MME and the S/P-GW is derived by using queueing theory. Section 4.5 shows the accommodation effect and performance evaluation of the proposed method. Section 4.6 concludes this chapter.

4.2 Related Study

Although there have been several related studies on this state-of-the-art technology, most research is targeted at signaling performance on the C-plane or packet processing delay on the D-plane. Evaluation of QoS performance of the vEPC in consideration of both the C-plane and the D-plane has not been studied.

Jeon et al. report that the vEPC can be provided as a Network-as-a-Service (NaaS) based on the NFV framework, and its use case could be a flexible on-demand traffic offloading service [46]. Some architecture modes are proposed, but the effect of traffic offloading is not evaluated in [46]. Taleb et al. compare implementation options of EPC entities [49], and mention that the implementation of all entities into a single vEPC server has the advantages of lower internal processing delay between each entity and high parallelization of the same components, while the resource management of C-plane and D-plane is a problem. Gonzalez et al. analyze the failure sources in the vEPC and discuss the service availability of the vEPC implemented on a data center network [50]. Basta et al. discuss the impact of an incremental functions migration to the vEPC on a cloud network [55]. Basta et al. mention that financial cost decreases, but data overhead and end-to-end delay increase as the number of migrated functions of the EPC increases.

Hussien and Elsayed focus on the end-to-end performance of M2M communications on the D-plane and propose applying MPLS tunneling into M2M communications in a vEPC network [56]. A simulation of M2M communication performance shows that the end-to-end delay can be reduced by applying MPLS tunneling to an external IP network. However, the C-plane is not considered in [56]. When a large number of smartphones or tablets transmit an attach request message, the MME cannot process the signaling of M2M devices within an allowable delay, and those M2M sessions may be blocked.

Hasegawa and Murata target increasing the capacity of M2M communications and propose a method for aggregating bearers in an EPC network [57]. Bearers are aggregated between M2M devices and eNodeBs and between eNodeBs and the S-GW. Hasegawa and Murata also propose applying an SDN and separating the S-GW and the P-GW into the C-plane and the D-plane. Numerical analysis showed that the combination of the bearer aggregation method and introducing an SDN increases the number of accommo-

dated M2M devices by 124%. However, all UEs are assumed to be M2M devices in [57], and communication from other devices is out of scope. M2M devices and smartphones have different traffic requirements on both the C-plane and the D-plane. If a large number of smartphones connect to a mobile core network, the broadband communication of smartphones increases the packet processing delay on the S-GW and the P-GW. The end-to-end delay of M2M communications also increases and may exceed the allowable delay.

Finally, Prados et al. have proposed Virtualized MME (vMME) and modeled the MME in VM host by using queueing theory [58]. The vMME is implemented as three components: Service Logics (SLs) for processing signaling packets, Front End (FE) for distributing arrived signaling packets, and State Database (SDB) for preserving each session state. vMME is modeled by using queueing theory, and signaling processing delay and utilization of the vMME is estimated. This estimation is used for studying the scalability of SLs' processing capability in accordance with data center scale. However, [58] targets only the signaling processing capability of the MME and user data packet processing delay on the S-GW and the P-GW is out of scope.

As mentioned in Sect. 4.1, this chapter proposes an optimal resource assignment method in a virtualized mobile core network. The targets of the proposed method are not only conventional 4G networks but also state-of-the-art 5G networks. The communications of M2M devices and other devices, e.g., smartphones and tablets, are distinguished. The purpose of this study is to minimize the blocking rate of M2M sessions, while ensuring that the mean service time on the D-plane does not exceed the allowable delay.

4.3 Model Settings

In this section, a mobile core network is modeled by using queueing theory. Figure 4.1 shows the network model, and both the C-plane and the D-plane components of EPC are migrated into a single vEPC server. The main components are listed below [59] [17]:

- **Evolved NodeB (eNodeB)** is physical hardware and directly connects to UEs.
- **Mobility Management Entity (MME)** authenticates connected users by processing signaling packets of UEs and accommodates authenticated sessions.
- **Serving Gateway (S-GW)** is an internal gateway and forwards user data packets to/from the P-GW or 2G/3G network.

- **Packet Data Network Gateway (P-GW)** is an external gateway and forwards user data packets to/from the S-GW or external IP network.
- **Home Subscriber Server (HSS)** is a database server and contains users' subscription data.

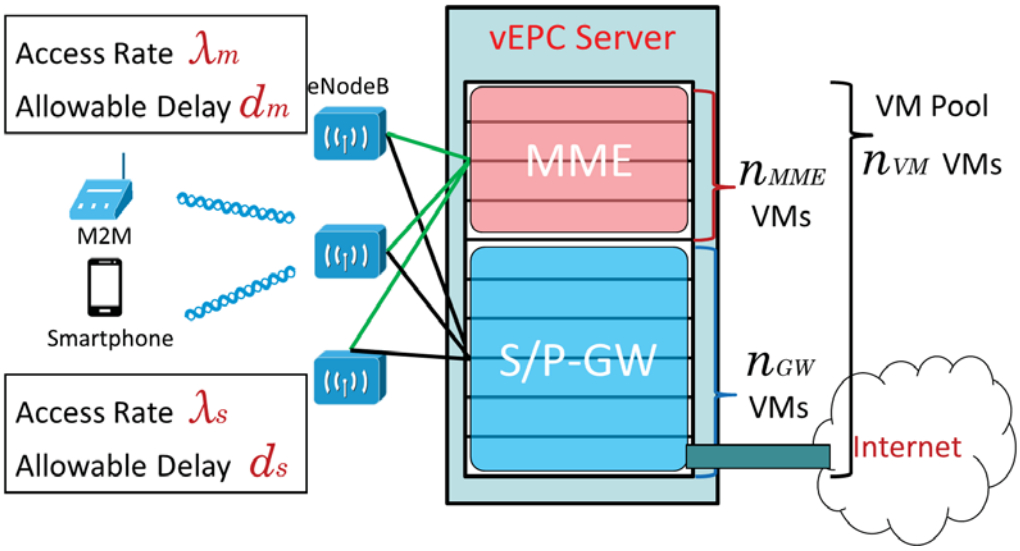


Fig. 4.1: Network model

For simplicity, it is assumed that the functions of the S-GW and the P-GW are migrated into a single gateway (S/P-GW) VNF on the vEPC server. It is also assumed that a HSS is located outside the vEPC server and is out of scope in this study. This is because the MME processes much more signaling packets than the HSS per each Attach/Detach request [17], especially per each Attach request, and thus the MME becomes the bottleneck on the C-plane [60]. For example, M2M communications and smartphone communications are distinguished on the HSS, by referring the International Mobile Subscriber Identity (IMSI) of the SIM card inserted into each UE.

As shown in Fig. 4.1, the vEPC server has a VM pool for n_{VM} VM resources. The MME and S/P-GW VNFs are assigned to this VM pool by using n_{MME} VM resources for the MME and n_{GW} VM resources for the S/P-GW, respectively. In NFV, the performance of VNFs depends on the hardware resources of the VM host server, e.g., the CPU clock, the number of cores or threads, RAM capacity, etc. In this study, the number of VM resources n_{VM} corresponds to the resource granularity of the vEPC server. Access from M2M devices follows a Poisson distribution with an average of λ_m, and that from

smartphones follows a Poisson distribution with an average of λ_s. M2M communications have an allowable delay d_m, and smartphone communications have an allowable delay d_s, on the S/P-GW.

As mentioned in Sect. 4.1, congestion on the C-plane occurs on the MME. The cause of congestion is not only because of the processing delay of signaling packets, but also because of hardware resource shortages for storing ongoing sessions. Thus, a comprehensive model is required for the MME. In this study, the MME is defined as a session pool for accommodating sessions of M2M communications and smartphone communications, and modeled as an $M_1, M_2/M_1, M_2/n_{MME}N/n_{MME}N$ heterogeneous queueing loss system in Fig. 4.2. It is assumed that the maximum number of accommodated sessions is in proportion to the number of MME VM resources n_{MME}, and N sessions can be accommodated per each MME VM resource. Thus, the MME has a session pool for $n_{MME}N$ sessions. Arrival distributions of M2M sessions and smartphone sessions are the same as those of access birth. The service time of M2M sessions follows an exponential distribution with an average of $1/\mu_m$, and that of smartphone sessions follows an exponential distribution with an average of $1/\mu_s$. To accommodate delay-sensitive M2M communications, signaling packets from M2M devices should be processed more quickly than those of smartphones. Therefore, it is assumed that M2M devices require more resources than smartphones, and n_m session resources are assigned to each accommodated M2M session. An incoming M2M session can be accommodated if there are equal to or more than n_m session resources available, and an incoming smartphone session can be accommodated unless the session pool is full of ongoing sessions. Otherwise, incoming sessions are blocked. The blocking rate of M2M sessions is defined as Br_m and that of smartphone sessions is defined as Br_s, respectively. It is also assumed that the ratio of required session resources is in inverse proportion to the ratio of the allowable delay, and n_m is defined that $n_m = \lfloor \frac{d_s}{d_m} \rfloor$.

Finally, the S/P-GW is modeled as an $M_1, M_2/D/1$ queueing delay model in Fig. 4.3. As mentioned in Sect. 4.1, on the one hand M2M devices frequently exchange signaling packets on the C-plane and transmit a small number of user data packets on the D-plane. On the other hand, smartphones transmit a larger number of user data packets on the D-plane, despite the signaling frequency of smartphones being much less than that of M2M devices. On the basis of these backgrounds, it is assumed that M2M devices transmit one packet, and smartphones transmit n_s packets for each session, respectively.

Arrivals of user data packets to the S/P-GW depend on the blocking rates on the MME because user data packets of blocked sessions cannot be transmitted to the S/P-GW. User data packet arrivals from M2M devices follow a Poisson distribution with an average of $(1 - Br_m)\lambda_m$, and those of smartphones follow a Poisson distribution with an average of

$(1 - Br_s)n_s\lambda_s$. In this study, the bottleneck on the D-plane is the packet processing delay on the S/P-GW, and thus, it is assumed that the service rate of user data packets follows a deterministic distribution with a constant value $n_{GW}\mu_{GW}$.

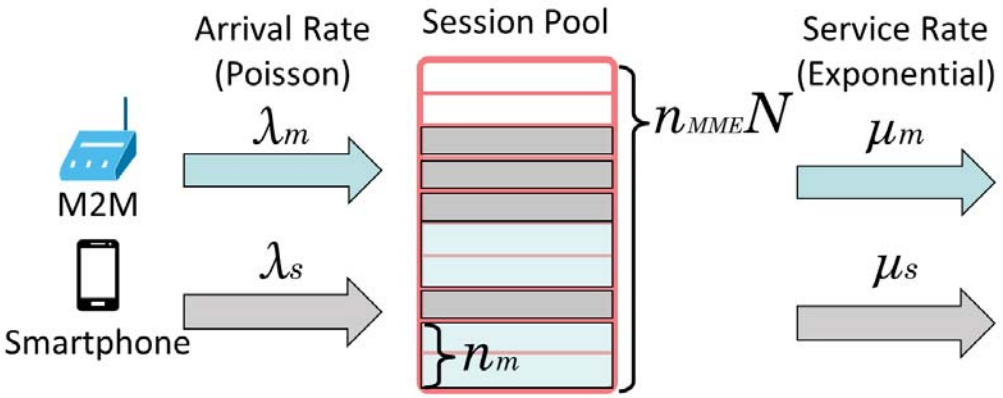


Fig. 4.2: Queueing model of MME

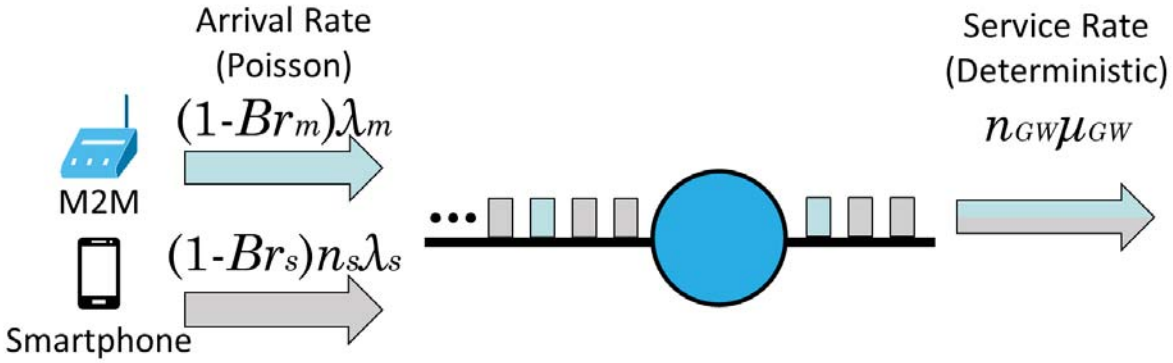


Fig. 4.3: Queueing model of S/P-GW

4.4 Proposed Method

This section proposes an optimal resource assignment of the MME and the S/P-GW on the vEPC server on the basis of the proposed queueing model in Sect. 4.3. The vEPC-ORA method [61] minimizes the blocking rate of M2M sessions Br_m, while the mean packet processing time W does not exceed each allowable delay d_m, d_s.

