A Study on Distributed Mobility Management Scheme in Mobile Networks

Submitted to

Graduate School of Information Science and Technology

Osaka University

January 2017

Hua YANG

List of publication

Journal papers

1. Hua Yang, Naoki Wakamiya, Masayuki Murata, Takanori Iwai, and Satoru Yamano, "Autonomous and distributed mobility management in mobile core networks," *Wireless Networks*, DOI: 10.1007/s11276-016-1274-3, pp. 1–18, April 2016.

International Conference Papers

1. Hua Yang, Naoki Wakamiya, Masayuki Murata, Takanori Iwai, and Satoru Yamano, "An autonomous and distributed mobility management scheme in mobile core networks," in *Proceedings of 9th EAI International Conference on Bio-inspired Information and Communications Technologies (BICT 2015)*, pp. 35–42, December 2015.

2. Hua Yang, and Naoki Wakamiya, "A distributed mobility management scheme in flat mobile architecture," in *Proceedings of the 31st International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2016)*, pp. 29–32, July 2016.

Oral Presentations

1. Hua Yang, and Naoki Wakamiya, "A study on autonomous and distributed mobility management scheme in mobile networks – Proposal and Evaluation of Mobility Management Scheme based on Attractor Selection Model –," *ITRC-NWGN 2014*, Ako City, Hyogo, August 2014.

2. Hua Yang, and Naoki Wakamiya, “A study on autonomous and distributed mobility management scheme in mobile networks – proposal and evaluation of mobility management scheme based on attractor selection model –,” *Technical Report of IEICE (IN2014-90)*, vol. 114, no. 307, pp. 51–56, November 2014.

3. Hua Yang, and Naoki Wakamiya, “Autonomous and distributed mobility management scheme in mobile networks,” *ITRC-NWGN 2015*, Kobe City, Hyogo, September 2015 (in Japanese).

4. Hua Yang, and Naoki Wakamiya, “Proposal of self-organized mobility management server selection in mobile networks,” *ITRC-NWGN 2016*, Wakayama City, Wakayama, September 2016 (in Japanese).

Preface

Recently, with proliferation of mobile terminal devices such as multifunction smart phones and portable computers as well as fast development of IoT (Internet of Things) and M2M (Machine-to-Machine) technologies, mobile and wireless communication systems are experiencing challenges of exploding mobile data traffic, considerable number of devices, and heterogeneous applications. On the other hand, as Average Revenue Per User (ARPU) of M2M terminal devices is much lower than cellular phones, it is required to reduce the communication and management cost in mobile networks to accommodate considerable number of M2M terminal devices. However, the current mobile network architecture which provides all terminals with uniform services such as mobility management by centralized control cannot satisfy those emerging requirements.

The current 3.9G LTE/EPC (Long Term Evolution / Evolved Packet Core) network adopts the centralized architecture where U-plane (User Plane) has a functional and structural hierarchy that consists of a PGW (Packet Data Network Gateway) providing connectivity from UEs to external networks, SGWs (Serving Gateways) working as a local mobility anchor point for intra-system handovers, and base stations called eNBs (evolved Node Bs). In addition, an MME (Mobility Management Entity) usually deployed together with each SGW performs C-plane mobility management for all UEs in the corresponding TA (Tracking Area). As a corollary of centralized control, a problem of concentration of traffic and load on these nodes arises. In addition, communication between UEs and these nodes, which are distant from UEs, consumes bandwidth and introduces response

delay.

To tackle this problem, there have been several research activities on distributed mobile core network architecture such as DMM (Distributed Mobility Management). Recent years, partial DMM is a hot topic which adopts a flat architecture and shortens the distance from gateways to UEs by anchoring U-plane traffic closer to user devices. Regarding full DMM, there are only some blueprints for extending the partial DMM to full DMM and they are defective and immature. In addition, a software distributed architecture of DMME (Distributed Mobility Management Entity) was proposed to implement distributed mobility management in the C-plane. However, servers specifically dedicated to distribution of C-plane tasks need to be pre-allocated and load balancing among servers is not considered.

In this thesis, first, we introduce a novel architecture of distributed mobility management and an autonomous and adaptive mobility management scheme which distributes mobility management functions of the LTE/EPC MME on nodes in a mobile core network and allocate them to UEs (User Equipments) in accordance with their mobility characteristics and a management policy, which contributes to improvement of the performance of mobile network systems by considerably reducing the C-plane communication and management cost. Specifically, we propose a distributed mobility management scheme which can be realized on the current hierarchical architecture of 3.9G LTE/EPC (Long Term Evolution / Evolved Packet Core) networks.

In order to mitigate concentration of control traffic at MMEs, we propose a virtualized MME, namely, ADMME (Autonomous Distributed Mobility Management Entity) which can be configured at any nodes of a PGW, SGWs and eNBs by virtualization technology. Taking into account mobility characteristics of a UE, an appropriate ADMME is dynamically allocated. When an allocated ADMME called current ADMME receives a C-Plane request from a UE, it determines an ADMME which succeeds management of the requesting UE basing on the information it collects

if the current ADMME is not appropriate any more. For dynamic and adaptive selection of an appropriate ADMME with multiple performance measures such as delay, load, and C-plane signalling overhead, we adopt a biologically-inspired adaptation algorithm called attractor selection, which is a mathematical model of adaptive behavior of biological systems to dynamically changing environmental conditions. In the attractor selection model, a dynamic system finds a solution to maximize a scalar, called activity, which expresses the goodness of the current solution for the current conditions. In our scheme, by defining the activity based on multiple performance measures, each ADMME autonomously makes a decision on delegation of mobility management. Through simulation experiments, we confirmed that our proposal could accomplish lower delay, higher load balancing, and lower C-plane overhead comparing to other methods including the current standard under a variety of mobility scenarios.

However, because of the current hierarchical LTE/EPC architecture, the control messages are sent from UEs along the hierarchy, e.g. from an eNB to an SGW, then to a PGW, and finally to the MME at another SGW. As a result, the system performance has notable limitations on reduction of response delay and C-plane overhead. Therefore, next we design a novel flat mobile network architecture. More specifically, a mobile network is comprised of a RO (Regional Office) corresponding to a central office of PGW, LOs (Local Offices) serving local areas, and APs (Access Points) corresponding to eNBs. Moreover, MAs (Mobility Anchors) used for U-plane anchor points are deployed in LOs. For the sake of more flexible and dynamic distribution and allocation of C-plane management tasks, we separate C-plane mobility management tasks of the LTE/EPC MME into two parts, i.e. mobility management functions and storage of context information of UEs. Furthermore, they are virtualized as VDMMEs (virtualized and distributed mobility management entities) and DMDs (distributed mobility databases), and distributed over a mobile core network. Then, an appropriate pair of those virtualized servers is allocated or launched at APs, LOs, or a RO

for a UE taking into account its mobility characteristics, response delay, and load status of servers, and C-plane overhead.

For dynamic and autonomous allocations of VDMMEs and DMDs in accordance with UE characteristics and network status in a flat mobile network, we adopt the response threshold model which is a mathematical model derived from self-organized division of labor in social insects. In the model, each individual autonomously and stochastically determines whether to perform a task or not depending on the stimulus or demand of the task and its inherent threshold or hesitation against the task. In our scheme, we extend the model to incorporate multiple factors, such as delay, number of candidates, and C-plane overhead. We confirmed that our proposal can mitigate delay to the half or even one-fifth and C-plane overhead by more than 41.7% from a partial distributed mobility management scheme proposed in other literature through simulation experiments under three mobility scenarios.

In this thesis, we first proposed an architecture and a scheme for distributed mobility management conforming to the current 3.9G LTE/EPC standards and then moved to a flat architecture for further flexible and adaptive mobility management leaving away from the current standards. Our main idea is to adopt bio-inspired algorithms to accomplish autonomous decision-making by individual servers and mitigate C-plane overhead. Based on obtained results we confirmed superiority of our proposals to other existing methods, which open up new perspectives on architecture design of a sustainable mobile core network.

Acknowledgments

This thesis could not have been accomplished without the assistance of many people, and I would like to acknowledge all of them.

First of all, I would like to express my great gratitude to my superviser, Professor Naoki Wakamiya, for his continuous support and insightful comments throughout my Ph.D, for his patience, motivation, and immense knowledge. His generous guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I am heartily grateful to the members of my thesis committee, Professor Hideo Matsuda, Professor Hiroshi Shimizu, and Professor Taro Maeda of Graduate School of Information Science and Technology, Osaka University, for their multilateral reviews, perceptive comments and encouragement.

Also, I would like to express my sincere appreciation for Professor Yoshichika Yoshioka of Biofunctional Imaging Lab., Immunology Frontier Research Center (IFReC), Osaka University and Professor Masayuki Murata of Graduate School of Information Science and Technology, Osaka University. Without their continuous advices and supports, I would not have entered the Ph.D. program. My sincere thanks also goes to Dr. Takanori Iwai and Dr. Satoru Yamano of Cloud System Research Labs, NEC Corporation for their support from the prospective of mobility management architecture.

Furthermore, I must acknowledge Professor Masayuki Murata, Associate Professor Junnosuke Teramae, Assistant Professor Masahumi Hashimoto, Assistant Professor Yuichi Ohshita, Specially Appointed Assistant Professor Naomi Kuze, Specially Appointed Assistant Professor Lu Chen of Graduate School of Information Science and Technology, Osaka University, and Dr. Kenji Leibnitz of National Institute of Information and Communications Technology, for their valuable comments and suggestions on my study.

I express my appreciation to all of past and present colleagues, friends, and secretaries of the Bio-system analysis Laboratory, Graduate School of Information Science and Technology, Osaka University.

I cannot conclude my acknowledgement without expressing my thanks to my parents and family. Thank you for your giving me invaluable supports throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Recently, with proliferation of mobile terminal devices such as multifunction smart phones and portable computers as well as fast development of IoT (Internet of Things) and M2M (Machine to Machine) technologies, mobile and wireless communication networks have been becoming ubiquitous and indispensable in our daily lives, which leads to exploding mobile data traffic, considerable number of devices, and heterogeneous applications [1] [2]. On the other hand, despite the considerable number of M2M devices, the amount of data traffic and communication frequency between M2M devices are considerably lower than cellular phones. Accordingly, ARPU (Average Revenue Per User or Unit) of M2M devices is much lower comparing to cellular phones [3]. However, they require almost the same control and management cost in the current mobile network architecture, which provides all terminals with uniform services such as mobility management by centralized control independently of their communication and mobility characteristics. Consequently, it is required to reduce the communication and management cost in mobile networks to accommodate considerable number of M2M terminal devices. The centralized mobility management architecture in a current mobile core network cannot satisfy those emerging requirements.

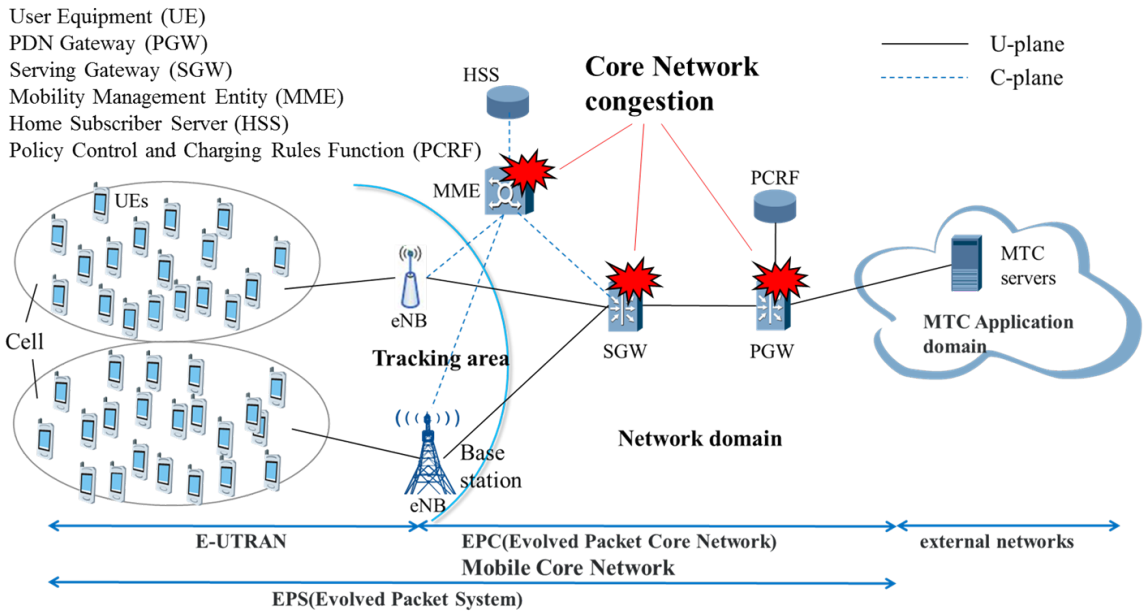

Figure 1.1: Centralized architecture of 3.9G LTE/EPC network. U-plane has a functional and structural hierarchy that consists of a PGW providing connectivity from UEs to external networks, SGWs working as a local mobility anchor point for intra-system handovers, and base stations called eNBs. In C-plane, an MME usually deployed together with each SGW performs mobility management for all UEs in the corresponding TA.