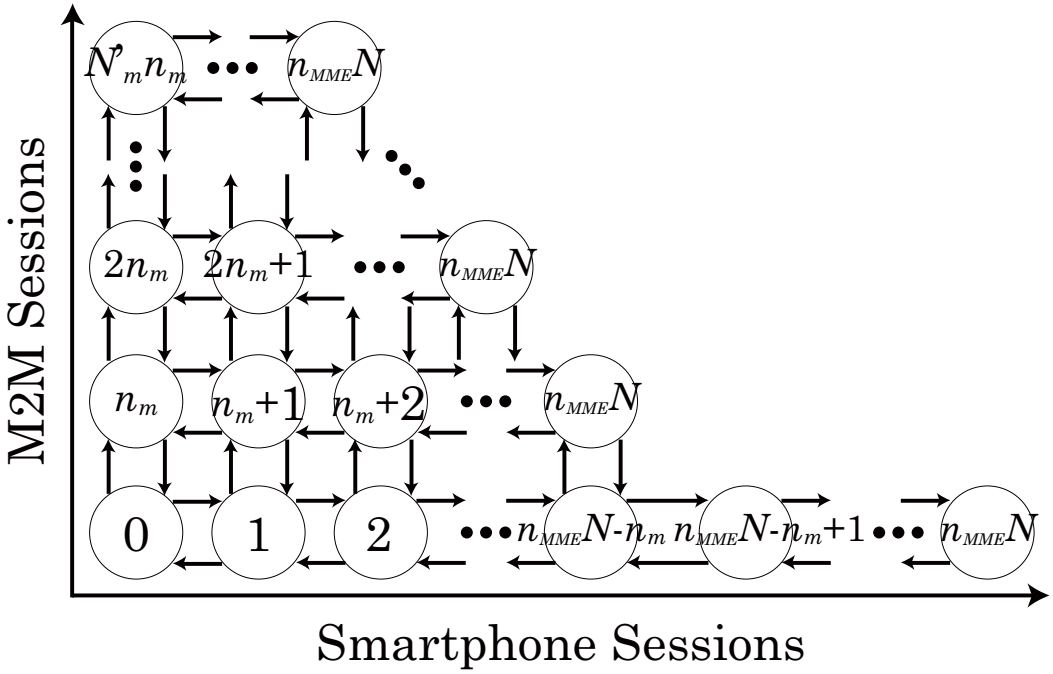
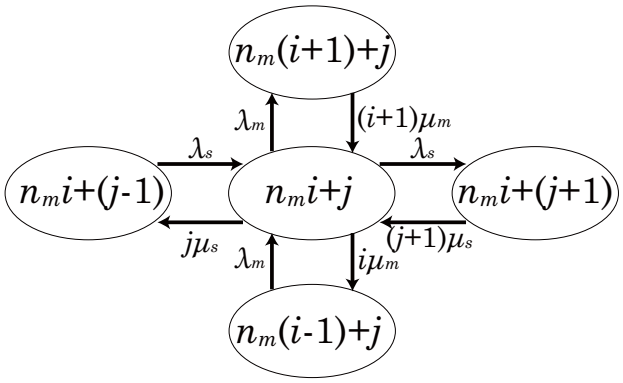
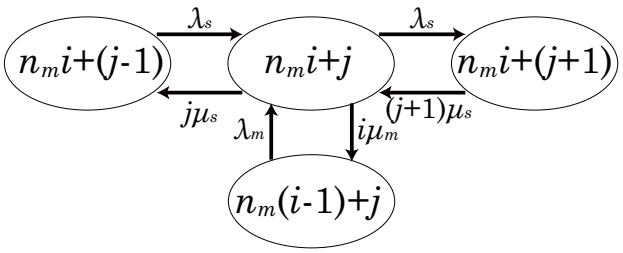


Fig. 4.4: State transition diagram of session pool

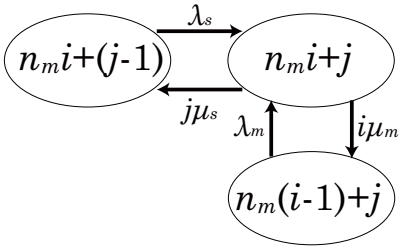
Figure 4.4 shows the state transition diagram of the session pool on the MME [62]. When the number of MME sessions is i and that of smartphone sessions is j, the state probability is defined as $P(i,j)$. Let N'_m be the maximum number of accommodatable M2M sessions and $N'_m = \lfloor \frac{n_{MME}N}{n_m} \rfloor$. $P(i,j)$ is defined as $P(i,j) = 0$ when $i < 0$ or $j < 0$. State transition events in infinitesimal time (Δt) are limited to neighboring states because both the arrival process and the service time satisfy the Markov property. Therefore, state transitions are listed below and the state transition diagram for each condition is shown in Fig. 4.5:



(a) $0 \leq n_m i + j < n_{MME} N - n_m + 1$



(b) $n_{MME} N - n_m + 1 \leq n_m i + j < n_{MME} N$



(c) $n_m i + j = n_{MME} N$

Fig. 4.5: State diagram for each condition

1. When $0 \leq n_m i + j < n_{MME}N - n_m + 1$

$$\begin{aligned}(\lambda_m + \lambda_s + i\mu_m + j\mu_s)P(i,j) \\ = \lambda_m P(i-1,j) + \lambda_s P(i,j-1) + (i+1) \\ \mu_m P(i+1,j) + (j+1)\mu_s P(i,j+1)\end{aligned} \tag{4.1}$$

2. When $n_{MME}N - n_m + 1 \leq n_m i + j < n_{MME}N$

$$\begin{aligned}(\lambda_s + i\mu_m + j\mu_s)P(i,j) \\ = \lambda_m P(i-1,j) + \lambda_s P(i,j-1) \\ +(j+1)\mu_s P(i,j+1)\end{aligned} \tag{4.2}$$

3. When $n_m i + j = n_{MME}N$

$$\begin{aligned}(i\mu_m + j\mu_s)P(i,j) \\ = \lambda_m P(i-1,j) + \lambda_s P(i,j-1)\end{aligned} \tag{4.3}$$

By Eqs. (1)–(3) and the normalization condition $\sum_{0 \leq n_m i + j \leq n_{MME}N} P(i,j) = 1$, state probability $P(i,j)$ is calculated as:

$$P(i,j) = \frac{a_m^i}{i!}\frac{a_s^j}{j!}P(0,0) \tag{4.4}$$

$$\text{where} \quad a_m = \frac{\lambda_m}{\mu_m}, a_s = \frac{\lambda_s}{\mu_s}, \tag{4.5}$$

$$P(0,0) = \left[\sum_{i=0}^{N'_m} \sum_{j=0}^{n_{MME}N - n_m i} \frac{a_m^i a_s^j}{i!j!}\right]^{-1} \tag{4.6}$$

On the one hand, the blocking rate of M2M sessions Br_m and the blocking rate of smartphone sessions Br_s are also calculated by summation of Eq. 4.4:

$$Br_m = \left(\sum_{i=0}^{N'_m-1} \sum_{j=n_{MME}N-n_m(i+1)+1}^{n_{MME}N-n_m i} \frac{a_m^i a_s^j}{i!j!} + \frac{a_m^{N'_m}}{N'_m!} \sum_{j=0}^{k} \frac{a_s^j}{j!} \right) P(0,0) \tag{4.7}$$

$$\text{where} \quad k = n_{MME}N \mod n_m \tag{4.8}$$

$$Br_s = \left(\sum_{i=0}^{N'_m} \frac{a_m^i}{i!} \frac{a_s^{n_{MME}N-n_m i}}{(n_{MME}N - n_m i)!} \right) P(0,0) \tag{4.9}$$

On the other hand, the mean packet processing time on S/P-GW W can be calculated by the Pollaczek-Khinchine formula:

$$W = \frac{\rho}{1-\rho} \frac{1+C_s^2}{2} h \tag{4.10}$$

When the utilization of S/P-GW ρ is defined as

$$\rho = \frac{(1-Br_m)\lambda_m + (1-Br_s)n_s\lambda_s}{n_{GW}\mu_{GW}}, \tag{4.11}$$

the squared coefficient of variation $C_s^2 = v/h^2 = 0$ because the variance of deterministic service time $v = 0$ and mean service time $h = 1/n_{GW}\mu_{GW}$. Thus, W can be calculated as follows:

$$W = \frac{\rho}{2(1-\rho)n_{GW}\mu_{GW}} \tag{4.12}$$

In this study, M2M communications have stricter allowable delay d_m than that of smartphone communications d_s, and allowable delays of those two types of devices satisfy $d_m < d_s$. Therefore, mean packet processing time W follows the inequality below:

$$\frac{\rho}{2(1-\rho)n_{GW}\mu_{GW}} \leq d_m \tag{4.13}$$

$$\begin{aligned} \iff \quad & 2\mu_{GW}^2 d_m n_{GW}^2 \\ & -2((1-Br_m)\lambda_m + (1-Br_s)n_s\lambda_s)\mu_{GW} d_m n_{GW} \\ & -((1-Br_m)\lambda_m + (1-Br_s)n_s\lambda_s) \geq 0 \end{aligned} \tag{4.14}$$

From Eq. 4.14 and $n_{GW} > 0$, the number of S/P-GW VM resources n_{GW} follows the inequality below:

$$n_{GW} \geq \frac{\Lambda + D}{2\mu_{GW}} \tag{4.15}$$

$$\text{where} \quad \Lambda = (1-Br_m)\lambda_m + (1-Br_s)n_s\lambda_s, \tag{4.16}$$

$$D = \sqrt{\Lambda^2 + \frac{2\Lambda}{d_m}} \tag{4.17}$$

As a result, the optimal VM assignment of MME, S/P-GW $S^* = (n^*_{MME}, n^*_{GW})$ is solved as:

$$S^* \in \underset{n_{MME}, n_{GW}}{\arg\min}\ Br_m$$
$$s.t.\ n_{MME} + n_{GW} = n_{VM},\ n_{GW} \geq \frac{\Lambda + D}{2\mu_{GW}} \tag{4.18}$$

Since both n_{MME} and n_{GW} are natural numbers, Eq. 4.18 can be rewritten as Eq. 4.19.

$$n^*_{MME} = n_{VM} - \left\lceil \frac{\Lambda + D}{2\mu_{GW}} \right\rceil, n^*_{GW} = \left\lceil \frac{\Lambda + D}{2\mu_{GW}} \right\rceil \tag{4.19}$$

4.5 Numerical Evaluation

Finally, the blocking rate reduction effect of the vEPC-ORA method is evaluated by numerical evaluations. The blocking rates of M2M sessions and smartphone sessions on the MME, and the mean packet processing time on the S/P-GW, are calculated by a brute force calculation. The traffic characteristics of blocking rates or the mean packet processing time are investigated for various traffic situations in this section.

It is assumed that the mean service time of M2M sessions $1/\mu_m = 1$ [sec] and the mean service time of smartphone sessions $1/\mu_s = 60$ [sec] throughout this chapter. Subsections 4.5.1–4.5.4 are partially based on the author's previous study [61] and Subsect. 4.5.6 is partially based on the author's previous study [63].

4.5.1 Blocking Rates and Mean Packet Processing Time

Figure 4.6 shows the relationship among the number of MME VM resources n_{MME} and blocking rates Br_m, Br_s for $\lambda_m = \lambda_s = 50$ [/sec], $d_m = 1$ [msec], $d_s = 10$ [msec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor = 10$, $n_s = 500$, $N = 250$, $\mu_{GW} = 1500$ [pps] and $n_{VM} = 50$. The rest of the VM resources are assigned to the S/P-GW. In addition, mean packet processing time W [sec] is plotted on a semilogarithmic scale.

In Subsects. 4.5.1–4.5.5, as an example of the number of session pools per MME VM resource N, a 1/1000-scale value of a call processing blade on a Mobile-service Switching Center (MSC) in 3G networks [64] is used and $N = 250$. The role of a MSC call processing blade is almost equivalent to the MME in 4G and 5G networks. This study also assumes a 1/1000-scale wire speed of a 1000BASE-T Ethernet switch, as an example of the service rate of user data packets per S/P-GW VM resource μ_{GW}. According to the minimum Ethernet frame size 64 [byte], the preamble size 8 [byte] and the interframe gap 12 [byte], μ_{GW} is configured as: 1000 [Mbps] / ((64 + 8 + 12) $\times$ 8 [bits] $\times$ 1000) = 1488.0... $\fallingdotseq$ 1500 [pps].

These values are some examples, and various values can be applied to parameters N and μ_{GW}, as both the hardware performance of the vEPC server and the software performance of each vEPC entity VNFs are improved.

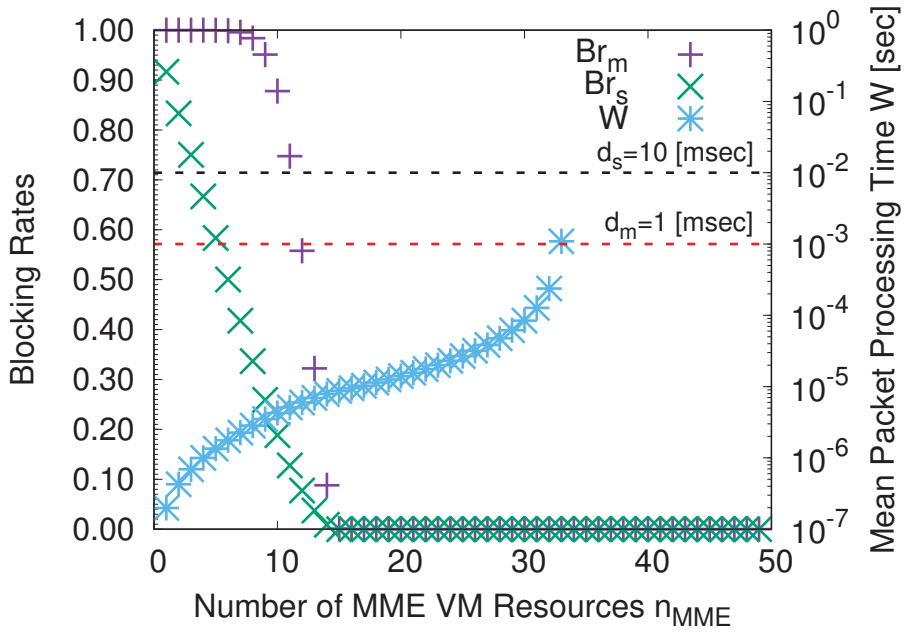


Fig. 4.6: Blocking rates and mean packet processing time ($\lambda_m = 50$ [/sec], $\lambda_s = 50$ [/sec])

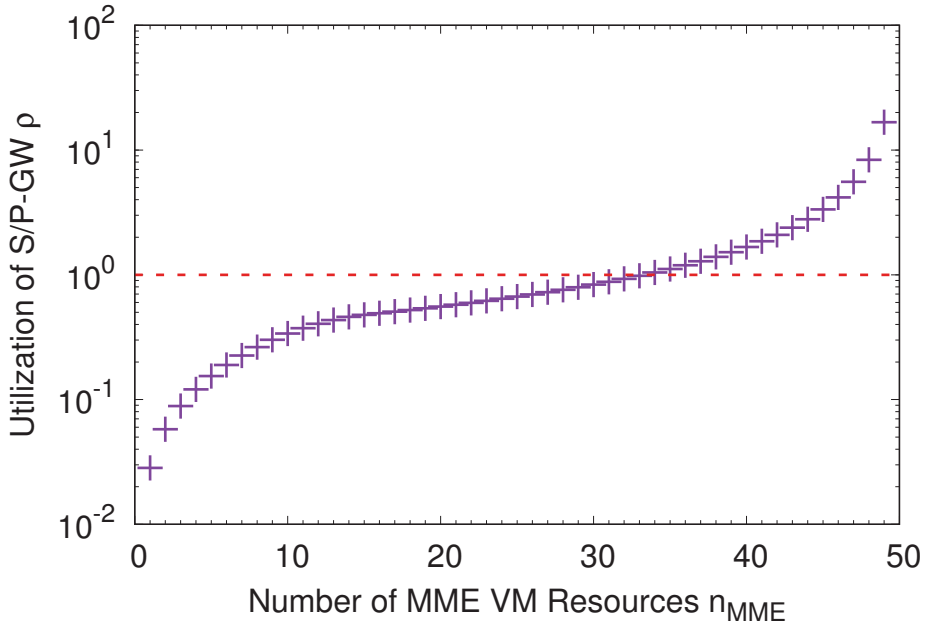


Fig. 4.7: Utilization of S/P-GW ($\lambda_m = 50$ [/sec], $\lambda_s = 50$ [/sec])

As shown in Fig. 4.6, a lot of sessions of both M2M and smartphone communications are blocked when the resources of the MME are not enough. These blocking rates are reduced as n_{MME} increases and reduced to almost zero when $n_{MME} \geq 15$. However, mean packet processing time W gradually increases as n_{MME} increases, and the packet service rate of S/P-GW $n_{GW}\mu_{GW}$ decreases. W exceeds the allowable delay of M2M communications d_m when $n_{MME} = 33$ and diverges to infinity when $n_{MME} \geq 34$. This is because the arrival rate of sessions exceeds the service rate on the S/P-GW. Figure 4.7 shows the relationship between n_{MME} and the utilization of S/P-GW ρ on a semilogarithmic scale. In the proposed $M_1, M_2/D/1$ queueing model, the mean packet processing time can be calculated by Eq. 4.12 only when $\rho < 1$. The optimal resource assignment $S^* = (32, 18)$ derived from the vEPC-ORA method reduces the blocking rates of both M2M and smartphone sessions with $Br_m = 9.9 \times 10^{-272}$ and $Br_s = 3.1 \times 10^{-273}$, while the mean packet processing time is kept within the allowable delay of M2M communications with $W = 0.24$ [msec]. Since there are enough VM resources in this traffic situation, both M2M and smartphone communications can be accommodated even though the resource assignment is not optimal. For example, the resource assignment $S = (15, 35)$ gives $Br_m = 0.0012, Br_s = 0.00010$ and $W = 0.0087$ [msec]. The evaluation results are theoretical value, but both the blocking rates and mean packet processing time are reduced for practical use, in the range of $15 \leq n_{MME} \leq 32$. Mobile network operators may

choose an appropriate resource assignment S according to the result of the brute force calculation.

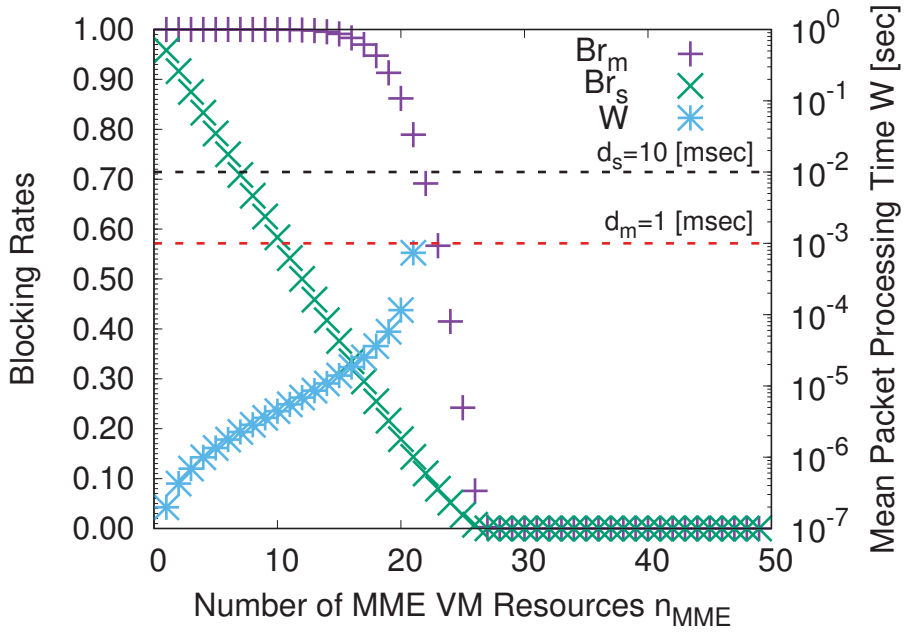


Fig. 4.8: Blocking rates and mean packet processing time ($\lambda_m = 50$ [/sec], $\lambda_s = 100$ [/sec])

On the other hand, Fig. 4.8 shows the relationship between n_{MME} and Br_m, Br_s, W for $\lambda_m = 50$ [/sec], $\lambda_s = 100$ [/sec] while other parameters are the same as Fig. 4.6. In this traffic condition, there are no resource assignments which give both the better blocking rates and the better mean packet processing time, because the amount of incoming traffic is over the capacity of the vEPC server. It is needed to operate the vEPC-ORA method to derive the optimal resource assignment under such a heavy congestion, so that as many M2M sessions and smartphone sessions as possible are accommodated, while ensuring that the mean packet processing time does not exceed the allowable delays.

4.5.2 Access Rate

Second, Fig. 4.9 shows the relationship between the ratio of arrival rates and the optimal number of MME VM resources n^*_{MME} for $d_m = 1$ [msec], $d_s = 10$ [msec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor = 10$, $n_s = 500$, $N = 250$, $\mu_{GW} = 1500$ [pps] and $n_{VM} = 50$. The sum of access rates

is fixed as $\lambda_m + \lambda_s = 100$ [/sec], and there are nine values for (λ_m, λ_s), from $(10, 90)$ to $(90, 10)$. λ_m is shown in the horizontal axis, and blocking rates Br_m, Br_s are also plotted in the vertical axis. Figure 4.10 shows the relationship between the ratio of arrival rates and mean packet processing time W under the same traffic condition.

As shown in Fig. 4.9, about 59% of M2M sessions and about 8% of smartphone sessions are blocked only when $(\lambda_m, \lambda_s) = (10, 90)$ [/sec]. This is because a lot of incoming smartphone sessions require many packet processing resources from the S/P-GW, and thus, the session pool on the MME decreases. Figure 4.10 shows that W is kept within d_m and d_s in every traffic situation. If more VM resources are assigned to MME, W exceeds d_m or even diverges to infinity and cannot satisfy the traffic requirements of M2M sessions. These results mean that a much greater resource capacity for the vEPC server is required especially when the number of smartphone broadband communications is much larger than the number of M2M communications.