As Fig. 1.1 illustrates, the 3.9G LTE/EPC (Long Term Evolution / Evolved Packet Core) network adopts a centralized architecture, where U-plane (User Plane) has a functional and structural hierarchy that consists of a PGW (Packet Data Network Gateway) providing connectivity from UEs (User Equipment) to external networks, SGWs (Serving Gateways) working as a local mobility anchor point for intra-system handovers, and base stations called eNBs (evolved Node Bs) [4]. In addition, an MME (Mobility Management Entity) usually deployed together with each SGW performs C-plane mobility management for all UEs in the corresponding TA (Tracking Area). Another two auxiliary nodes in a mobile core network are Home Subscription Server (HSS), and Policy Control and Rules Function (PCRF), which provide additional network functions like management of subscription-related information, and policy enforcement and flow-based charging

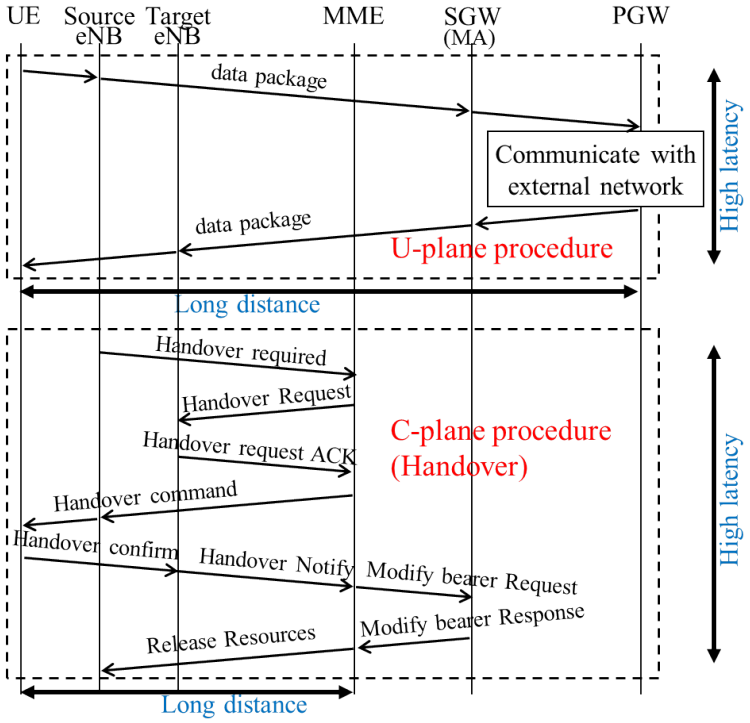

Figure 1.2: Example for U-plane and C-plane procedures when there is a Handover. For U-plane, to communicate with an external network a UE needs to send a data package along the hierarchy, first to the nearest eNB, next to the connecting SGW, and finally to the external network via a PGW, additionally following a reverse path response message reaches the UE. Regarding Handover procedure in C-plane, MME communicates with eNBs for handover by a series of C-plane signalling messages and modify the bearer between eNBs and SGW.

function. We can easily find from Fig. 1.1 that with the substantial increase of number of UEs the centralized architecture in a current mobile core network would face critical problems such as traffic congestion in PGW/SGW and overload in MME. Furthermore, communication between UEs and these nodes, which are distant from UEs, leads to excessive bandwidth consumption and high latency in both of U-plane and C-plane as shown in Fig. 1.2. This figure gives a specific example of signalling procedures of U-plane and C-plane when there is a handover. As for U-plane, to communicate with an external network a UE needs to send a data package along the hierarchy, first to the nearest eNB(source eNB), next to the connecting SGW, and finally to the external network

via a PGW, additionally following a reverse path response message reaches the UE via a target eNB due to a handover, which causes high latency. Regarding Handover procedure in C-plane shown in Fig. 1.2, which is triggered when a connected UE enters into a new cell, eNBs and MME communicates with each other for handover by a series of C-plane signalling messages. After a UE hands over to the target eNB, the MME server sends a modify bearer request to the SGW to switch the bearer from the source eNB to the target eNB. Through this example we can easily know that although the size of a C-plane signalling message is much smaller comparing to a data package in U-plane, a long distance and frequent communication between eNBs and MMEs also results in high latency and bandwidth consumption.

To tackle this problem, a direction toward distributed network architecture has drawn a lot of attentions in industry, academia, and government, being led by for example METIS (Mobile and wireless communications Enablers for Twenty-twenty Information Society) of Europe [5] [6]. Recently DMM(Distributed Mobility Management) solution is a hot topic in IETF and 3GPP which adopts flat mobile network architecture [7] [8]. Mobility management may be distributed in different levels and it could be partial distributed or full distributed. A partial DMM shortens the distance from gateways to UEs by anchoring U-plane traffic closer to user devices. Regarding full DMM, there are only some blueprints for extending the partial DMM to full DMM which implements the distribution of mobility management in both C-plane and U-plane, and they are defective and immature. In addition, a software-based distributed architecture of DMME(Distributed Mobility Management Entity) was proposed to implement distributed mobility management in the C-plane [9]. However, servers specifically dedicated to distribution of C-plane tasks need to be pre-allocated and its allocation scheme is designed only for minimizing signalling latency in C-plane without considering other indexes such as load balancing among servers, and C-plane overhead. Therefore for highly flexible management of a mobile core

network, network virtualization technologies such as SDN (Software-Defined Networking) and NFV (Network Functions Virtualization) are considered to be incorporated [10]. Virtualization enables a high freedom of choice in topology and functional layout.

1.2 Objective

Based on these arguments, we first design a novel architecture of distributed mobility management and propose an autonomous and adaptive mobility management scheme which distributes mobility management tasks on nodes in a mobile core network and allocates them in accordance with mobility characteristics of UEs and a management policy. In order to mitigate concentration of control traffic at MMEs, we propose a virtualized MME, namely, ADMME (Autonomous Distributed Mobility Management Entity) which can be deployed at any nodes in mobile network, i.e. a PGW, SGWs, and eNBs, by virtualization technology. Taking into account mobility characteristics of a UE, an appropriate ADMME is dynamically allocated, which contributes to improve the performance of mobility management in mobile networks by considerably reducing C-plane communication and management cost. Specifically, we propose a distributed mobility management scheme which can be realized on the current hierarchical architecture of 3.9G LTE/EPC networks based on a bio-inspired adaptation algorithm named attractor selection, which is a mathematical model of adaptive behavior of biological systems to dynamically changing environmental conditions, to accomplish adaptive allocation of appropriate ADMMEs according to multiple objectives.

However, because of the current hierarchical LTE/EPC architecture, the control messages are sent from UEs along the hierarchy, e.g. an eNB to an SGW, then to a PGW, and finally to the MME at another SGW. As a result, the system performance has notable limitations on mitigation of response delay and C-plane overhead. For the sake of more flexible and dynamic distribution

and allocation of C-plane management tasks, we design a novel flat architecture, in which data traffic in U-plane is distributed among mobility anchors closer to UEs based on DMM technique, additionally, C-plane mobility management functions of the LTE/EPC MME are separated into two parts, i.e. virtualized and distributed mobility management entities (VDMMEs) and distributed mobility databases (DMDs) distributed over a mobile core network. VDMMEs and DMDs can be allocated or launched at any nodes of APs, LOs, or a RO for a UE by virtualized technology. Moreover, for dynamic and autonomous allocation of appropriate VDMMEs and DMDs for UEs according to their mobility characteristics, response delay, load status of servers, and C-plane overhead, we propose a mechanism and an algorithm based on the response threshold model which intrinsically combines autonomous, high scalability, adaptability and robustness.

1.3 Outline of Thesis

Autonomous and Distributed Mobility Management in Mobile Core Networks

As the configuration of this thesis shown in Fig. 1.3, in order to tackle the problems of centralized control in current 3.9G LTE/EPC networks, in Chapter 2, we first present a novel architecture of distributed mobility management in C-plane in the mobile core network, which employs virtualized mobility management entities called ADMMEs (Autonomous Distributed Mobility Management Entity). Note that ADMMEs can be configured at any nodes of a PGW, SGWs and eNBs by virtualization technology. In addition, to assign an appropriate ADMME to a UE in accordance with mobility characteristics of the UE and a management policy, we propose an autonomous and adaptive ADMME selection scheme, which can be realized on the current hierarchical architecture of 3.9G LTE/EPC networks. When an allocated ADMME called current ADMME receives a C-Plane request from a UE, it determines an ADMME which succeeds management of the requesting UE basing on the information it collects, e.g. mobility characteristics,

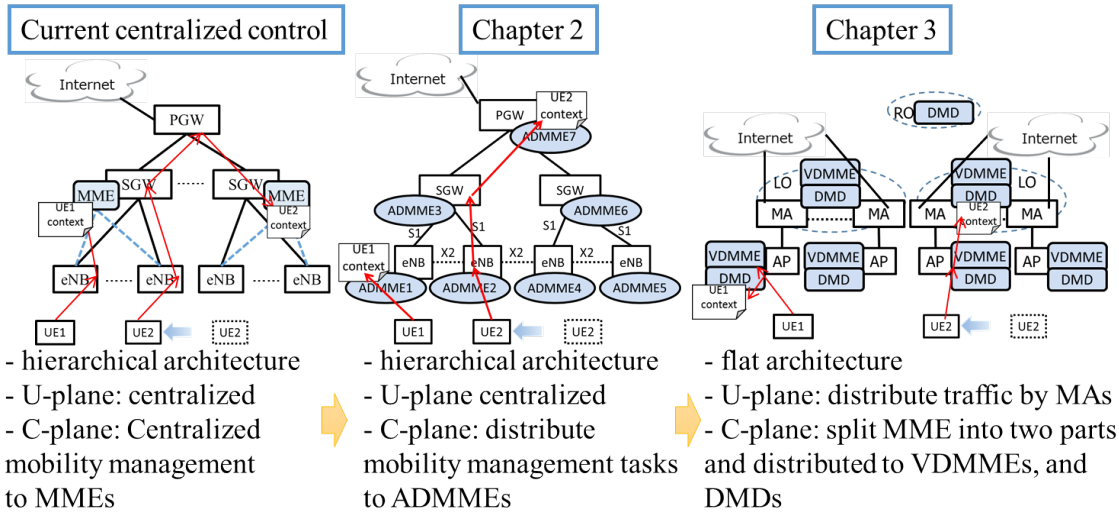

Figure 1.3: Configuration of the thesis. First, in Chapter 2, in order to tackle the problems of centralized control in current 3.9G LTE/EPC networks we propose an autonomous and distributed mobility management scheme which can be realized in current hierarchical architecture. To further reduce the response delay and C-plane overhead, then Chapter 3 moves from the distributed mobility management in hierarchical network architecture to flat architecture in which U-plane is also distributed.

response delay, and load status of servers, and C-plane overhead, if the current ADMME is not appropriate any more.

An ADMME selection algorithm should adaptively find a solution of a multi-objective optimization problem in accordance with the dynamically changing location and mobility of UEs and load status of ADMMEs. By general distributed algorithms, the relationship between servers and tasks becomes more complicated, correspondingly computation becomes more complicated. For example, S. Kraus et al. [11] presented some algorithms of distributed task allocation for cooperative agents. Although the distributed algorithm can achieve high performance close to the optimal, the complexity of the algorithm depends on the number of the agents and the number of types of tasks. Accordingly, if we adopt it to ADMME allocation, complexity of the algorithm increases a lot which causes substantial overhead of computation and communication, especially for

large-scale mobile networks accommodating numerous UEs. Additionally, a biologically-inspired algorithm can enable potential optimization of task allocation simply by utilizing fluctuations under dynamically changing constraints. Therefore, our selection algorithm adopts the bio-inspired attractor selection model, which is a mathematical model of adaptive behavior of biological systems to dynamically changing environmental conditions [12]. For its high adaptability, it is considered well-suited to the targeted problem than other metaheuristics and has been applied to a variety of network control.

In the attractor selection model, a dynamic system finds a solution to maximize a scalar, called activity, which expresses the goodness of the current solution for the current conditions. In our scheme, defines a scalar α called activity based on those performance measures, which expresses the goodness of current selection of an ADMME for a UE. When condition changes and the current ADMME becomes inappropriate, activity α decreases, accordingly, other ADMMEs will be selected randomly until an appropriate one is selected. Therefore, each ADMME autonomously makes a decision on delegation of mobility management for a UE to another to maximize the activity.

Simulations are run based on the server configuration, the number of UEs and the mobility model in realistic mobile core networks. Through simulation experiments, we confirmed that our proposal could accomplish lower delay, higher load balancing, and lower C-plane overhead comparing to other methods including the current standard under a variety of mobility scenarios. Particularly, it was verified that our proposal could accomplish more than 63% performance improvement comparing to the current method from viewpoints of delay, load balancing, and C-plane overhead under a dynamic mobility scenario.

A Distributed Mobility Management Scheme based on Response Threshold Model in Flat Mobile Architecture

As mentioned in 1.2, there are notable limitations on system performance, i.e. mitigation of response delay and C-plane overhead, due to the hierarchical architecture of the current standards. In order to increase the efficiency and further reduce the communication and management cost in both U-plane and C-plane in mobile networks, we move our research of distributed mobility management from hierarchical architecture to flat architecture of mobile networks in Chapter 3 as Fig. 1.3.

We first design a novel flat mobile network architecture based on DMM solution [7]. More specifically, a mobile network is comprised of a RO (Regional Office) corresponding to a central office of PGW, LOs (Local Offices) serving local areas, and APs (Access Points) corresponding to eNBs. Moreover, MAs (Mobility Anchors) used for U-plane anchor points are deployed in LOs. For the sake of more flexible and dynamic distribution and allocation of C-plane management tasks, we separate C-plane mobility management tasks of the LTE/EPC MME into two parts, i.e. mobility management functions and storage of context information of UEs. Furthermore, they are virtualized as VDMMEs (virtualized and distributed mobility management entities) and DMDs (distributed mobility databases), and distributed over a mobile core network. Then, an appropriate pair of those virtualized servers is allocated or launched at APs, LOs, or a RO for a UE taking into account its mobility characteristics, response delay, and load status of servers, and C-plane overhead.

We can easily find that in our flat architecture which has much different functional and structural hierarchy from the current 3.9G LTE/EPC network, there are more servers and most of them are deployed closer to UEs than hierarchical architecture. Sometimes multiple distributed servers are appropriate for serving one UE in flat architecture. Accordingly, the algorithm is required to accomplish autonomous, adaptive, and consistent decision-making by each individual server for serving UEs and realize multi-objectives optimization as a whole in our flat architecture. If we adopt other algorithms of distributed task allocation such as analyzed in [11] for VDMME and

DMD allocation, one member server received a request from a UE will determine to serve for the UE or transfer it to another member server based on the effort level for optimizing the whole system performance. As the algorithm become more complicated due to the large number of servers and tasks, the overhead of computation and communication is considerable for solving the problem of optimization of task allocation to VDMMEs and DMDs. In contrast, bio-inspired stochastic algorithms can achieve potential optimization with high computationally efficient. Therefore, we also adopt a bio-inspired algorithm for autonomous and adaptive allocation of appropriate VDMMEs and DMDs for UEs, which has high adaptability and flexibility.

Chapter 2 adopted the attractor selection model for autonomously and adaptively allocating appropriate ADMMEs for UEs in the current hierarchical architecture of 3.9G LTE/EPC networks. If the current ADMME is not appropriate, by using the attractor selection, a new appropriate ADMME cannot be selected by the current server instantly except after several random walks, while these random walks can be tolerated. If we use a response threshold model for ADMME selection, the convergence time is a little faster than attractor selection. However, its C-plane overhead of computation for the decision-making of each member server and communication between multiple selected servers and the current ADMMEs increase a lot due to long distances between servers in hierarchical architecture. Conversely, as there are much better solutions of allocation of VDMMEs and DMDs due to a large combination space in our flat architecture comparing to the allocation of appropriate ADMMEs, random walks before finally allocating an appropriate pair of those virtualized servers based on the attractor selection model leads to considerable convergence time and overhead. When we apply the response threshold model to implement dynamic and autonomous allocations of VDMMEs and DMDs, the appropriate servers are easier to be selected which contributes to accomplish a higher and faster convergence comparing to the attractor selection model. Therefore, to implement autonomous and adaptive decision-making on whether to become

a candidate or not by member VDMMEs and member DMDs, we adopt a biologically-inspired mathematical model of division of labors in insect societies [13].

In response threshold model, an idle individual becomes a worker for performing a task with a probability of $s^2/(s^2 + \theta^2)$, while a worker stops performing a task with probability q. The scalar s is called stimulus, corresponding to the demand that a task is performed. s increases with a certain constant rate while decreases proportionally to the number of workers. θ is an inherent threshold of the individual, meaning the degree of hesitation in performing a task. In our scheme, we extend the model to incorporate multiple factors, such as delay, number of candidates, and C-plane overhead. Each server on receiving a request from a UE autonomously and stochastically determines whether to become a candidate to supply the service of mobility management or not for the UE based on the response threshold model. Once the current DMD and an HSS receives the notifications from the candidate VDMMEs and DMDs, they will select one which satisfies the conditions respectively. If the allocated VDMME and DMD is appropriate for the UE, the stimulus for other members to become a candidate decreases. Therefore, the mobility management tasks of a UE can be allocated to a appropriate pair of VDMMEs and DMDs to reduce the stimulus for achieving a high performance.

Efficiency of our proposal is evaluated by the simulation experiments under the scenarios based on the reality. We confirmed that our proposal can mitigate delay to the half or even one-fifth and C-plane overhead by more than 41.7% from a partial distributed mobility management scheme proposed in other literature through simulation experiments under three mobility scenarios.

Chapter 2

Autonomous and Distributed Mobility Management in Mobile Core Networks

2.1 Introduction

In recent years, with proliferation of IoT and M2M technologies, the number of M2M terminal devices such as sensors and actuators are increasing exponentially [14], which further leads to the huge growth of mobile data traffic in a mobile communication network. Despite the large number of M2M devices, the amount of traffic and communication frequency between M2M devices are considerably lower than mobile phones. Therefore, ARPU of M2M devices is expected to be very low. However they require almost the same control and management cost in the current mobile network architecture, which offers uniform services such as mobility management.

The 3.9G LTE/EPC networks adopt the centralized architecture where an SGW handles the U-Plane, i.e. user traffic, and an MME supports the most relevant mobility management functions in the C-Plane of connected UEs. As a corollary of centralized control, it suffers from concentration of not only used data but also control traffic to MMEs. More specifically, an MME becomes easily overloaded by providing each of considerable number of M2M devices with full mobility management functions similar to regular mobile phones despite completely different traffic and mobility characteristics. In addition, physical distances between those management nodes in an

EPC network and UEs results in excessive bandwidth consumption and introduces large response delay in both of U-plane and C-plane. To tackle the problem, a direction toward distributed network architecture has drawn a lot of attentions in industry, academia, and government, being led by for example METIS (Mobile and wireless communications Enablers for Twenty-twenty Information Society) of Europe [5] [6], such as DMM [7] [8]. However they mainly consider the U-Plane.

In this chapter, we first propose novel architecture of autonomous and adaptive distribution of mobility management tasks in C-plane among nodes, i.e. a PGW (Packet data network GateWay), SGWs, and eNBs (evolved NodeB). That is, our proposed architecture is compromised with the current centralized architecture of the 3.9G LTE/EPS (Evolved Packet System) network. For flexible and adaptive distribution of mobility management tasks, we introduce ADMME, a virtual node or a virtual machine which has the same functionality of the LTE/EPC MME. In our proposed architecture, ADMMEs can be deployed at any node in a mobile core network with a help of virtualization technology. They communicate with each other by using for example the S10 and X2 interface to exchange control messages and deliver context information required for UE mobility management.

By introducing virtualized management entities called ADMMEs, dynamic and flexible assignment of mobility management tasks to nodes becomes possible. However, we need an algorithm and a mechanism to select and appoint an appropriate ADMME for each UE which dynamically changes location and mobility. Therefore, we propose a scheme to dynamically and adaptively select an ADMME appropriate for a UE based on its mobility characteristics and a management policy, i.e. delay mitigation or load balancing, of a mobile core network. In our proposal, when an ADMME receives a C-plane request from a UE in an Attach, TAU, or Handover procedure, it determines whether to delegate mobility management of the UE to other ADMME by using information about delay, load status of nodes, and C-plane overhead of ADMME switching.

Delay is measured by embedding timestamps to the request, load status is collected from nodes, and C-plane overhead is estimated at the ADMME. As delay between a UE and a node reflects their distance, the response time and the bandwidth consumption in the C-Plane can be reduced by appointing an ADMME closer to a UE as a serving ADMME. However, greedy delay minimization to select an eNB of a cell where a mobile UE resides causes considerable C-plane overhead due to frequent ADMME switching to handle UE mobility. For adaptive selection of an ADMME under dynamically changing constraints, we adopt a biologically-inspired heuristics, called the attractor selection model [12]. It is a mathematical model of behaviour of biological systems that can adapt themselves to dynamically changing environment without well-designed adaptation rules. Through simulation experiments using three mobility scenarios, we show the superiority of our proposal to five other methods.

In the rest of this chapter, first, in Section 2.2 we state current research and development activities on the 5G mobile network and distributed management. Next we propose our architecture of distributed mobility management in Section 2.3 and an ADMME selection scheme in Section 2.4. Then we evaluate our proposal through comparison with other methods in Section 2.5. Finally, we summarize this chapter and describe future work in Section 2.6.

2.2 Related Work

As mentioned in Section 1, problems caused by the centralized architecture in the current 3.9G LTE/EPC network are considered challenging by various organizations in many regions [5] [6]. To tackle the problems, DMM approaches have been recently investigated by IETF (Internet Engineering Task Force) and so on [7] [15]. IETF has established a working group for DMM and published drafts and RFCs (Request for Comments). We analyze activities on mobility management and then classify their approaches in Table 2.1 [7] [15] [16].

Table 2.1: Classification of mobility management approaches. "Partially" means distribution of only U-Plane and "Fully" distributes both of C-Plane and U-Plane.

	Centralized	Distributed	
		Partially (U-Plane Only)	Fully (C/U-Plane)
Host-based	MIPv6, NEMO, HMIPv6, MOBIKE		Dynamic Mobile IP Anchoring, DMME
Network-based	PMIPv6	Distributed Mobility Anchoring, Dynamic Tunneling for DMM, etc.	

In the centralized approaches, both of U-Plane and C-Plane managements are done by a single entity per UE, e.g. an HA (Home Agent) in MIPv6 (Mobile IPv6) and an LMA (Local Mobility Anchor) in PMIPv6 (Proxy Mobile IPv6). In contrast, there are multiple management entities in distributed approaches [16]. "Partially" means distribution of only U-Plane and "Fully" distributes both of C-Plane and U-Plane. Mobility management approaches can also be classified to two types in accordance with whether mobility management is based on client mobile IP or proxy mobile IP, i.e. Host-based and Network-based respectively as in the vertical axis in Table 2.1. Compared with host-based approaches, network-based approaches have three advantages of more flexible network management by having mobility functions in network-side apparatus only [17]. First, additional functions do not affect existing behaviour of UEs. Second, network-based approaches can reduce Handover-related latency since they avoid overhead over the air. Finally, network-based approaches can keep location of UEs in a confidential manner [17]. Considering these advantages of network-based approaches, PMIPv6 is adopted as one of mobility management methods in the current 3.9G LTE/EPC network [18].