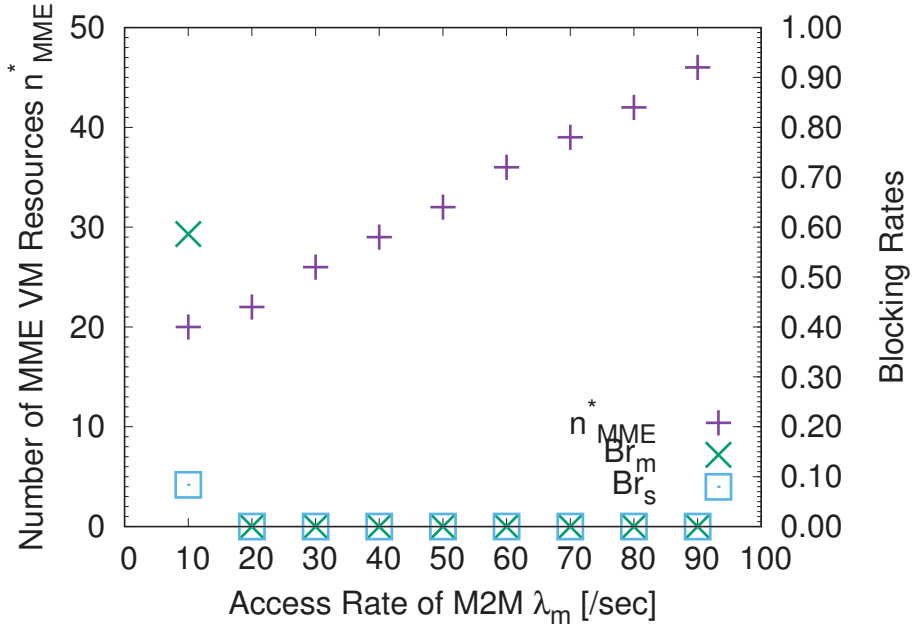


Fig. 4.9: Relationship between access rate and optimal resource assignment

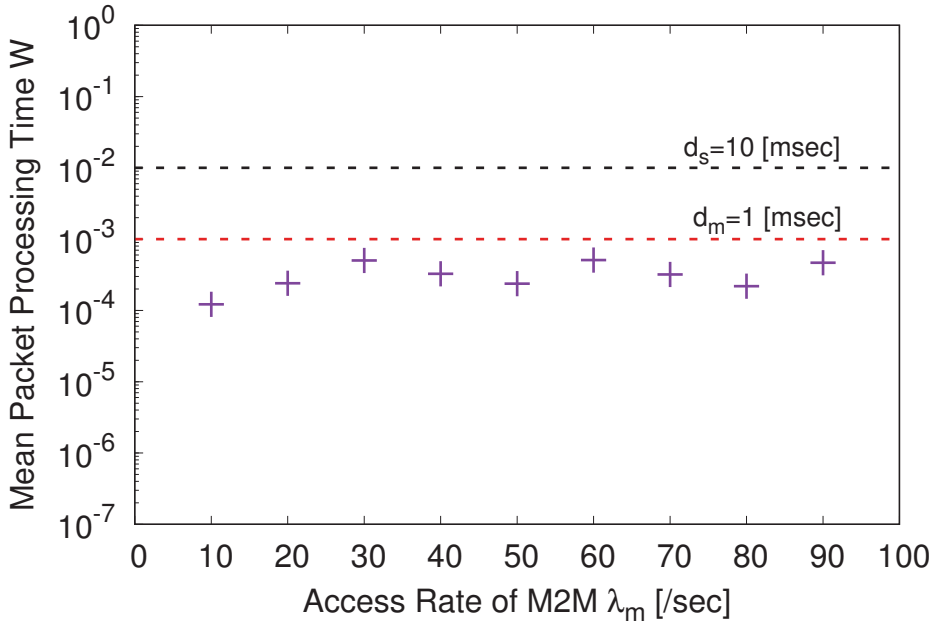


Fig. 4.10: Relationship between access rate and mean packet processing time

4.5.3 Resource Capacity of vEPC Server

As mentioned in the last subsection, the resource capacity of the vEPC server is the most important factor in accommodating sessions and satisfying traffic requirements. This subsection investigates the effect of resource capacity on QoS of both communication types. Figure 4.11 shows the relationship among the number of VM resources n_{VM} and blocking rates Br_m, Br_s under optimal resource assignment S^* for $\lambda_m = \lambda_s = 50$ [/sec], $d_m = 1$ [msec], $d_s = 10$ [msec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor = 10$, $n_s = 500$, $N = 250$ and $\mu_{GW} = 1500$ [pps]. n_{VM} is given ten values from 5 to 50. Mean packet processing time W is also plotted on a semilogarithmic scale in the vertical axis.

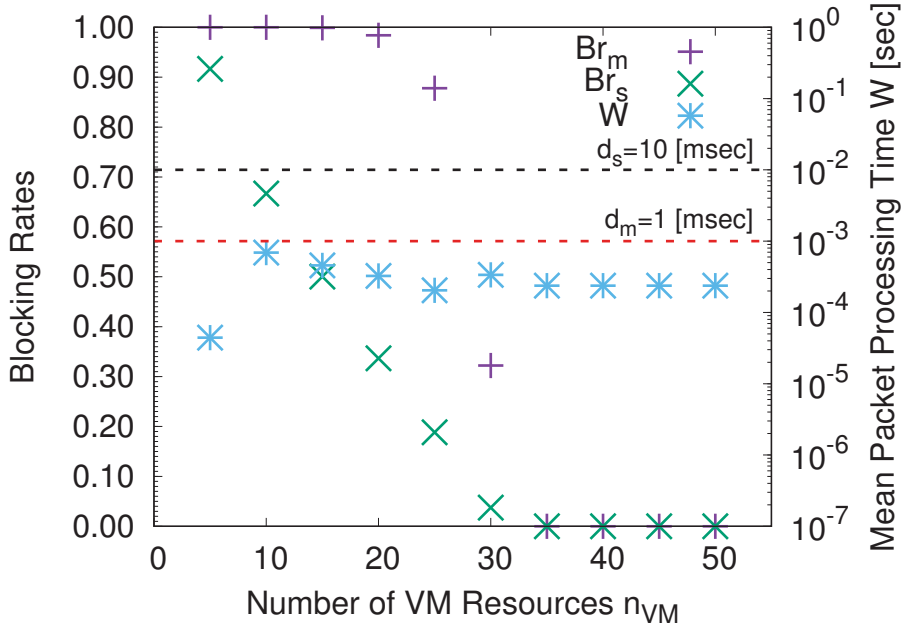


Fig. 4.11: Effect of resource capacity of vEPC server on QoS

Figure 4.11 shows that W is kept within d_m and d_s. However, resource shortages on the MME cause a lot of blockings for both M2M sessions and smartphone sessions. Moreover, most M2M sessions are blocked when $n_{VM} \leq 20$. To accommodate most M2M sessions, VM resources of at least $n_{VM} = 35$ are required in this traffic situation. The calculation of Br_m, Br_s, W for various n_{VM} can be applied for estimation of the vEPC server hardware requirements.

4.5.4 User Data Packet Rate per Smartphone Session

In this subsection, the relationship among the user data packet rate per smartphone session n_s and the QoS of both communication types is investigated. In smartphone communications, there are various types of transmission rates and bandwidth requirements. If there are some broadband communications with a large packet rate, a large number of user data packets arrive at the S/P-GW and may cause congestion even though the number of smartphone sessions is small. Figure 4.12 shows the effect of user data packet rate per smartphone session n_s on Br_m, Br_s and W under optimal resource assignment S^* for $\lambda_m = \lambda_s = 50$ [/sec], $d_m = 1$ [msec], $d_s = 10$ [msec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor = 10$, $N = 250$, $\mu_{GW} = 1500$ [pps] and $n_{VM} = 50$. n_s is given ten values, from 100 to 1000. W is plotted on a semilogarithmic scale.

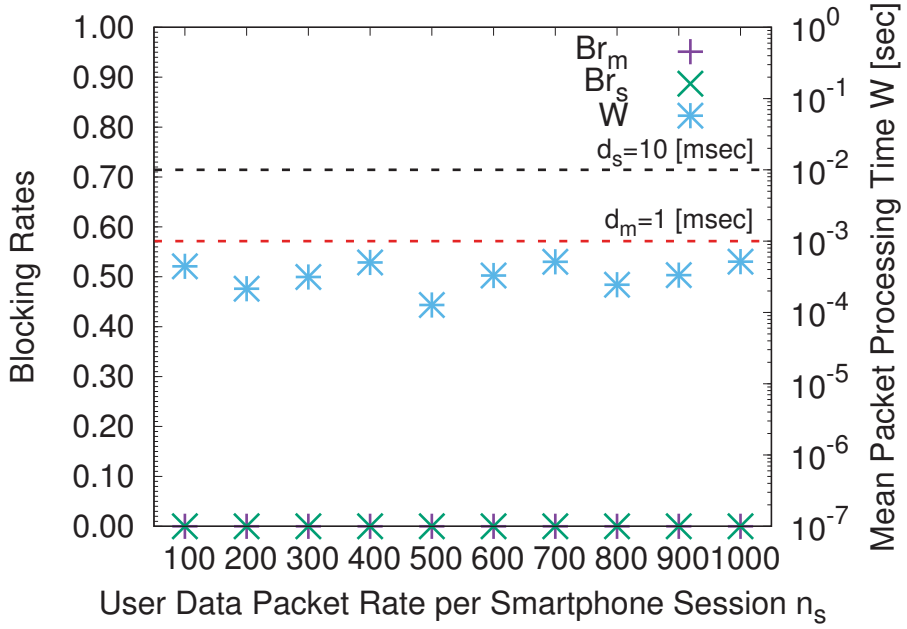


Fig. 4.12: Effect of user data packet rate per smartphone session on QoS

As shown in Fig. 4.12, both blocking rates Br_m and Br_s are reduced to almost zero by the vEPC-ORA method. Mean packet processing time W is also reduced to be less than d_m and d_s. These results mean that the vEPC-ORA method does not depend on the user data packet rate of smartphone communications, as long as the vEPC server has enough resources. The proposed optimal resource assignment can accommodate both M2M sessions and various types of smartphone communications such as VoLTE, video streaming, game applications, etc.

4.5.5 Allowable Delay

This subsection investigates the effect of the allowable delay of each session type d_m, d_s on the optimal resource assignment derived from the vEPC-ORA method.

Figure 4.13 shows the relationship among the M2M allowable delay d_m and blocking rates Br_m, Br_s under optimal resource assignment S^* for $\lambda_m = \lambda_s = 50$ [/sec], $1/\mu_m = 1$ [sec], $1/\mu_s = 60$ [sec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor$, $N = 250$, $\mu_{GW} = 1500$ [pps] and $n_{VM} = 50$. The smartphone allowable delay is fixed as $d_s = 10$ [msec] and the M2M allowable delay d_m is given nine values from 1 [msec] to 9 [msec]. Figure 4.14 shows the relationship between d_m and mean packet processing time W under the same traffic condition.