Recently, active researches on partial DMM have been carried out [8] [19] [20] [21]. Examples are Distributed Mobility Anchoring (DMA) as an extension of PMIPv6 [22] and Dynamic Tunneling for DMM as a network-based DMM [23]. Regarding full DMM, IETF DMM Working group shows a blueprint for extension of partial DMM to full DMM [16] and summarizes its

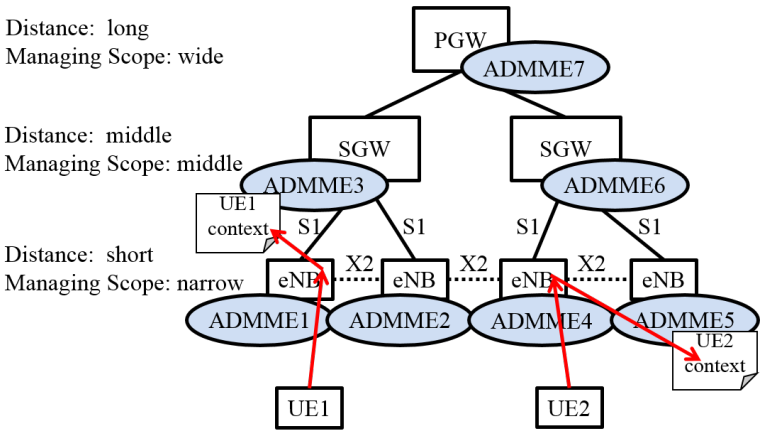

Figure 2.1: Conceptual architecture of distributed mobility management. ADMME has the same functionality of the LTE/EPC MME. Each node, such as PGW, SGW, and eNB, has one ADMME, but there could be nodes with multiple ADMMEs.

requirements [15]. Dynamic Mobile IP Anchoring (DMIPA) [24] is an example of host-based full DMM which considers distribution of mobility management functionalities among UEs and access routers for flat network architecture. Another example is DMME [9], which proposes distribution of MME. However, it cannot handle migration of UEs in ECM_CONNECTED state.

Considering the above, we take a network-based approach. Specifically, this chapter focuses on distribution of C-plane, which has not been well investigated yet in other literatures. We are going to incorporate U-plane distribution in the next step of our research.

2.3 Distributed Mobility Management Architecture

In this section, we introduce our distributed mobility management architecture, which can handle exploding mobile data traffic, considerable number of devices, and heterogeneous applications by using ADMME.

2.3.1 Conceptual Architecture

Figure 2.1 illustrates our conceptual architecture of distributed mobility management. In the figure,

each node, such as PGW, SGW, and eNB, has one ADMME, but there could be nodes with multiple ADMMEs. It also is possible that some nodes are unable to serve ADMMEs due to resource limitation. Each ADMME is responsible for mobility management of connected UEs under a certain management policy. For example, ADMME3 on an SGW manages UE1 and maintains its context information. The maximum number of UEs per ADMME depends on the capacity of a host node and available bandwidth.

2.3.2 New Functions Required in Mobile Networks

The following three functions are newly required to realize our proposed architecture and enable dynamic and adaptive distribution of ADMMEs while minimizing influence on the current 3GPP specifications.

- ADMME distributed deployment function
  - Each UE can communicate with any ADMME distributed in mobile networks.
- ADMME switching function
  - Each ADMME can autonomously decide a UE to manage and delegate management of the UE to other ADMME if needed. We call the delegation of management task ADMME switching.
- Information collection function
  - Each ADMME can collect information required to decide ADMME switching.

Details of the three functions are explained below. First, any ADMME must be able to communicate with any UE and delegate mobility management tasks to other ADMME at any time to have the same functionality as standard MME. For this purpose, it is necessary (1) to ensure IP

reachability between a UE and eNB or ADMME, and (2) to dynamically update mapping between a GUMMEI (Globally Unique MME Identity) and an IP address in an eNB as described below.

In 3GPP specifications [25], an eNB translates a GUMMEI specified by a requesting UE to an IP address of the corresponding MME and communicates with the MME using S1AP (S1 Application Protocol) messages [26] [27]. Among interconnected eNBs an X2 logical interface is also available and messages are exchanged by using X2AP (X2 Application Protocol). Since an ADMME can be deployed at any node at any time, an eNB needs to update the latest mapping between GUMMEIs and IP addresses. Possible solutions are (1) to acquire a mapping table from an HSS (Home Subscriber Server) or an operation server periodically or (2) to acquire a mapping table from a DNS (Domain Name System) server on receiving a request from a UE, for example. Regarding assignment of an IP address, an ADMME can use either of (1) the same IP address as a physical node which the ADMME is deployed on or (2) a different IP address. In the former case, a TCP or UDP port can be used to distinguish between an ADMME and a physical node.

Second, in 3GPP specifications [25], there are opportunities of MME switching in attach procedure for UE registration, TAU (Tracking Area Update) procedure for location update, and X2/S1 handover procedure. Similarly in our proposal, on receiving a request from a UE during either of these procedures, an ADMME decides whether to perform ADMME switching based on the ADMME switching policy and information about candidate ADMMEs, called possible ADMMEs. Possible ADMMEs are those received or forwarded a request message of the UE in the current and past procedures. Since the TAU procedure is initiated not only when a UE moves from one TA to another but also when a TAU timer expires, ADMME switching would be carried out even for a non-moving UE. Details will be given in Section 2.4.1.

Table 2.2 shows parameters used as an ADMME switching policy. Parameters listed in "Selection Priority" are taken into account to determine an ADMME appropriate for mobility

Table 2.2: ADMME switching policy. An ADMME switching policy would additionally include for example an interval of information collection and a period that collected information is valid.

Category	Information
Selection Priority	delay, load, switching frequency, signalling overhead, communication and mobility characteristics of UE
Algorithm	attractor selection, random selection, linear programming, etc.
Others	information collection interval, information validity period

management of a requesting UE. The appropriateness of selection depends on the management policy of mobile core networks, e.g. delay minimization, load balancing, and minimization of C-plane overhead. A variety of algorithms can be used for adaptive ADMME switching as far as it provides satisfactory solutions in regard to the management policy. In this chapter, we propose an effective algorithm based on the attractor selection model and verify its superiority by comparing to other algorithms. Finally, an ADMME switching policy would additionally include for example an interval of information collection and a period that collected information is valid. An ADMME switching policy is delivered to ADMMEs in advance or as needed.

When an ADMME decides to delegate management tasks of a requesting UE to another ADMME, an ADMME switching procedure is performed. For example in Fig. 2.2, as indicated as procedure (1), UE1 sends a request of Attach, TAU, or Handover to ADMME2, which is responsible for mobility management of UE1 at this time. We call an ADMME which is responsible for mobility management of a UE and a request from the UE is destined to a current ADMME. On receiving the request, ADMME2 evaluates the necessity of switching based on an ADMME switching policy and decides to delegate the role to ADMME3 in this case. It sends an ADMME switching request with context information of UE1 to ADMME3 in procedure (2). Then in procedure (3) ADMME3 performs a switching process such as U-plane routing. Finally in procedure (4) ADMME2 sends a

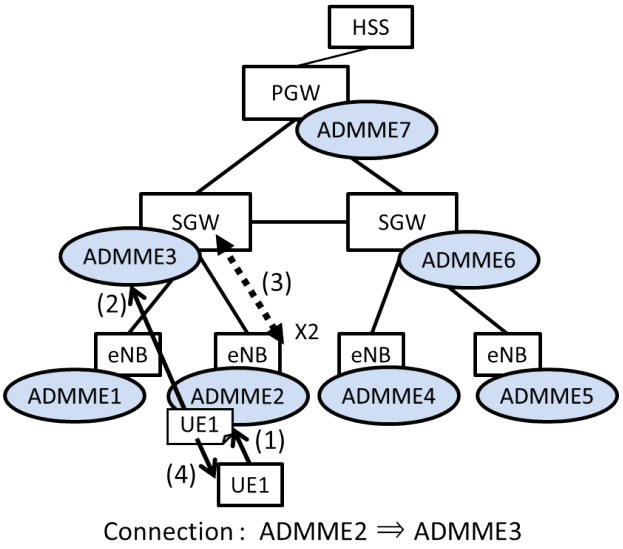

Figure 2.2: ADMME switching procedure. (1) UE 1 sends a request to ADMME 2, namely current ADMME. ADMME2 evaluates and decides to delegate the role to ADMME3. (2) It sends an switching request to ADMME 3 with context information of UE 1. (3) ADMME3 performs a switching process. (4) ADMME2 sends a response to UE1.

response to UE1 to notify ADMME switching.

Details of signalling in a case that TAU triggers ADMME switching are illustrated in Fig. 2.3. After making a decision to perform ADMME switching, a current ADMME (ADMME2) sends a Switching Request message together with UE context information to a new ADMME (ADMME3). The new ADMME sends a Modify Bearer Request message to a designated SGW to switch a connection between the SGW and the current SGW to the new SGW. The request message contains an IP address of the new ADMME and a TEID (Tunnel Endpoint ID) to manage EPS bearer. Then the SGW returns a Modify Bearer Response message. A reason why those messages are indicated by dashed lines is that changing a bearer connection is not required in this example. On receiving the response, the new ADMME sends a Switching Response message to the current ADMME. At the same time, it sends a Update Location Request message to an HSS to notify the ADMME switching. The HSS tells the current ADMME to discard the UE context information by sending

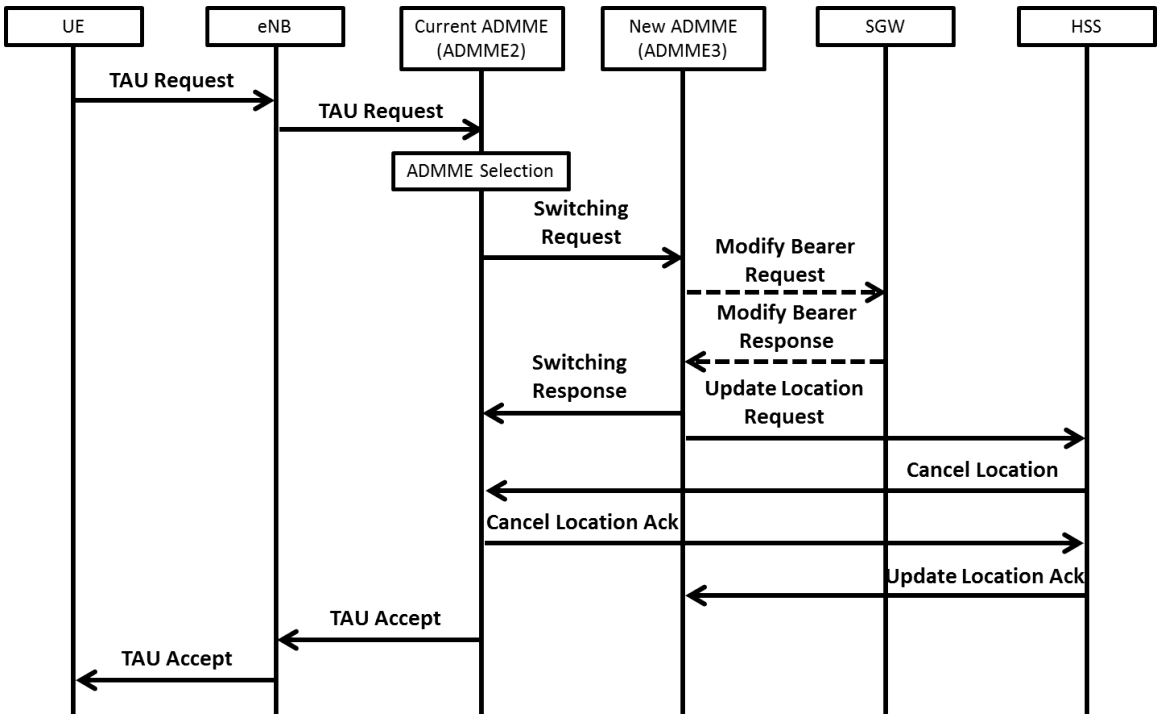

Figure 2.3: Tracking Area Update. Details of signalling in a case that TAU triggers ADMME switching.

a Cancel Location message. After receiving a Cancel Location Ack message from the current ADMME, the HSS sends a Update Location Ack to the new ADMME to notify completion of updating location information. Finally, the current ADMME sends a TAU Accept message to the UE via the eNB. In the case that ADMME switching does not occur, messages are exchanged as shown in Fig. 2.4, which follows the same steps as the 3GPP specifications.