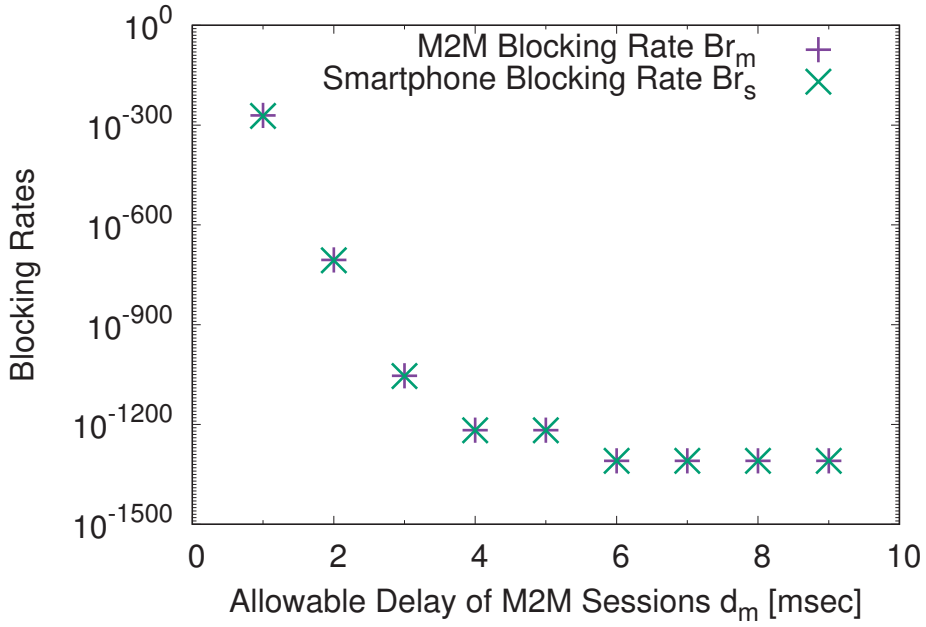


Fig. 4.13: Effect of M2M Allowable Delay on Blocking Rates

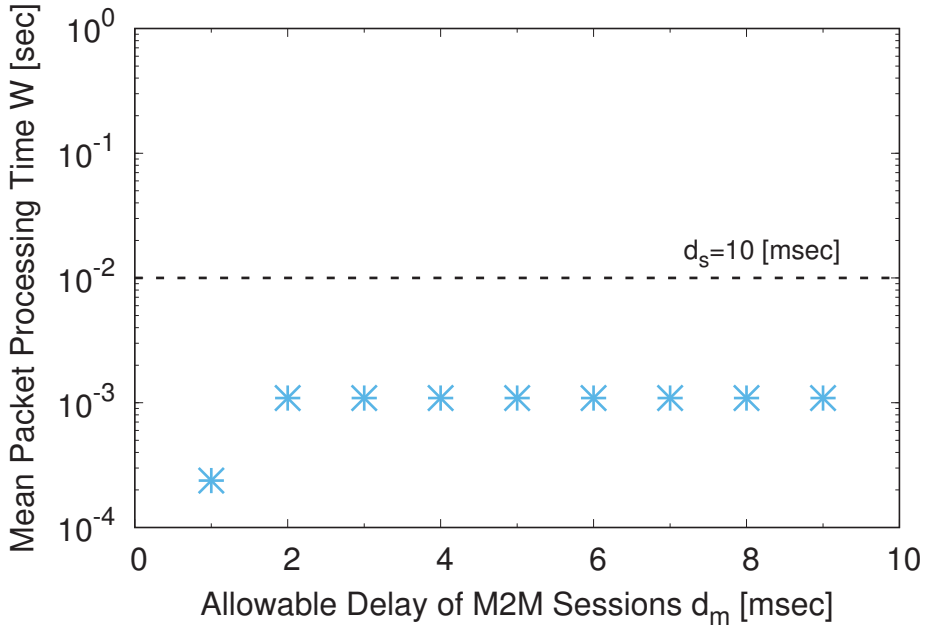


Fig. 4.14: Effect of M2M Allowable Delay on Mean Packet Processing Time

Fig. 4.13 shows that both blocking rates are reduced to almost zero for each value of d_m and blocking rates gradually decrease as d_m increases. This is because the number of session resource required per M2M session n_m becomes smaller and gradually approaches to 1. Fig. 4.14 shows that W has a constant value $W = 1.09$ [msec] when $d_m \geq 2$ [msec]. This is because the optimal resource assignment of S/P-GW n^*_{GW} is kept as $n^*_{GW} = 33$ and the packet processing rate is constant when $d_m \geq 2$ [msec], while $n^*_{GW} = 32$ when $d_m = 1$ [msec].

On the other hand, Fig. 4.15 shows the relationship among the smartphone allowable delay d_s and Br_m, Br_s, W under optimal assignment S^* for $\lambda_m = \lambda_s = 50$ [/sec], $1/\mu_m = 1$ [sec], $1/\mu_s = 60$ [sec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor$, $N = 250$, $\mu_{GW} = 1500$ [pps] and $n_{VM} = 50$. The M2M allowable delay is fixed as $d_m = 10$ [msec] and the smartphone allowable delay d_s is given ten values from 10 [msec] to 100 [msec].

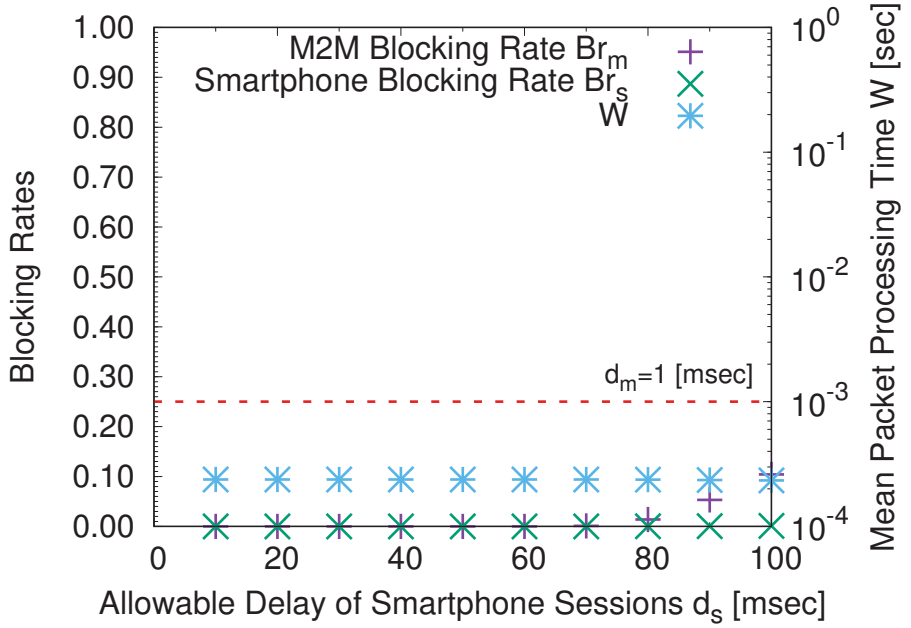


Fig. 4.15: Effect of Smartphone Allowable Delay on QoS

As shown in Fig. 4.15, the mean packet processing time W is reduced below d_m for each value of d_s, by the optimal resource assignment derived from the vEPC-ORA method. The smartphone blocking rate Br_s is also reduced to almost zero for each d_s, but the M2M blocking rate Br_m gradually increases in the range of $d_s \geq 80$ [msec]. When $d_s = 100$ [msec], Br_m especially increases to $Br_m = 0.104$ and about 10% of M2M

sessions are blocked. This is because the value of n_m increases as d_s increases, and the number of session resource required per M2M session becomes larger. In this chapter, it is assumed that the value of n_m is in inverse proportion to the ratio of allowable delay. However, if the smartphone allowable delay is much larger than the M2M allowable delay, too much session resources may be assigned to each M2M session based on this assumption. Since there have been some studies on performance evaluation of the implemented vEPC server [65] [66], the difference between signaling performance requirements for both M2M and smartphone communications in the real vEPC network will be taken into account, and the network model will be expanded for future work.

4.5.6 Resource Granularity

In the vEPC-ORA method, it is assumed that the hardware resources of a vEPC server can be separated into units of a natural number n_{VM}, and the vEPC server is modeled as a VM pool of n_{VM} VM resources. However, although the brute-force calculation of resource assignments depends on n_{VM} and requires an exponential time, the effect of resource granularity (number of VM resources) on QoS has not been studied.

For this reason, this subsection evaluates the effect of resource granularity n_{VM} on the optimal resource assignment S^* derived from the vEPC-ORA method [63]. The total hardware resources of MME and S/P-GW is fixed as $n_{VM}N = 12500$ and $n_{VM}\mu_{GW} = 75000$ [pps]. The number of session pools per MME VM resource N and the service rate of user data packets per S/P-GW VM resource μ_{GW} differ according to the resource granularity n_{VM}. The brute force calculation is processed in the system with dual socket Intel® Xeon® E5-2640v2 Processor @ 2.0 GHz.

Figures 4.16-4.18 show the evaluation results for $\lambda_m = \lambda_s = 50$ [/sec], $d_m = 1$ [msec], $d_s = 10$ [msec], $n_m = \lfloor \frac{d_s}{d_m} \rfloor = 10$, $n_s = 500$. n_{VM} is given 17 values, 2, 3, ..., 9, 10, 15, ..., 45 and 50. Figure 4.16 shows that the calculation time is almost in proportion to the VM resource granularity n_{VM}. However, Fig. 4.17 shows that blocking rate reduction effect with a fine-grained resource granularity is not remarkable. Figure 4.18 shows that the mean packet processing time becomes close to the allowable delay of M2M sessions d_m, as the resource granularity n_{VM} increases.

As shown in the evaluation results, the vEPC-ORA method with any VM resource granularity n_{VM} reduces not only both blocking rates to almost zero, but also mean packet processing time below d_m=1 [msec]. In particular, the calculation time with $n_{VM} = 3$ is only 2.4% of that with $n_{VM} = 50$, while both blocking rates with $n_{VM} = 3$ are below 10^{-10}. The vEPC-ORA method with coarse-grained resource granularity derives the optimal resource assignment of MME and S/P-GW in a practical calculation time, even though the calculation of blocking rates requires an exponential time.