Finally, the information collection function collects information required for decision of ADMME switching, including ADMME load, inter-node delay, and UE characteristics information. ADMME load information corresponds to load status of an ADMME in C-plane. Examples of ADMME load information are summarized in Table 2.3. ADMME load information can be collected periodically or as needed by exchanging load information request and response messages

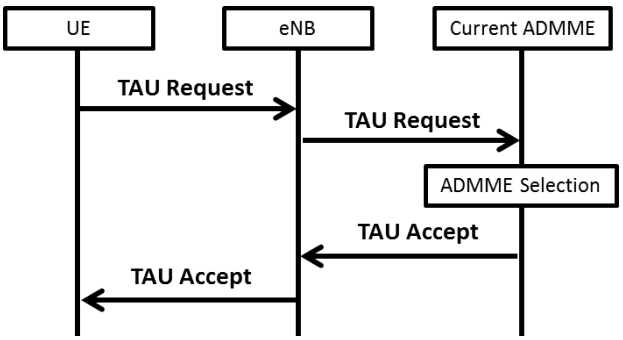

Figure 2.4: Tracking Area Update. Details of signalling during TAU procedure in a case that ADMME switching does not occur.

Table 2.3: ADMME load information. ADMME load information corresponds to load status of an ADMME in C-plane; periodically collected.

Category	Information
Number of UE connections	Maximum and current number of UE connections
Amount of signaling	Amount of signaling messages exchanged with each UE and other nodes

between ADMMEs.

Inter-node delay information means delay between physical nodes which ADMMEs are deployed on. From a viewpoint of delay minimization, it would be better to select an ADMME with the smallest delay or the shortest distance to a UE as a current ADMME. Table 2.4 summarizes examples of inter-node delay information. There are two types of information, i.e. delay-based and topology-based ones. Inter-node delay information is obtained by (1) exchanging measurement messages between ADMMEs or (2) using regular request and response messages exchanged between an ADMME and a UE. In the latter case, each message has to carry pairs of an IP address and a timestamp of all nodes which it traverses.

As for UE characteristics information, see Table 2.5 for examples. They are helpful in finding the most appropriate ADMME in regard to the management policy. For example from a viewpoint

Table 2.4: Inter-node delay information. There are two types of information, i.e. delay-based and topology-based ones.

Category	Information
Measurement Value	Delay between nodes
Route Information	Hop count between nodes, physical location of node

Table 2.5: UE characteristics information. They are helpful in finding the most appropriate ADMME in regard to the management policy.

Category	Information
Mobility Characteristics	Handover interval for each UE
Communication Characteristics	Communication interval, state transition interval for each UE

of delay minimization, an eNB managing a cell where a UE resides is the best. However, such selection causes frequent ADMME switching and incurs considerable C-plane overhead for a highly mobile UE. In this case, a node at a higher level, e.g. an SGW and a PGW, is considered more appropriate. Since those UE-related information is not always available, our ADMME selection scheme proposed in Section 2.4 does not rely on UE characteristics information.

We should note here that these three functions can be easily realized without significant change in 3GPP specifications or cost such as complexity and performance. For example, a function to manage and select an MME is specified in "4.3.8.3 MME selection function" of 3GPP TS 23.401 V13.3.0 (2015-06) [28]. Addition of information about possible ADMMEs to MME management information accomplishes the ADMME distributed deployment function. Regarding the ADMME switching function, existing MMEs have a similar mechanism to switch MMEs on receiving a TAU from a UE as specified "5.3.3 Tracking Area Update procedures" [28] and there is freedom of design of a selection algorithm. Exchange of information among MMEs and collection of information about UEs are specified in "5.3.3 Tracking Area Update procedures" and "5.7.2 MME" of 3GPP

TS 23.401 V13.3.0 (2015-06), respectively. Therefore, the information collection function can be easily implemented by adding optional information to the standard procedure without significant change in 3GPP specifications or performance degradation.

2.4 ADMME Selection Scheme

In this section, we propose an autonomous ADMME selection scheme which aims at delay minimization, load balancing, and C-plane overhead mitigation taking into account of mobility characteristics of UEs.

2.4.1 Outline of ADMME Selection Scheme

As explained, on receiving a request of Attach, TAU or Handover, a current ADMME selects an appropriate ADMME from a set of ADMMEs called possible ADMMEs by using an algorithm explained in the next section. Therefore the frequency of decision making of ADMME switching is the same as the frequency of requests, which is the same as in the current mobile core network. The algorithm uses three information, i.e. the history of ERDs (Estimated Response Delays), the load status of ADMMEs, and the C-plane overhead related to ADMME switching.

An ERD of an ADMME is an estimation of the sum of the duration from emission of a request message to its reception at the current ADMME and the delay from the current ADMME to the ADMME. Therefore, ERDs are measured every time the current ADMME receives an Attach, TAU, or Handover request from a UE. The history of ERDs is a list of tuples of an ADMME identifier and the corresponding ERD in the past W procedures. Possible ADMMEs include those located at nodes on a path from the requesting UE to the current ADMME, those recorded in the ERD history, and those on the nearest SGW and PGW.

See Fig. 2.5 and Table 2.6 as an example. In Fig. 2.5, we denote delays between a UE and the nearest eNB, between an eNB and a connected SGW, between an SGW and a PGW as Δ_0,

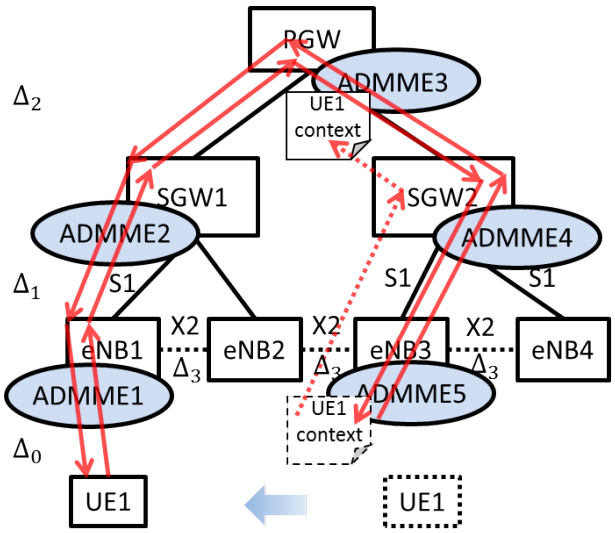

Figure 2.5: Example of UE movement and ADMME switching. Delays between a UE and the nearest eNB=Δ_0, between an eNB and a connected SGW=Δ_1, between an SGW and a PGW=Δ_2, between a pair of connected eNBs=Δ_3. ADMME5 is current ADMME; At time t_0 UE1 moves from the cell of eNB3 to eNB1. UE1 sends a request to the current ADMME, ADMME5 following the tree structure as indicated by red arrows.

Δ_1, and Δ_2, respectively. Furthermore, the delay between a pair of connected eNBs is denoted as Δ_3. Assume that UE1 first stays in a cell of eNB3. Several procedures are triggered in this cell and ADMME5 located at eNB3 is selected to be the current ADMME for UE1. At time t_0 UE1 moves to the eNB1's cell and its current ADMME is still at eNB3, i.e. ADMME5. Then UE1 sends the h-th request to ADMME5. Since there is no X2 interface between eNB1 and eNB3, a request follows the tree structure through eNB1, SGW1, PGW, and SGW2 as indicated by red arrows.

A request remembers a timestamp when it leaves a node. Now, timestamps that a request remembers on arriving at ADMME5 are; ADMME1=$t_0 + \Delta_0$, ADMME2=$t_0 + \Delta_0 + \Delta_1$, ADMME3=$t_0 + \Delta_0 + \Delta_1 + \Delta_2$, ADMME4=$t_0 + \Delta_0 + \Delta_1 + 2\Delta_2$, and ADMME5=$t_0 + \Delta_0 + 2\Delta_1 + 2\Delta_2$. By using those timestamps, ERDs (delay from UE to the current ADMME + delay from the current ADMME to the ADMME) are derived as: ADMME5=$\Delta_0 + 2\Delta_1 + 2\Delta_2$, ADMME4=$\Delta_0 + 3\Delta_1 + 2\Delta_2$, ADMME3=$\Delta_0 + 3\Delta_1 + 3\Delta_2$, ADMME2=$\Delta_0 + 3\Delta_1 + 4\Delta_2$, and

Table 2.6: An example of the history of ERDs. A request remembers a timestamp when it leaves a node. By using those timestamps, ERDs (delay from UE to the current ADMME + delay from the current ADMME to the ADMME) are derived.

Time	ERDs of possible ADMMEs				
	i=1	2	3	4	5
h	ADMME1	ADMME2	ADMME3	ADMME4	ADMME5
	$\Delta_0 + 4\Delta_1$ $+4\Delta_2$	$\Delta_0 + 3\Delta_1$ $+4\Delta_2$	$\Delta_0 + 3\Delta_1$ $+3\Delta_2$	$\Delta_0 + 3\Delta_1$ $+2\Delta_2$	$\Delta_0 + 2\Delta_1$ $+2\Delta_2$
$h-1$	ADMME5	ADMME4	ADMME3		
	Δ_0	0	0		
$h-2$	ADMME5	ADMME4	ADMME3		
	Δ_0	0	0		
$\cdots$	$\cdots$	$\cdots$	$\cdots$		
$h-$	ADMME5	ADMME4	ADMME3		
$W+1$	Δ_0	0	0		

ADMME1=$\Delta_0 + 4\Delta_1 + 4\Delta_2$. Those values are registered as an entry for the h-th procedure in the history of ERDs as shown in Fig. 2.6. Here, we name ERDs of ADMME1 through ADMME5 as $d_1(h)$ through $d_5(h)$ respectively. As will be explained in Section 2.4.2, we consider the sum of ERDs of the h-th procedure, items belonging to the same procedure can have arbitrary identifiers. In our definition, an ADMME closer to a UE than the current ADMME has a larger ERD. Since UE1 stays in the eNB3's cell from the $(h-W+1)$-th to the $(h-1)$-th procedures, the table contains ERDs of ADMME3 at the PGW, ADMME4 at SGW2, and ADMME5 at eNB3. To evaluate ERDs of ADMMEs on the PGW and SGW2, which are located behind the current ADMME, i.e. ADMME5, their ERDs are recorded as 0 in Table 2.6. If in the past procedure the current ADMME is ADMME4, the ERD of PGW which is behind the current ADMME is 0 in the table. Note that all the nodes in the Table 2.6 are possible ADMMEs.

The second information is the load status of ADMMEs. Load balancing is one of crucial issues to mitigate influence of a node failure and avoid excessive expenditure of node and network resources especially when we consider overhead in managing a large number of M2M devices. For ADMME selection aiming at load balancing, each ADMME collects load status information from

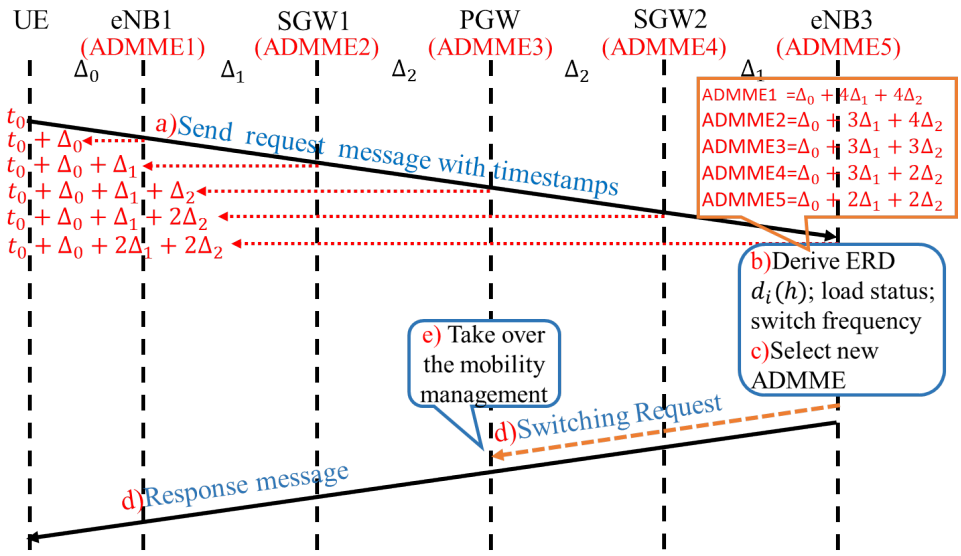

Figure 2.6: The h-th TAU/Handover procedure. Those ERD values are registered as an entry for the h-th procedure in the history of ERDs.

possible ADMMEs by using the Information Collection Function explained in Section 2.3.2. In the evaluation section, we use the ratio of the number of UEs to the capacity of an ADMME as the load status information.