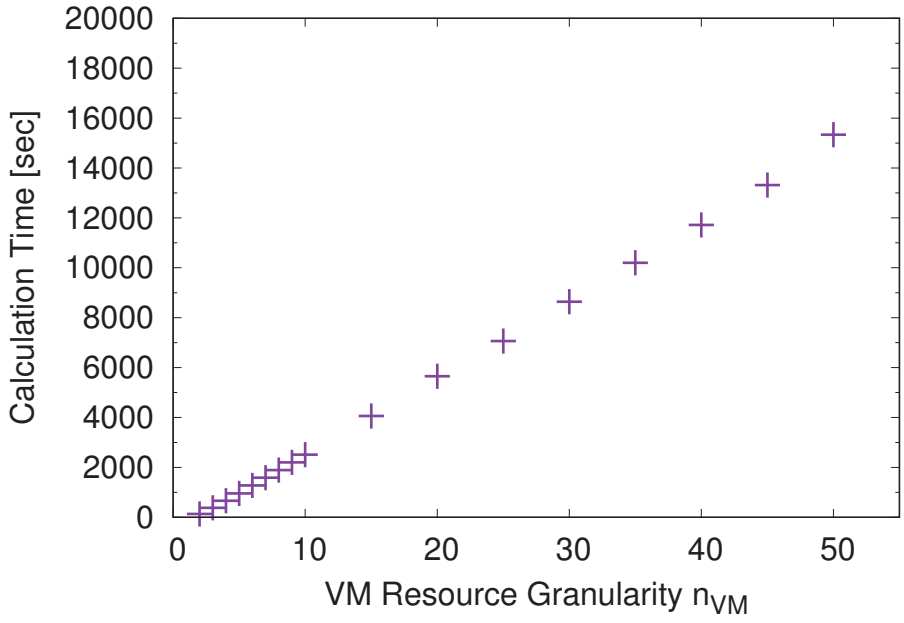


Fig. 4.16: Calculation Time

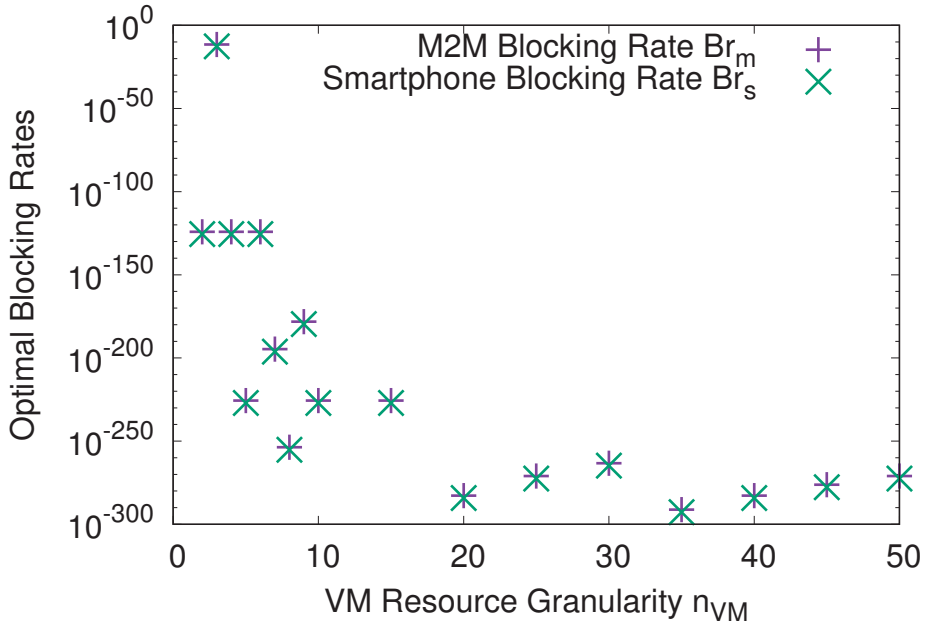


Fig. 4.17: Effect of Resource Granularity on Blocking Rates

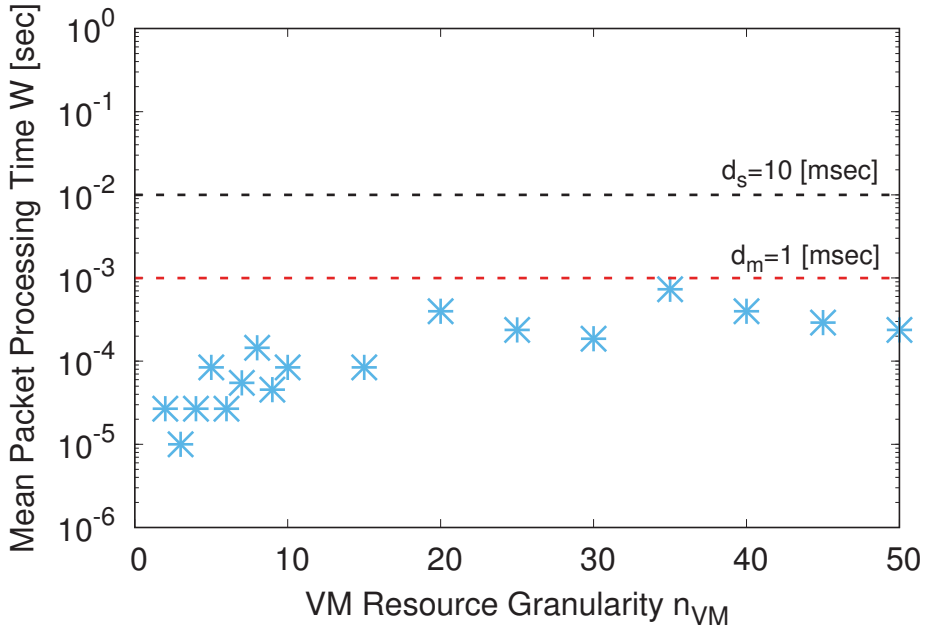


Fig. 4.18: Effect of Resource Granularity on Mean Packet Processing Time

4.6 Conclusion

This chapter proposed an optimal resource assignment method of the C-plane and the D-plane for vEPC mobile core networks. This study focused on delay-sensitive M2M communications in both 4G and 5G networks. The communications of M2M devices and smartphones are distinguished and the vEPC server is modeled with consideration of the different traffic demands of those two communication types. In this chapter, the MME on the C-plane is modeled as an $M_1, M_2/M_1, M_2/n_{MME}N/n_{MME}N$ heterogeneous queueing loss system, and the S/P-GW on the D-plane is modeled as an $M_1, M_2/D/1$ queueing delay system. The vEPC-ORA method derives the optimal resource assignment and accommodates more sessions of not only M2M devices but also smartphones, while the mean packet processing time on the S/P-GW is kept within the allowable delay of each communication type.

Numerical evaluations of optimal resource assignment showed that the vEPC-ORA method minimizes the blocking rates of M2M sessions and smartphone sessions, as long as the arrival rate of the user data packet does not exceed the packet service rate on the S/P-GW. Moreover, the relationship between the ratio of the access rate, the resource capacity of the vEPC server or the user data packet rate per smartphone session, and the QoS of M2M devices and smartphones are investigated. It is confirmed that the vEPC-ORA method accommodates both M2M sessions and smartphone sessions while ensuring that the allowable delay of each traffic situation is not exceeded, as long as the vEPC server has enough VM resources. Moreover, the resource granularity effect of a vEPC server on the optimal resource assignment of MME and S/P-GW was studied. Numerical evaluations showed that the vEPC-ORA method derives the optimal resource assignment in a practical calculation time.

In future work, the MME session pool will be divided into two session pools, and an incoming session will be accommodated separately in accordance with its communication type. Since the normal data transmission of smartphones has a much longer allowable delay than that of M2M communications, a call admission control (CAC) method in which M2M sessions are accommodated with priority and smartphone sessions are allowed to wait in a delay queue will also be studied.

Chapter 5

Conclusions

This thesis proposed two novel CAC methods to appropriately assign network resources and guarantee the QoS of both voice and data traffic in telecommunications networks. In the IP-based network environments, excess accommodation of incoming sessions causes heavy congestion on both the C-plane and the D-plane, and thus the QoS is degraded in the entire network. The proposed methods determine the maximum number of accommodatable sessions by using queueing theory, and control incoming sessions so that more communication traffic is accommodated with a guaranteed level of QoS.

Chapter 3 proposed a threshold relaxation and general call holding time limitation method for use during emergencies. In this chapter, various systems were theoretically modeled as a wired telephone exchange and a $M_1, M_2/M, D/s/s, th$ queueing loss system. The call-blocking reduction effect of the proposed method was compared with that of a conventional method. From the results of a computer simulation, the proposed method accommodates required emergency calls and suppresses the increase in call-blocking of general calls, by a collaboration of threshold relaxation and holding time limitation of general calls. In Chapter 3, it is assumed that general call users have a constant holding time especially when the time limit is relatively short. When the time limit is quite long, this assumption is not appropriate and the holding time follows a truncated exponential distribution i.e. Okada's [26]. The situation when the time limit is long will be considered, and this truncated exponential distribution will be applied to the proposed method. In addition, the result of computer simulations will be utilized for expanding the proposed method into dynamic control method of the threshold and general call holding time limit.

Chapter 4 proposed an optimal resource assignment method of the C-plane and the D-plane for vEPC mobile core networks. This study focused on delay-sensitive M2M communications in both 4G and 5G networks. The communications of M2M devices and smartphones were distinguished and the vEPC server was modeled with consideration of the different traffic demands of those two communication types. In this chapter, the MME on the C-plane was modeled as an $M_1, M_2/M_1, M_2/n_{MME}N/n_{MME}N$ het-

erogeneous queueing loss system, and the S/P-GW on the D-plane was modeled as an $M_1, M_2/D/1$ queueing delay system. The vEPC-ORA method derives the optimal resource assignment and accommodates more sessions of not only M2M devices but also smartphones, while the mean packet processing time on the S/P-GW is kept within the allowable delay of each communication type. Numerical evaluations of optimal resource assignment showed that the vEPC-ORA method minimizes the blocking rates of M2M sessions and smartphone sessions, as long as the arrival rate of the user data packet does not exceed the packet service rate on the S/P-GW. Moreover, the relationship between the ratio of the access rate, the resource capacity of the vEPC server or the user data packet rate per smartphone session, and the QoS of M2M devices and smartphones are investigated. It is confirmed that the vEPC-ORA method accommodates both M2M sessions and smartphone sessions while ensuring that the allowable delay of each traffic situation is not exceeded, as long as the vEPC server has enough VM resources. Moreover, the resource granularity effect of a vEPC server on the optimal resource assignment of MME and S/P-GW was studied. Numerical evaluations showed that the vEPC-ORA method derives the optimal resource assignment in a practical calculation time. For future work of Chapter 4, the MME session pool will be divided into two session pools, and an incoming session will be accommodated separately in accordance with its communication type. Since the normal data transmission of smartphones has a much longer allowable delay than that of M2M communications, a call admission control (CAC) method in which M2M sessions are accommodated with priority and smartphone sessions are allowed to wait in a delay queue will also be studied.

These proposed methods realize the efficient and reliable telecommunication network infrastructure. By guaranteeing the QoS of each session, users can use both voice and data communications with relief, even in a congested network condition. Moreover, the proposed methods are also effective in future network environments, where a lot of IoT/M2M devices and high data rate mobile devices coexist and are connected to the Internet, with much stricter QoS requirement.

As a basic study for the future work of this thesis, a project related to this study is ongoing. In this project, the author's CAC model is expanded to IP telephony networks. In VoIP communications, the bandwidth of each session frequently varies due to the IP packet multiplexing effect. [67] proposes a CAC method to determine the number of maximum VoIP sessions by using expected value of each accommodated sessions in an $M/M/s/s$ loss system. The packet queue in burst is also modeled as an $MMPP/M/1/K$ system, and the packet loss probability is also calculated. This CAC method maximizes the number of maximum VoIP sessions S so that the packet loss probability is lower than the required value p.

Moreover, another project related to Chapter 3 is ongoing. This project proposes a new delay system that permits some blocked general calls to wait in the waiting queue on a SIP server. In the IP network, a limited number of VoIP sessions can be accommodated and have their QoS guaranteed. However, this number is expected to be less than the limit of SIP servers. The remaining communication resource of SIP servers will be utilized as the waiting queue of awaiting general calls. When all the telephone lines in the SIP server are being used, an incoming general call is enqueued into the waiting queue, and a general call request on the head of the queue will be dequeued and accommodated into the system, after an ongoing general call is terminated. By this waiting queue, general call users are guaranteed for their communication opportunity, once their call are enqueued into the waiting queue. The ongoing study [68] shows that introducing the waiting queue reduces the call-blocking rate of general calls, even though the queue length is short.

Finally, for guaranteeing fairness for general call users, some general call users who retry calling after being blocked or having their calls terminated, are taken into account in another project. These general recalls should be distinguished from new general calls and should have a shorter holding time limit than new calls. In the ongoing study [69], the blocking of new general calls is defined as the cumulative result of several call trials. Then, this study proposes a method to configure a threshold in the trunk reservation, which can accommodate more repeated calls by permissive configuration of the allowable blocking rate of new general calls. Computer simulation of the proposed method shows that more repeated calls can be accommodated, satisfying the allowable blocking rate of new general calls, as the system is more permissible for call retries.

Based on the results of these ongoing studies, the proposed CAC methods will be implemented on a real network environment. Since there are some open source project of core network entities, such as Open IMS Core [70], OpenEPC [71] and Open5GCore [72], SDN/NFV-based telecommunication networks will be emulated on a virtualized infrastructure and the QoS-guaranteed accommodation of both real voice and data traffic will be confirmed.

Nevertheless, there are several problems yet to be studied for realizing a truly reliable communication environment in the IP-based telecommunications networks. In the field of wireless networks, the discussion on the cross-layer approach has been active in the recent decade [73]. When a large number of IoT/M2M devices are connected to the Internet in the era of cyber communication society, characteristics and congestion factors on the physical layer, such as frequency band allocation and mutual interference between devices, cannot be ignored to accommodate their high speed mobile traffic appropriately. In addition, the security issue of ongoing data traffic should be considered. If the command of IoT/M2M devices such as smart home / smart city facilities, connected cars and

connected medical platforms is seized by cyber attackers, hijacked communications will cause fatal accidents or even terrorisms in future cyberphysical systems. In order to facilitate our humans' lives in such IoT-based infrastructures, the telecommunication network architecture should be designed much more securely.

Furthermore, a mechanism of communication resource sharing should be studied to overcome the accommodation capability shortage in a virtualized core network entity. Recently, the demand of Mobile Edge Computing [74] has arisen for traffic load balancing and satisfying the delay sensitive communication traffic. To properly accommodate a large number of various communication traffic, not only the incoming traffic should be balanced according to each QoS requirement, but the accommodation and processing resource should also be shared among adjacent nodes for sudden traffic bursts.

Bibliography

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021." http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf, March 2017. (accessed Jan. 31, 2019)

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions", Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, Sep. 2013. doi: 10.1016/j.future.2013.01.010

[3] G. M. Lee, N. Crespi, J. K. Choi, and M. Boussard, "Internet of things," Evolution of Telecommunication Services, Springer, 2013, pp. 257-282. doi: 10.1007/978-3-642-41569-2_13

[4] 3GPP, "Service requirements for V2X services," TS 22.185, V15.0.0, July 2018.

[5] 3GPP, "Service requirements for enhanced V2X scenarios," TS 22.186, V15.3.0, July 2018.

[6] 3GPP, "Study on scenarios and requirements for next generation access technologies," TS 38.913, V15.0.0, Sept. 2018.