The last information that an ADMME uses to select an appropriate ADMME is the overhead to move UE contexts from the current ADMME to a new ADMME in ADMME switching. When the current ADMME considers that another ADMME is more suitable to perform mobility management of a requesting UE, it delegates the mobility management task of the requesting UE by sending a Switching Request message carrying UE context information required for mobility management as shown in Fig. 2.3. Since such switching takes time and consumes bandwidth, frequent switching should be avoided from a viewpoint of C-plane overhead.

After completion of ADMME switching, a current ADMME sends a response to a requesting UE. In Fig. 2.5, ADMME5 chooses ADMME3 on the PGW as a new ADMME. In this case, the response carries the updated context information of the UE with an identifier of ADMME3. Then,

UE1 can recognize a new ADMME and sends succeeding requests directly to ADMME3.

2.4.2 Autonomous and Adaptive ADMME Selection Algorithm

An ADMME selection algorithm should adaptively find a solution of a multi-objective optimization problem in accordance with the dynamically changing location and mobility of UEs and load status of ADMMEs. By general distributed algorithms, the relationship between servers and tasks becomes more complicated, correspondingly computation becomes more complicated. For example, S. Kraus et al. [11] presented some algorithms of distributed task allocation for cooperative agents. Although the distributed algorithm can achieve high performance close to the optimal, the complexity of the algorithm depends on the number of the agents and the number of types of tasks. Accordingly, if we adopt it to ADMME allocation, complexity of the algorithm increases a lot which causes substantial overhead of computation and communication, especially for large-scale mobile networks accommodating numerous UEs. Additionally, a biologically-inspired algorithm can enable potential optimization of task allocation simply by utilizing fluctuations under dynamically changing constraints. Therefore, in this thesis, our selection algorithm adopts a nonlinear mathematical model called attractor selection model. It follows a heuristics inspired from biological systems, which can adapt themselves to dynamically changing and even unknown surroundings [12]. For its high adaptability, it is considered well-suited to the targeted problem than other metaheuristics and has been applied to a variety of network control [29].

In the general form, the attractor selection model is expressed as $d\vec{x}/dt = f(\vec{x}) \cdot \alpha + \vec{\eta}$. $\vec{x}$ corresponds to the state of a system, whose dynamics is governed by an energy function f. $f(\vec{x})$ defines attractors, i.e. a set of states where a system converges to. $\vec{\eta}$ corresponds to internal and / or external noise causing fluctuation. α $(0 \leq \alpha \leq 1)$ is a scalar value called activity. Activity α reflects the goodness of state $\vec{x}$ in regard to the current condition. When α is large, that is, the state is appropriate for the condition, temporal dynamics of the system state is governed by the

energy function f. As a result, the state converges to a nearby attractor and the system stably stays there. When the condition changes and the state becomes inappropriate, the activity decreases first. Then, the noise term dominates dynamics and the state randomly changes, i.e. random walk. Once the state approaches an attractor appropriate for the new condition, the activity gradually increases. Consequently the influence of f becomes larger and the state will be entrained to the new attractor. As a result of the increased activity, finally the state reaches the new attractor and the system becomes stable again.

In summary, the attractor selection model contains heuristics combining deterministic dynamics corresponding to reinforcement of a solution and random search with mediation of the activity as feedback. The attractor selection model enables a system to find a state appropriate for the dynamically changing surrounding condition in an adaptive manner.

In our proposal, an ADMME maintains activity α for each UE it manages. The activity is defined to express the goodness of the current ADMME as a mobility management entity for the UE. In addition, an ADMME also maintains a vector $\vec{m}$=(m_1, m_2, $\cdots$, m_M) called a state vector for each UE it manages. m_i is called a state value and indicates the goodness of a possible ADMME i as a serving ADMME. M is the number of possible ADMMEs. Based on the definition explained in the previous section, a set of possible ADMME differs among UEs. In general, there are more possible ADMMEs, i.e. a larger M, for a UE with higher mobility. The current ADMME chooses a possible ADMME with the largest state value as a new ADMME in an ADMME selection process.

When an ADMME receives the h-th request from a UE, it first evaluates ERDs, load status, and C-plane overhead, and next calculates the activity by using the following equation.

$$\alpha(h) = \rho \cdot \alpha_{delay}(h) + (1 - \rho) \cdot \alpha_{load}(h), \tag{2.1}$$

where ρ ($0 \leq \rho \leq 1$) is a weight parameter to take a balance of $\alpha_{delay}(h)$ and $\alpha_{load}(h)$ reflecting a

management policy of a mobile core network.

$\alpha_{delay}(h)$ is a delay-based activity which is derived as,

$$\alpha_{delay}(h) = \left(\frac{\sum_{k=0}^{W-1} \frac{d_{cm}(h-k)}{k+1}}{\max\limits_{1 \le i \le M} \sum_{k=0}^{W-1} \frac{d_i(h-k)}{k+1}} \right)^{\varepsilon}, \text{if } h > W. \tag{2.2}$$

or

$$\alpha_{delay}(h) = \left(\frac{\sum_{k=0}^{h-1} \frac{d_{cm}(h-k)}{k+1}}{\max\limits_{1 \le i \le M} \sum_{k=0}^{h-1} \frac{d_i(h-k)}{k+1}} \right)^{\varepsilon}, \text{if } h \le W. \tag{2.3}$$

$d_i(h)$ is the ERD of possible ADMME i measured by the h-th request message and cm means the current ADMME. For example, $\sum_{k=0}^{W-1} d_{cm}(h-k)/k+1$ is the weighted sum of ERDs of the current ADMME measured in the past W procedures. By dividing by $k+1$, we take account later ERDs more than older ones. When the current ADMME has the largest sum among possible ADMMEs, $\alpha_{delay}(h)$ becomes 1.

A load-based activity $\alpha_{load(h)}$ is derived from the load status of ADMMEs as,

$$\alpha_{load}(h) = \frac{\min\limits_{1 \le i \le M} l_i(h)}{l_{cm}(h)}. \tag{2.4}$$

Here, $l_i(h)$ is the load or utilization of ADMME i. Although there are a variety of definitions of $l_i(h)$, in this paper we use the ratio as $l_i(h) = n_i(h)/c_i(h)$. $n_i(h)$ is the number of UEs that ADMME i manages. $c_i(h)$ is the capacity which is determined based on computational capacity, memory, and bandwidth of ADMME i, for example. When the load of the current ADMME is higher than any of possible ADMMEs, $\alpha_{load}(h)$ becomes small.

Then, an ADMME updates state values by using the following equation.

$$\frac{dm_i}{dt} = \frac{s(\alpha(h))}{1 + m_{max}^2 - m_i^2} - d(\alpha(h)) \cdot m_i + \eta_i. \tag{2.5}$$

where $m_{max} = \max_{1 \le j \le M}(m_j)$, $s(\alpha(h)) = \alpha(h)[\beta \cdot \alpha(h)^{\gamma} + \varphi^*]$, $d(\alpha(h)) = \alpha(h)$, and $\varphi^* =$

$1/\sqrt{2}$. η_i is the white Gaussian noise with mean of 0 and variance of 1. If state value m_i is the maximum, i.e. $m_i = m_{max}$ and α is large, $s(\alpha(h)) > d(\alpha(h))$ and thus m_i increases. At the same time, other smaller state values decrease. Eventually, state values reach the equilibrium where one state value m_i is the largest and other state values are close to 0. Therefore, a system whose dynamics is defined by Eq. 2.5 has M attractors, i.e. stable conditions, where one state value is the largest and the others are small. On the contrary, when α is so small that the amount of change given by the first two terms is smaller than the noise, state values randomly change.

Now we explain how adaptive selection of ADMME is accomplished by our algorithm. If the current ADMME is appropriate for a UE, α becomes high and dynamics of $\vec{m}$ is governed by the first two terms of the right-hand side of Eq. 2.5. As explained above, it pushes the largest state value, corresponding to the current ADMME, to increase while making the other state values decrease. As a consequence of reinforcement, the system will reach a stable state where one state value out of M is the largest and the others are small. State values are stably kept with small perturbation of a noise term.

On the contrary, if the current ADMME has a large delay or unfair load status, α becomes small and the dynamic system is not stable any more. By being driven by the noise term, state values randomly change and the role of mobility management of the requesting UE would be delegated to another ADMME. If an ADMME leading to smaller delay or fairer load status is selected, the activity eventually increases and selection becomes stable.

By using the activity defined by Eq. 2.1, an appropriate ADMME can be determined for a UE taking balance between delay and load and considering mobility characteristics of the UE. However, as we pointed in Section 2.1, one of the crucial issues of a mobile core network in the forthcoming era of IoT and M2M is explosive increase in C-plane overhead. Therefore, we further incorporate an algorithm to reduce C-plane overhead especially caused by ADMME switching.

One simple way is to limit the number of ADMME switching less than that of the current standard. In the current architecture, MMEs are located at SGWs independently of mobility of UEs. When a UE moves from one TA to another, management of the UE is switched from an MME of the previous TA to that of the new TA by moving the corresponding context information among them. Therefore, by counting the number of TAU triggered by movement between TAs, the number of UE context migrations in the current standard can be estimated. We refer the number of UE context migrations estimated at the h-th procedure for the past W procedures for a UE as $N_{SGW}(h)$. When the number of ADMME switching in the proposal expressed as $N(h)$ is more than $N_{SGW}(h)$, the current ADMME does not perform ADMME switching independently of a result of evaluation. We call this method Deterministic. On the contrary, the original method which does not use $N(h)$ is called Simple.

Deterministic can strictly limit the C-plane overhead of ADMME switching. However, it spoils adaptability and optimality of our proposal in distributed mobility management. To tackle the problem and allow the certain degree of freedom of switching even in the case that the switching frequency is larger than the existing standard method, i.e. $N(h) \geq N_{SGW}(h)$, we introduce another parameter $\delta(h)$. It is derived as,

$$\delta(h) = \frac{\max(N(h), N_{SGW}(h)) + 1}{N_{SGW}(h) + 1}. \tag{2.6}$$

Then $\delta(h)$ is combined with $\alpha(h)$ as,

$$\alpha(h) \leftarrow \min\{1, \alpha(h) \cdot \delta(h)\} \tag{2.7}$$

Figure 2.7 shows the relationship between $\delta(h)$ and $N(h)$ when $N_{SGW}(h) = 10$. As shown in the figure, when $N(h)$ is less than $N_{SGW}(h)$, $\delta(h)$ remains 1, which means there is no influence to activity α. When ADMME switching frequently occurs, $N(h)$ becomes larger than $N_{SGW}(h)$ and $\delta(h)$ linearly increases. As a result, ADMME switching is avoided even when $\alpha(h)$ derived

Table 2.7: Proposed algorithms for ADMME selection. Summarization of alternatives of our ADMME selection algorithms.

Name	Activity	Deterministic control
Simple	$\alpha = \rho \cdot \alpha_{delay} + (1 - \rho) \cdot \alpha_{load}$	None
Deterministic	$\alpha = \rho \cdot \alpha_{delay} + (1 - \rho) \cdot \alpha_{load}$	ADMME selection stops if $N \geq N_{SGW}$
Delta	$\alpha \leftarrow \min\{1, \alpha \cdot \delta\}$	None

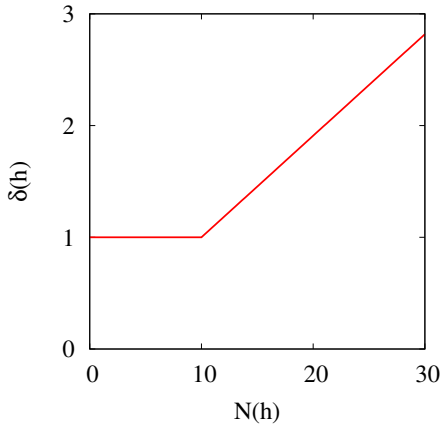

Figure 2.7: Relation between $\delta(h)$ and $N(h)$. $N_{SGW}(h) = 10$. When $N(h)$ is less than $N_{SGW}(h)$, $\delta(h)$ remains 1, which means there is no influence to activity α. When ADMME switching frequently occurs, $N(h)$ becomes larger than $N_{SGW}(h)$ and $\delta(h)$ linearly increases.

by Eq. 2.1 is very small. We call this method Delta. We summarize alternatives of our ADMME selection algorithms in Table 2.7.