[7] K. Knightson, N. Morita and T. Towle, "NGN architecture: generic principles, functional architecture, and implementation," IEEE Communications Magazine, vol. 43, no. 10, pp. 49–56, Oct. 2005. doi: 10.1109/MCOM.2005.1522124

[8] ITU-T Rec. Y.2001, "General overview of NGN" Dec. 2014.

[9] "PSTN Migration: General Outlook," https://www.ntt-west.co.jp/news_e/1011/pdf/101102a_1.pdf, Nov. 2010. (accessed Jan. 31, 2019)

[10] T. Aoyama, "A new generation network: Beyond the Internet and NGN," IEEE Communications Magazine, vol. 47, no. 5, pp. 82–87, May 2009. doi: 10.1109/MCOM.2009.4939281

[11] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky and S. Uhlig, "Software-Defined Networking: A comprehensive survey," in Proceedings of the IEEE, vol. 103, no. 1, pp. 14-76, Jan. 2015. doi: 10.1109/JPROC.2014.2371999

[12] ETSI, "Network Functions Viratualisation (NFV); Architectural Framework," GS NFV 002, Dec. 2014.

[13] Ministry of Internal Affairs and Communications, "Maintaining Communications Capabilities during Major Natural Disasters and other Emergency Situations: Final Report," `http://www.soumu.go.jp/main_content/000146938.pdf`, Dec. 2011. (accessed Jan. 31, 2019)

[14] Y. Shibata, N. Uchida and N. Shiratori, "Analysis of and proposal for a disaster information network from experience of the Great East Japan Earthquake," IEEE Communications Magazine, vol. 52, no. 3, pp. 44–50, March 2014. doi: 10.1109/MCOM.2014.6766083

[15] M. Kobayashi, "Experience of infrastructure damage caused by the Great East Japan Earthquake and countermeasures against future disasters," IEEE Communications Magazine, vol. 52, no. 3, pp. 23–29, March 2014. doi: 10.1109/MCOM.2014.6766080

[16] ITU-T Rec. E. 107, "Emergency Telecommunications Service (ETS) and interconnection framework for national implementations of ETS," Feb. 2007.

[17] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," TS 23.401, Sept. 2016.

[18] 3GPP, "System Architecture for the 5G System," TS 23.501, June 2018.

[19] Y. Ran, "Considerations and suggestions on improvement of communication network disaster countermeasures after the Wenchuan Earthquake," IEEE Communications Magazine, vol. 49, no. 1, pp. 44–47, Jan. 2011. doi: 10.1109/MCOM.2011.5681013

[20] Y. Shibata, N. Uchida and N. Shiratori, "Problem analysis and improvement of disaster information network and system from experiences of the great East Japan

Earthquake," Proc. IEEE R10-HTC, pp. 188–193, Aug. 2013. doi: 10.1109/R10-HTC.2013.6669039

[21] J. S. Huang and Y. N. Lien, "Challenges of emergency communication network for disaster response," Proc. IEEE ICCS 2012, pp. 528–532, Nov. 2012. doi: 10.1109/ICCS.2012.6406204

[22] J. Ni, D. H. K. Tsang, S. Tatikonda and B. Bensaou, "Threshold and reservation based call admission control policies for multiservice resource-sharing systems," Proc. IEEE INFOCOM 2005, vol. 2, pp. 773–783, Mar. 2005. doi: 10.1109/INFCOM.2005.1498309

[23] P. McGregor, J. Szeto and F. Suraci, "Performance modeling of high probability call-completion features," Proc. 7th International Conference on Comp. and Commun. Networks, pp. 81–90, Oct. 1998. doi:10.1109/ICCCN.1998.998764

[24] S. Komorita, Y. Kitatsuji and H. Yokota, "Congestion-based automatic calling for improving call establishment in VoLTE," Proc. IEEE CCNC 2013, pp. 521–527, Jan. 2013. doi: 10.1109/CCNC.2013.6488493

[25] K. Okada, "Limiting the holding time in mobile cellular systems during heavy call demand periods in the aftermath of disasters," IEICE Trans. Fundamentals, vol. E85–A, no. 7, pp. 1451–1462, July 2002.

[26] K. Okada, "Limiting the holding time considering emergency calls in mobile cellular phone systems during disasters," IEICE Trans. Commun., vol. E89–B, no. 1, pp. 57–65, Jan. 2006. doi: 10.1093/ietcom/E89-B.1.57

[27] J. Zhou and C. Beard, "Comparison of combined preemption and queuing schemes for admission control in a cellular emergency network," Proc. IEEE WCNC 2006, pp. 122–128, Apr. 2006. doi: 10.1109/WCNC.2006.1683451

[28] J. Zhou and C. Beard, "Tunable preemption controls for a cellular emergency network," Proc. IEEE WCNC 2007, pp. 3647–3652, Mar. 2007. doi: 10.1109/WCNC.2007.668

[29] J. Zhou and C. Beard, "A controlled preemption scheme for emergency applications in cellular networks," IEEE Trans. Veh. Technol., vol. 58, Issue 7, pp. 3753–3764, Sept.2009. doi: 10.1109/TVT.2009.2014634

[30] K. Nakano, K. Kawamura, N. Karasawa, M. Sengoku and S. Shinoda, "Traffic Characteristics of Dynamic Channel Assignment under Non Uniform Traffic Distribution," Proc. VTC'97, vol. 3, pp. 1465–1469, May 1997. doi: 10.1109/VETEC.1997.605572

[31] N. Caceres, L. M. Romero, F. G. Benitez and J. M. del Castillo, "Traffic Flow Estimation Models Using Cellular Phone Data," IEEE Trans. Intell. Syst., vol. 13, no. 3, pp. 1430–1441, Sept. 2012. doi: 10.1109/TITS.2012.2189006

[32] K. Tanabe, S. Miyata, K. Baba and K. Yamaoka, "Why are so many lines still reserved for emergency telephone calls in emergency situations?," Proc. NETWORKS 2014, pp.1–6, Sept. 2014. doi: 10.1109/NETWKS.2014.6959249

[33] K. Tanabe, S. Miyata, K. Baba and K. Yamaoka, "Threshold configuration of emergency trunk reservation considering traffic intensity for accepting more general telephone calls," Proc. RNDM 2014, pp. 165–170, Nov. 2014. doi: 10.1109/RNDM.2014.7014947

[34] K. Tanabe, S. Miyata and K. Yamaoka, "Accepting more general telephone calls in emergency situations by limiting general call holding time under trunk reservation control," Proc. IEEE CQR 2014, May 2014. (On USB Flash drive) doi: 10.1109/CQR.2014.7152634

[35] S. Y. Lien, K. C. Chen and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," IEEE Communications Magazine, vol. 49, no. 4, pp. 66–74, April 2011. doi: 10.1109/MCOM.2011.5741148

[36] 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 Tutorials, vol. 17, no. 4, pp. 2347–2376, 4th Quart. 2015. doi: 10.1109/COMST.2015.2444095

[37] N. Khan, J. Mišić and V. B. Mišić, "VM2M: An overlay network to support vehicular traffic over LTE," Proc. IWCMC 2016, pp. 13–18, Sept. 2016. doi: 10.1109/IWCMC.2016.7577026

[38] N. Khan, J. Mišić and V. Mišić, "Priority based VM2M communications over LTE," Proc. IEEE CAMAD 2016, pp. 177–182, Oct. 2016. doi: 10.1109/CAMAD.2016.7790354

[39] S. Wahle, T. Magedanz and F. Schulze, "The OpenMTC framework – M2M solutions for smart cities and the internet of things," Proc. IEEE WoWMoM 2012, pp. 1–3, June 2012. doi: 10.1109/WoWMoM.2012.6263737

[40] M. Corici, H. Coskun, A. Elmangoush, A. Kurniawan, T. Mao, T. Magedanz and S. Wahle, "OpenMTC: Prototyping Machine Type communication in carrier grade operator networks," in IEEE GLOBECOM Workshops, pp. 1735–1740, Dec. 2012. doi: 10.1109/GLOCOMW.2012.6477847

[41] M. Mazzola, G. Schaaf, F. Niewels and T. Kurner, "Exploration of centralized Car2X-systems over LTE," Proc. IEEE VTC 2015-Spring, pp. 1–5, May 2015. doi: 10.1109/VTCSpring.2015.7145836

[42] U. Pützschler, "LTE and Car2x: Connected cars on the way to 5G." `https://www.cambridgewireless.co.uk/media/uploads/resources/Mobile\%20Broadband\%20Group/06.04.16/MobileBroadband-06.04.16-Nokia-Uwe_Putzschler.pdf`, April 2016. (accessed Jan. 31, 2019)

[43] M. R. T. Hossain, M. A. Shahjalal and N.F. Nuri, "Design of an IoT based autonomous vehicle with the aid of computer vision," Proc. ECCE 2017, pp. 752–756, Feb. 2017. doi: 10.1109/ECACE.2017.7913003

[44] M. Gerla, E.K. Lee, G. Pau and U. Lee, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds," Proc. IEEE WF-IoT 2014, pp. 241–246, March 2014. doi: 10.1109/WF-IoT.2014.6803166

[45] J. Khoury, R. Ramanathan, D. McCloskey, R. Smith and T. Campbell, "RadarMAC: Mitigating radar interference in self-driving cars," Proc. IEEE SECON 2016, pp. 1–9, June 2016. doi: 10.1109/SAHCN.2016.7733011

[46] S. Jeon, D. Corujo and R. L. Aguiar, "Virtualised EPC for on-demand mobile traffic offloading in 5G environments," Proc. IEEE CSCN 2015, pp. 275–281, Oct. 2015. doi: 10.1109/CSCN.2015.7390457

[47] Verizon, "SLA for Internet dedicated services." `https://enterprise.verizon.com/content/dam/resources/support/2017/sla_global-sla-for-internet-dedicated-services.pdf`. (accesed Jan. 31, 2019)

[48] 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. doi: 10.1109/MNET.2014.6963800

[49] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis and T. Magedanz, "EASE: EPC as a service to ease mobile core network deployment over cloud," IEEE Network, vol. 29, no. 2, pp. 78–88, March 2015. doi: 10.1109/MNET.2015.7064907

[50] A. Gonzalez, P. Grønsund, K. Mahmood, B. Helvik, P. Heegaard and G. Nencioni, "Service Availability in the NFV Virtualized Evolved Packet Core," Proc. IEEE GLOBECOM 2015, pp. 1–6, Dec. 2015. doi: 10.1109/GLOCOM.2015.7417254

[51] R. Martínez, A. Mayoral, R. Vilalta, R. Casellas, R. Munoz, S. Pachnicke, T. Szyrkowiec and A. Autenrieth, "Integrated SDN/NFV orchestration for the dynamic deployment of mobile virtual backhaul networks over a multilayer (packet/optical) aggregation infrastructure," IEEE/OSA J. Opt. Commun. Netw., vol. 9, no. 2, pp. A135–A142, Feb. 2017. doi: 10.1364/JOCN.9.00A135

[52] Mavenir, "Virtual Evolved Packet Core (vEPC)." `https://mavenir.com/solutions/cost-reduction/virtualized-evolved-packet-core-vepc`. (accessed Jan. 31, 2019)

[53] NEC, "vEPC Solutions : Network Functions Virtualization (NFV)." `http://www.nec.com/en/global/solutions/tcs/vepc/index.html`. (accessed Jan. 31, 2019)

[54] N. Nikaein and S. Krco, "Latency for real-time machine-to-machine communication in LTE-based system architecture," Proc. European Wireless 2011, pp. 263–268, April 2011.

[55] 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," Proc. IEEE SDN4FNS 2013, pp. 1–7, Nov. 2013. doi: 10.1109/SDN4FNS.2013.6702532

[56] H. M. Hussien and H. A. Elsayed, "Performance evaluation of virtualized LTE-EPC data plane with MPLS core using PPBP machine-to-machine traffic," IEICE Trans. Commun., vol. E99-B, no. 2, pp. 326–336, Feb. 2016. doi: 10.1587/transcom.2015ITP0013

[57] G. Hasegawa and M. Murata, "Joint bearer aggregation and control-data plane separation in LTE EPC for increasing M2M communication capacity," Proc. IEEE GLOBECOM 2015, pp. 1–6, Dec. 2015. doi: 10.1109/GLOCOM.2015.7417359

[58] J. Prados, J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado and J. Lopez-Soler, "Modeling and dimensioning of a virtualized MME for 5G mobile networks," IEEE Trans. Veh. Technol., vol. 66, no. 5, pp.4383–4395, May 2017. doi: 10.1109/TVT.2016.2608942

[59] 3GPP, "Network architecture," TS 23.002, Sept. 2016.

[60] A. Ksentini, T. Taleb, X. Ge and H. Honglin, "Congestion-aware MTC device triggering," Proc. IEEE ICC 2014, pp. 294–298, June 2014. doi: 10.1109/ICC.2014.6883334

[61] K. Tanabe, H. Nakayama, T. Hayashi and K. Yamaoka, "An optimal resource assignment for C/D-plane virtualized mobile core networks," Proc. IEEE ICC 2017, pp. 1–6, May 2017. doi: 10.1109/ICC.2017.7997202

[62] L. Gimpelson, "Analysis of mixtures of wide- and narrow-band traffic," IEEE Transactions on Communication Technology, vol. 13, no. 3, pp. 258–266, Sept. 1965. doi: 10.1109/TCOM.1965.1089121

[63] K. Tanabe, H. Nakayama, T. Hayashi and K. Yamaoka, "A study on resource granularity of vEPC optimal resource assignment," Proc. IEEE/ACM IWQoS 2017, pp. 1–2, June 2017. doi: 10.1109/IWQoS.2017.7969176

[64] NEC, "Advanced Mobile-services Switching Centre MX5380 Series." `http://jpn.nec.com/tcs/amsc/`. (accessed Jan. 31, 2019)

[65] B. Hirschman, P. Mehta, K.B. Ramia, A.S. Rajan, E. Dylag, A. Singh and M. Mcdonald, "High-performance evolved packet core signaling and bearer processing on general-purpose processors," IEEE Network, vol. 29, no. 3, pp. 6–14, May 2015. doi: 10.1109/MNET.2015.7113219

[66] M. Corici, I. Gheorghe-Pop, E. Cau, A. A. Corici and T. Magedanz, "A benchmarking methodology for virtualized packet core implementations," Proc. IEEE CSCN 2016, pp. 1–6, Oct. 2016. doi: 10.1109/CSCN.2016.7785156

[67] Ryota Murakami, Kazuki Tanabe, Ken-ichi Baba and Katsunori Yamaoka, "VoIP admission control to increase QoS-guaranteed sessions by considering state probability," Proc. IEEE PIMRC 2018, Sept. 2018. doi: 10.1109/PIMRC.2018.8580986

[68] Kenta Kawai, Kazuki Tanabe, Katsunori Yamaoka and Ken-ichi Baba, “Emergency trunk reservation control using waiting queue for accommodating more general calls,” Proc. ICNC 2019. (to be published)

[69] Takumi Kada, Takuya Okamoto, Kazuki Tanabe and Katsunori Yamaoka, “Improvement of accommodation efficiency in emergency communication network by introducing permissible number of transmission trials,” IEICE Technical Report IN2016-147, vol.116, no. 485, pp. 299–304, March 2017.

[70] Open IMS Core, `https://www.openimscore.com/`. (accessed Jan. 31, 2019)

[71] OpenEPC, `https://www.openepc.com/`. (accessed Jan. 31, 2019)

[72] Open5GCore, `https://www.open5gcore.org/`. (accessed Jan. 31, 2019)

[73] V. Srivastava and M. Motani, “Cross-layer design: a survey and the road ahead,” IEEE Communications Magazine, vol. 43, no. 12, pp. 112–119, Dec. 2005. doi: 10.1109/MCOM.2005.1561928

[74] W. Shi, J. Cao, Q. Zhang, Y. Li and L. Xu, “Edge computing: vision and challenges,” in IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, Oct. 2016. doi: 10.1109/JIOT.2016.2579198

Achivements

Fellowships

- 2016–2019 Japan Society for the Promotion of Science (JSPS) Research Fellow DC1

Awards and Honors

- IEEE/ACM International Symposium on Quality of Service (IWQoS 2017) Best Demo Award, June 2017.

Publications forming part of the thesis

A. Journal papers

[A-1] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi and Katsunori Yamaoka, "vEPC Optimal Resource Assignment Method for Accommodating M2M Communications," IEICE Transactions on Communications, vol. E101–B, no. 3, pp. 637–647, Mar. 2018.

[A-2] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, "Threshold Relaxation and Holding Time Limitation Method for Accepting More General Calls under Emergency Trunk Reservation," IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, vol. E99–A, no. 8, pp. 1518–1528, Aug. 2016.

B. International conferences (With review)

[B-1] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, "A Study on Resource Granularity of vEPC Optimal Resource Assignment," in Proc. of IEEE/ACM IWQoS 2017, June 2017.

[B-2] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “An Optimal Resource Assignment for C/D-plane Virtualized Mobile Core Networks,” in Proc. of IEEE ICC 2017, May 2017.

[B-3] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, “Threshold configuration of emergency trunk reservation considering traffic intensity for accepting more general telephone calls,” in Proc. RNDM 2014, pp. 165–170, Nov. 2014.

[B-4] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, “Why Are So Many Lines Still Reserved for Emergency Telephone Calls in Emergency Situations?,” in Proc. of Networks 2014, Sept. 2014.

[B-5] Kazuki Tanabe, Sumiko Miyata and Katsunori Yamaoka, “Accepting More General Telephone Calls in Emergency Situations by Limiting General Call Holding Time under Trunk Reservation Control,” in Proc. of IEEE CQR 2014, May 2014. (on USB Flash Drive)

C. Domestic conferences

[C-1] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “Reduction Effect of vEPC Optimal Resource Assignment Calculation Time by Resource Granularity,” in Proc. of IEICE Society Conference B-7-41, Sept. 2017.

[C-2] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “Relationship between Resource Granularity of vEPC Server and Optimal Resource Assignment,” IEICE Technical Report ICM2017-11, vol.117, no. 114, pp. 39–44, Jul. 2017.

[C-3] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “Characteristic of Heterogeneous Traffic Accommodation under Optimal Resource Assignment of vEPC,” in Proc. of IEICE General Conference B-7-7, March 2017.

[C-4] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “Effects of Mobile Traffic on vEPC Optimal Resource Assignment,” IEICE Technical Report IN2016-151, vol.116, no. 485, pp. 323–328, March 2017.

[C-5] Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “A Study on Optimal Resource Assignment for C/D-plane Virtualization of vEPC,” IEICE Technical Report ICM2016-34, vol.116, no. 324, pp. 55–60, Nov. 2016.

[C-6] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Ketsunori Yamaoka, “The Effectiveness of Accepting New General Calls by Rejecting General Retry Calls under Emergency Trunk Reservation Control,” in Proc. of IEICE General Conference B-7-65, March 2016.

[C-7] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, “Configuring Appropriate Threshold Considering Arriving Traffic Intensity in Emergency Situations,” in Proc. of IEICE General Conference B-7-12, March 2015.

[C-8] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, “A Threshold Configuration Method of Emergency Trunk Reservation Control Considering Arriving Traffic Intensity,” IEICE Technical Report CQ2014-85, vol.114, no. 298, pp. 83–88, Nov. 2014.

[C-9] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, “The Effectiveness of Threshold Relaxation Trunk Reservation Control,” in Proc. of IEICE Society Conference B-7-31, Sept. 2014.

[C-10] Kazuki Tanabe, Sumiko Miyata, Ken-ichi Baba, and Katsunori Yamaoka, “A Threshold Relaxation Method for Accepting More General Telephone Calls in Emergency Situation,” IEICE Technical Report MoNA2014-33, vol.114, no. 210, pp. 13–18, Sept. 2014.

[C-11] Kazuki Tanabe, Sumiko Miyata, and Katsunori Yamaoka, “The Efficiency of Limiting Holding Time of General Telephone Calls under Emergency Trunk Reservation Control,” in Proc. of IEICE General Conference B-7-90, March 2014.

[C-12] Kazuki Tanabe, Sumiko Miyata, and Katsunori Yamaoka, “A Study of Limiting Holding Time of General Telephone Calls and Threshold Control under Emergency Condition,” IEICE Technical Report IN2013-109, vol.113, no. 363, pp. 59–64, Dec. 2013.

Publications relevant to the thesis but not forming part of it

A. Journal papers

[A-1] Takuya Kosugiyama, Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi and Katsunori Yamaoka, “A Flow Aggregation Method under Allowable Delay Limitation in SDN,” IEICE Transactions on Communications, vol. E101–B, no. 3, pp. 795–804, March 2018.

B. International conferences (With review)

[B-1] Kenta Kawai, Kazuki Tanabe, Katsunori Yamaoka and Ken-ichi Baba, "Emergency Trunk Reservation Control Using Waiting Queue for Accommodating More General Calls," in Proc. of ICNC 2019. (to be published)

[B-2] Ryota Murakami, Kazuki Tanabe, Ken-ichi Baba and Katsunori Yamaoka, "VoIP Admission Control to Increase QoS-Guaranteed Sessions by Considering State Probability," in Proc. of IEEE PIMRC 2018, Sept. 2018.

[B-3] Takuya Kosugiyama, Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi and Katsunori Yamaoka, "A Flow Aggregation Method Based on End-to-End Delay in SDN," in Proc. of IEEE ICC 2017, May 2017.

C. Domestic conferences

[C-1] Kenta Kawai, Kazuki Tanabe, Katsunori Yamaoka, and Ken-ichi Baba, "Threshold Setting Method by Using Waiting Queue for General Calls under Emergency Trunk Reservation Control," in Proc. of IEICE Society Conference B-7-15, Sept. 2018.

[C-2] Ryota Murakami, Kazuki Tanabe, Ken-ichi Baba, and Katsunori Yamaoka, "A Method to Determine QoS-Guaranteed Connections with State Probability in VoIP Admission Control," in Proc. of IEICE General Conference B-7-26, March 2018.

[C-3] Kenta Kawai, Kazuki Tanabe, Katsunori Yamaoka, and Ken-ichi Baba, "Effect of Blocking Reduction by Using Waiting Queue for General Calls under Emergency Trunk Reservation Control," in Proc. of IEICE General Conference B-7-25, March 2018.

[C-4] Ryota Murakami, Kazuki Tanabe, Ken-ichi Baba, and Katsunori Yamaoka, "VoIP Admission Control for Guaranteeing QoS with State Probability," IEICE Technical Report IN2017-144, vol.117, no. 460, pp. 327-332, March 2018.

[C-5] Kenta Kawai, Kazuki Tanabe, Katsunori Yamaoka, and Ken-ichi Baba, "Emergency Trunk Reservation Control Using Waiting Queue for General Calls," IEICE Technical Report IN2017-116, vol.117, no. 460, pp. 159–164, March 2018.

[C-6] Takuya Kosugiyama, Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, "A Reduction Effect on Number of Flows by Aggregating Flows Satisfying Allowable Delay in SDN," in Proc. of IEICE General Conference B-7-1, March 2017.

[C-7] Takumi Kada, Takuya Okamoto, Kazuki Tanabe, and Katsunori Yamaoka, “Improvement of Accommodation Efficiency in Emergency Communication Network by Introducing Permissible Number of Transmission Trials,” IEICE Technical Report IN2016-147, vol.116, no. 485, pp. 299–304, March 2017.

[C-8] Takuya Kosugiyama, Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “A Characteristic of Flow Aggregation under End-to-End Allowable Delay Limitation in SDN,” IEICE Technical Report IN2016-124, vol.116, no. 485, pp. 163–168, March 2017.

[C-9] Takuya Kosugiyama, Kazuki Tanabe, Hiroki Nakayama, Tsunemasa Hayashi, and Katsunori Yamaoka, “A Flow Aggregation Method Based on End-to-End Delay in SDN,” IEICE Technical Report ICM2016-33, vol.116, no. 324, pp. 49–54, Nov. 2016.

[C-10] Kazuki Tanabe, Masaya Endo, Hiroyuki Minami, and Katsunori Yamaoka, “A Study on Number of Accommodated VoIP Sessions with QoS Satisfaction,” in Proc. of IEICE Society Conference B-7-19, Sept. 2016.