2.5 Evaluation and Discussion

In this section, we evaluate our proposals from viewpoints of delay, load, and C-plane overhead using three mobility scenarios. We used our simulation program written in C++. In order to evaluate our selection mechanism, procedures illustrated in Figs. 2.3 and 2.4 are simulated, but other detailed mechanisms such as bearer setup between eNB and SGW are omitted.

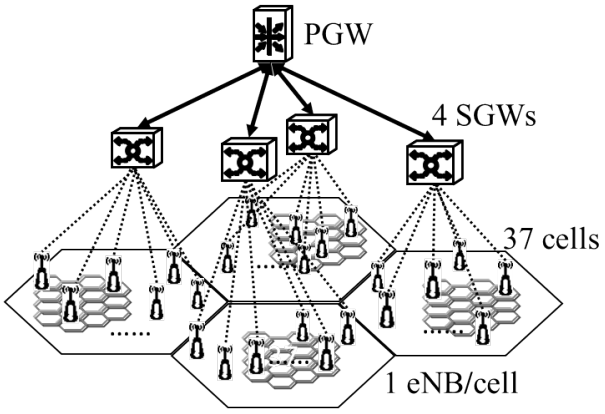

Figure 2.8: Simulation topology. Torus TA topology. 1 PGW, 4 SGWs/PGW, 1 SGW/TA, 37 cells/TA, 1 eNB/cell, 100 UEs/cell, Total: 153 ADMMEs, 14800 UEs.

2.5.1 Simulation Setting

A mobile core network used for simulation experiments has one PGW, four SGWs, and 37 eNBs per SGW as illustrated in Fig. 2.8. As in many other papers, we assume each eNB covers a hexagonal cell of diameter Φ and cells of 37 eNBs connected with an SGW organize a large hexagonal TA [30] [31]. We consider torus topology and each TA shares borders with all the other TAs. In each cell 100 UEs are located at the beginning of a simulation run. Therefore there are 14800 UEs. As for delays we set $\Delta_0 = 2$ ms, $\Delta_1 = 20$ ms, $\Delta_2 = 3$ ms, and $\Delta_3 = 4$ ms taking into account their average physical distances. To demonstrate load balancing performance, we empirically set the capacity of ADMMEs deployed at a PGW, an SGW, and an eNB as 8000, 4000, and 200 UEs per ADMME, respectively. Initially an ADMME located at an eNB of a cell where a UE resides is selected as the current ADMME in our proposals. However, it can be any other place in reality since an appropriate ADMME is adaptively and dynamically selected as time passes. Parameters used are $\beta = 10$, $\gamma = 10$, and $W = 5$ which are determined based on preliminary experiments. We change ρ as 0, 0.5, and 1 to investigate the influence of a weight parameter.

All UEs are attached throughout a simulation run, but only 30% out of them randomly selected

at time step are connected and the remaining 70% are in the idle state. In the evaluation, a time step corresponds to 10 minutes. They move from one cell to another based on a stay timer which is set at random following the Gaussian distribution with average of T_s and variance of 1 at the beginning of a simulation run. The initial value of stay timer of a UE is set at random from 0 to T_s. When a stay timer expires a UE moves to randomly selected one of neighbor cells. A TAU timer for periodic tracking area update is identically 30 min, but an initial value is set at random for each UE to avoid synchronized TAU.

We consider three mobility scenarios with different T_s setting. In Scenario 1, in order to observe the behaviors of ADMME selection corresponding to different mobility level of UEs to verify that our proposal can allocate appropriate ADMMEs according to UE mobility characteristic, we consider a homogeneous and constant mobility for all UEs with T_s=0.5, 1, 2, or 10 hrs. We also consider immobile UEs, i.e. $T_s = \infty$. Scenario 2 considers heterogeneous but constant UE mobility to investigate the influence of different UE mobility and confirm that our scheme can accomplish adaptive allocation of appropriate ADMMEs, in which 50% of UEs use $T_s = 0.5$ hrs and the other half use $T_s = 10$ hrs during the whole simulation run. In Scenario 3, UEs have heterogeneous and dynamic mobility, which is close to reality to show that our proposal can be applied in practice. A UE with dynamic mobility changes its T_s to randomly selected one from ∞ (immobile), 0.5, 1, 2, and 10 hrs every 8 hrs. For Scenario 3, we consider five different cases by changing the ratio of UEs with dynamic mobility as 100%, 90%, 80%, 70%, and 60%. In the latter four cases, remaining UEs are all immobile assuming sensor nodes for example. Therefore, by comparing the results of UEs of the same mobility with Scenario 1, we can confirm that our proposal can accomplish adaptive and dynamic allocation of appropriate ADMMEs and be applied in practice.

We compare six methods, i.e. Delta, Simple, Deterministic, PPGW, PSGW, and PeNB. PPGW represents a case without ADMME switching, in which ADMMEs for all the UEs are persistently

located at the PGW. PSGW corresponds to the current 3GPP/LTE standard, where an ADMME is always located at an SGW nearest to a UE. PeNB shows an extreme and greedy case where an ADMME is located at an eNB nearest to a UE, which should lead to delay minimization.

For comparison, we consider four measures. In the next section we show averages over 100 simulation runs. The first is the average response delay, which is the average duration from emission of a request from a UE to reception of a response. The second is the fairness of load. We use the Jain's fairness index [32] as,

$$\frac{(\sum_{i\in\{PGW,SGW,eNB\}} \bar{l}_i)^2}{3 \cdot \sum_{i\in\{PGW,SGW,eNB\}} \bar{l}_i^2}, \tag{2.8}$$

where $\bar{l}_i$ is the average load of nodes of type $i \in \{PGW, SGW, eNB\}$ at the end of a simulation run. The third is the number of ADMME switching per time step per UE. Here, time step is 10 minutes in our simulations. For Scenario 1, we also evaluate the average C-plane signalling overhead, which is the average number of messages multiplied by the average number of hops which a message takes per UE per hour for C-plane mobility management. We consider messages both for regular C-plane mobility management exchanged between a UE and a current ADMME and for ADMME switching. Therefore, if a current ADMME is distant from a UE or the location of a current ADMME frequently changes, the C-plane signalling overhead becomes large.

In order to evaluate relative performance, we consider normalization of delay, fairness, and C-plane overhead by the following equation in Scenario 3,

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{2.9}$$

X corresponds to either of average response delay and the average number of ADMME switching of a selected scheme. X_{min} and X_{max} are the minimum and maximum among schemes respectively. Therefore, the normalized value X' ranges from 0 to 1, where 0 means X is the minimum and 1 means the maximum respectively. That is, a smaller X' is better. In the case of fairness, because a

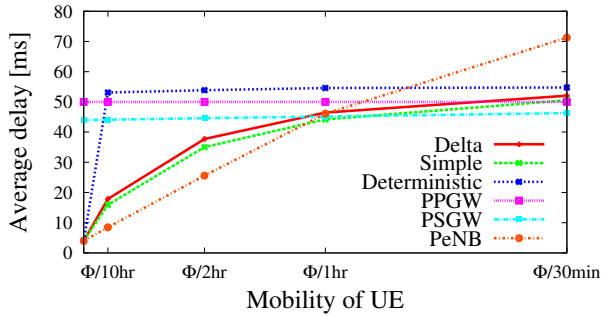

Figure 2.9: Average delay in Scenario 1. $\rho = 1$ means our methods only consider delay. The average delay of PeNB is the smallest for UEs with low mobility, but it increases in proportion to the mobility. The average delay is kept almost constant for PSGW. For PPGW the average delay is the same among UEs with different mobility.

larger value means better performance, we use,

$$X' = 1 - \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{2.10}$$

2.5.2 Results and Discussion

First Figs. 2.9, 2.10, and 2.11 show results of Scenario 1 with $\rho = 1$. Therefore, our methods only consider delay in Eq. 2.1. The x-axis shows the mobility of UEs expressed as a ratio of the cell diameter Φ to the average stay time T_s. 0 means immobile. Obviously, the average delay of PeNB is the smallest for UEs with low mobility, but it increases in proportion to the mobility. That is because, the location of a current ADMME frequently changes to keep up with a fast-moving UE as shown in Fig. 2.11 and ADMME switching causes additional response delay. Similarly, as shown in Fig. 2.11, the number of ADMME switching increases proportionally to the UE mobility in the case of PSGW. However, the average delay is kept almost constant. It is because in PSGW the delay for ADMME switching between two SGWs is much smaller than in PeNB due to the tree topology. Regarding PPGW, since there is no ADMME switching, the average delay is the same among UEs with different mobility as shown in Fig. 2.9. Figure 2.10 shows that PPGW, PSGW, and PeNB have

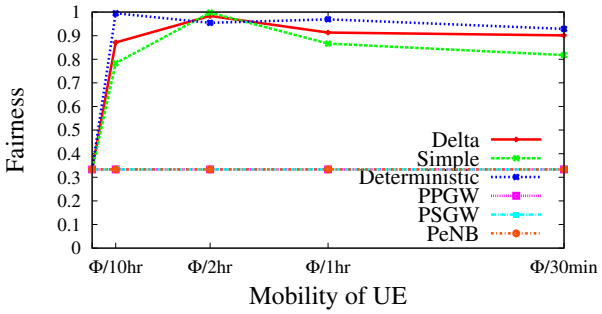

Figure 2.10: Fairness in Scenario 1. $\rho = 1$ means our methods only consider delay. PPGW, PSGW, and PeNB have identically low fairness as a result of concentration of ADMMEs to nodes of a specific type.

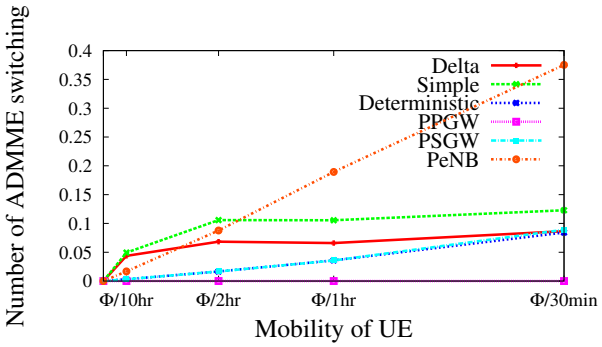

Figure 2.11: Number of ADMME switching in Scenario 1. $\rho = 1$ means our methods only consider delay. The number of ADMME switching increases proportionally to the UE mobility in the case of PSGW. The location of a current ADMME frequently changes to keep up with a fast-moving UE.

identically low fairness as a result of concentration of ADMMEs to nodes of a specific type.

Among our proposals, Delta and Simple achieve delay smaller than PSGW for UEs with lower mobility. If a UE is immobile, an ADMME at the nearest eNB is selected and delay is minimized. For a UE with moderate mobility, an ADMME at an SGW serving a TA where a UE resides is likely to be selected. When a UE frequently moves from one TA to another, an ADMME at the PGW would be selected to avoid frequent and costly ADMME switching. That is, our proposed algorithm is not greedy in delay minimization. It does not always prefer the nearest eNB as a location of an ADMME even with $\rho = 1$. Assume that an ADMME at the nearest eNB is once selected. Because

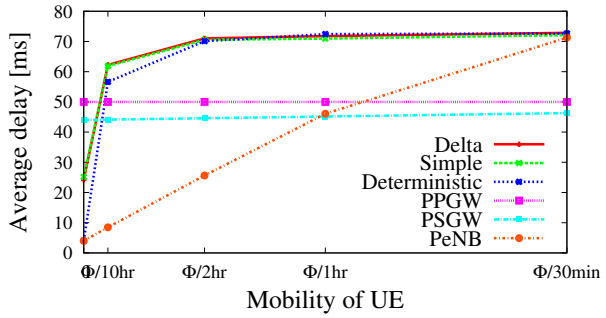

Figure 2.12:]
[Average delay in Scenario 1 ($\rho = 0$)]Average delay in Scenario 1. $\rho = 0$ means our methods only consider load balancing. Delta, Simple, and Deterministic result in larger delay than PPGW, PSGW and PeNB except the immobile case.

of mobility, a UE moves to another cell. Then, increased response delay decreases the activity and as a result a random walk will be initiated to find a better location. Since it is necessarily not a good choice to appoint an ADMME at an eNB of the new cell causing the same problem very soon due to the high mobility of the UE, selection converges to a solution to select an ADMME at an SGW.

As for Deterministic, because of strict restriction on ADMME switching, the average delay cannot be reduced enough except the immobile case, in which there are no ADMME switching. On the contrary, Delta can stochastically change a serving ADMME even with a large $\delta(h)$ for $N(h) \geq N_{SGW}(h)$, which results in as small delay as Simple. At the same time, the number of ADMME switching in Delta is reduced comparing to Simple as shown in Fig. 2.11. Interestingly, even with $\rho = 1$, the fairness of our proposed methods is very high as shown in Fig. 2.10. This is because that in the random search phase, an ADMME on nodes with higher capacity, i.e. the PGW and SGWs, is selected often due to the hierarchical topology.

Next Figs. 2.12, 2.13, and 2.14 show results of Scenario 1 with $\rho = 0$, where our methods only consider load balancing. Results of PPGW, PSGW, and PeNB are the same as in Fig. 2.9 through Fig. 2.11. It is apparent that Delta, Simple, and Deterministic result in larger delay than PPGW, PSGW and PeNB except the immobile case. It is because our proposal is likely to select ADMMEs

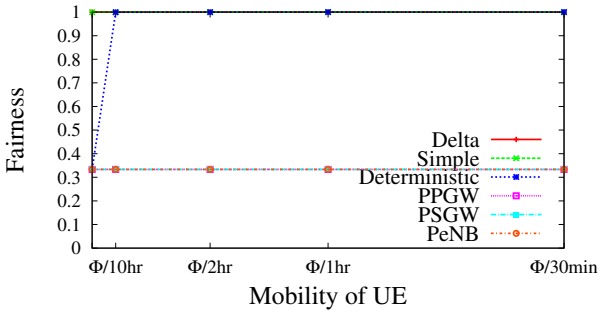

Figure 2.13: Fairness in Scenario 1. $\rho = 0$ means our methods only consider load balancing. As a result of sacrifice of delay, load is fully balanced among nodes in Delta and Simple.

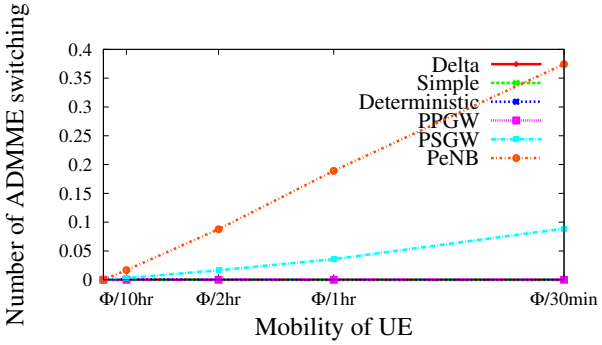

Figure 2.14: Number of ADMME switching in Scenario 1. $\rho = 0$ means our methods only consider load balancing. PeNB which has the minimum delay for UEs with low mobility suffers from the largest C-plane overhead. In Delta, Simple, and Deterministic, there is no or quite few ADMME switching after convergence, since the fairness is already satisfied.

on a PGW and SGWs, i.e. distant nodes having larger capacity than eNBs, for the sake of load balancing. A reason why the average delay of immobile UEs is smaller than PSGW and larger than PeNB is that an ADMME selected for each UE is mostly located at either of the nearest eNB or the nearest SGW. On the contrary, because of mobility, an ADMME is not necessarily located at a nearest node in the other cases. As a result of sacrifice of delay, load is fully balanced among nodes in Delta and Simple as shown in Fig. 2.13. A reason why Deterministic cannot achieve high fairness for immobile UEs is that there is no ADMME switching. As shown in Fig. 2.14, PeNB which has the minimum delay for UEs with low mobility suffers from the largest C-plane overhead.

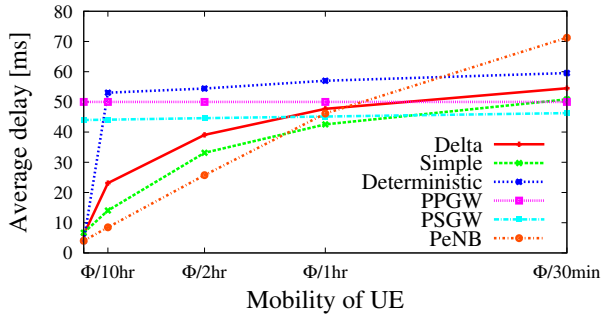

Figure 2.15: Average delay in Scenario 1. $\rho = 0.5$ means both of delay and load are considered in our methods. For low mobility cases average delay is more dominant in ADMME selection. Deterministic inferiors to the others from viewpoint of delay. The superiority of Simple to Delta in this figure becomes larger than in Fig. 2.9.

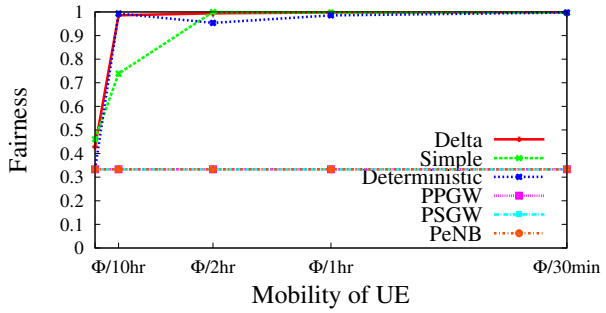

Figure 2.16: Fairness in Scenario 1. $\rho = 0.5$ means both of delay and load are considered in our methods. Delta and Simple achieve as high fairness as with $\rho = 0$ except for low mobility cases.

On the contrary, in Delta, Simple, and Deterministic, there is no or quite few ADMME switching after convergence, since the fairness is already satisfied.

Then Figs. 2.15, 2.16, and 2.17 show results of Scenario 1 with $\rho = 0.5$ where both of delay and load are considered in our methods. By comparing with cases of $\rho = 1$ and $\rho = 0.5$, we can find that the average delay and the number of ADMME switching per UE are similar to those with $\rho = 1$ while the fairness is close to that with $\rho = 0$. Regarding fairness, Delta and Simple achieve as high fairness as with $\rho = 0$ except for low mobility cases, where average delay is more dominant in ADMME selection. Apparently Deterministic inferiors to the others from viewpoint of delay,

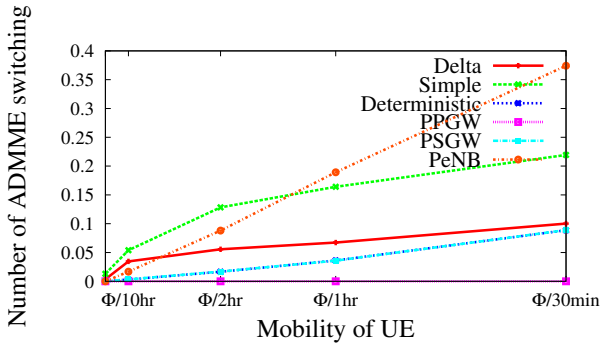

Figure 2.17: Number of ADMME switching in Scenario 1. $\rho = 0.5$ means both of delay and load are considered in our methods. Deterministic has the smallest C-plane overhead. Simple leads to much higher C-plane overhead than Delta.

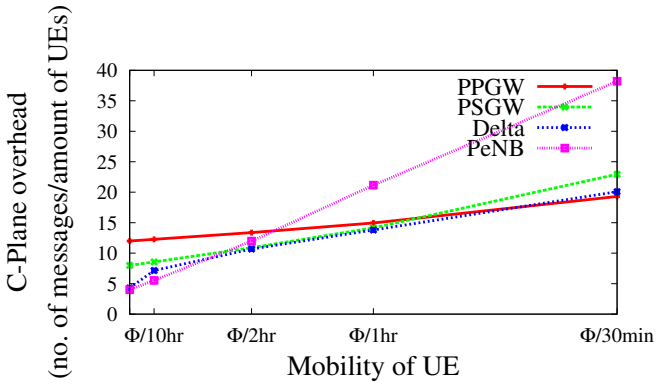

Figure 2.18: Average C-plane signalling overhead in Scenario 1. $\rho = 0.5$ means both of delay and load are considered in our methods. Average C-plane signalling overhead is the average number of messages multiplied by the average number of hops which a message takes per UE per hour for C-plane mobility management including messages both for regular C-plane mobility management exchanged between a UE and a current ADMME and for ADMME switching.

while it has the smallest C-plane overhead. The superiority of Simple to Delta in Fig. 2.15 becomes larger than in Fig. 2.9, but Simple leads to much higher C-plane overhead than Delta as shown in Fig. 2.17.

Finally, we compare the average C-plane signalling overhead of Delta, PPGW, PSGW, and PeNB in Scenario 1 in Fig. 2.18. Although the distance between a current ADMME and a UE is close and thus the C-plane signalling overhead for regular mobility management is small with

PeNB, frequent ADMME switching incurs considerable overhead. As a result, the average C-plane signalling overhead increases more than the other methods as the mobility of UE becomes higher. On the contrary, there is no ADMME switching with PPGW. Therefore, the C-plane signalling overhead contains only regular control messages. Since highly mobile UEs send TAU and Handover requests more often, the overhead increases in proportion to the UE mobility even with PPGW. Intuitively, as being supported by Fig. 2.18, the best methods are PeNB for immobile and low mobility UEs, PSGW for UEs with moderate mobility, and PPGW for highly mobile UEs, respectively. It is notable that Delta draws the lower envelope of these three lines. It proves that our proposal can successfully select an ADMME appropriate for a UE taking into account mobility characteristics.

For Scenario 2 where UEs with stay timer T_s=0.5 hrs and 10 hrs equally coexist, we set $\rho = 0.5$. Results are summarized in Fig. 2.19. There are five or three sets of bars in the figures. A set of bars at the center correspond to results averaged over all UEs in Scenario 2 (S2). Two sets of bars labelled as S2 are averages of UEs of $T_s = 0.5$ hrs and $T_s = 10$ hrs, respectively. For comparison purposes, we also show results of Scenario 1 (S1). The confidence interval of 95% is shown for each bar. Since it is not possible to derive the fairness index for each of stay timer settings, the second graph in Fig. 2.19 has only three sets of bars.

We can easily find that the sets of bars in T_s=0.5 hrs and T_s=10 hrs in Scenario 2 are close to those in Scenario 1 respectively. It means that our proposal can provide UEs with mobility management control appropriate to their mobility characteristics even in a heterogeneous case. One significant difference is in the number of ADMME switching of Simple for mobile UEs with T_s=0.5 hrs. A reason why Simple has less ADMME switching for mobile UEs in Scenario 2 is segregation. As discussed, ADMMEs on SGWs and eNBs are likely to be selected for UEs with low mobility and those on a PGW and SGWs are preferred for UEs with high mobility as a result

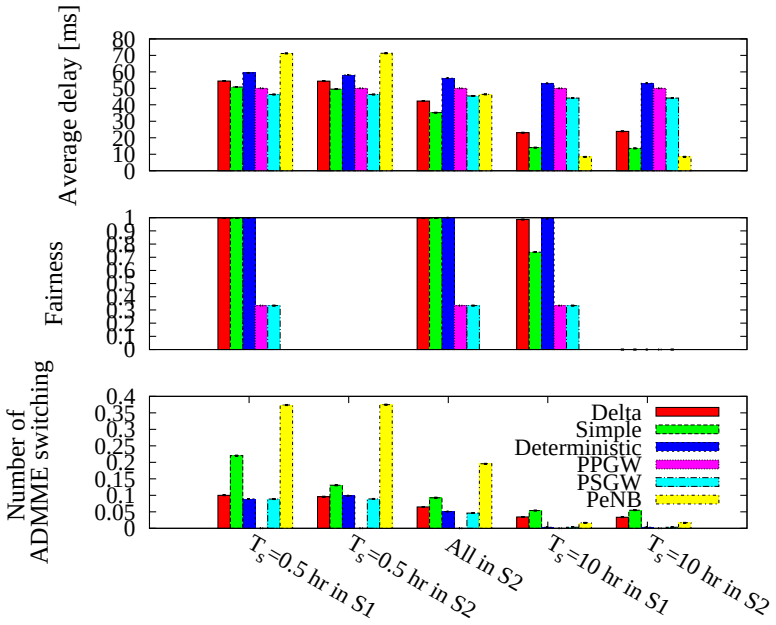

Figure 2.19: Results in Scenario 2 v.s. Scenario 1. $\rho = 0.5$ means both of delay and load are considered in our methods. Scenario 2: heterogeneous mobility, 50% UEs(7400): T_s=0.5 hours, the left 50%(7400): 10 hours. A set of bars at the center correspond to results averaged over all UEs in Scenario 2 (S2). For comparison purposes, we also show results of Scenario 1 (S1). The confidence interval of 95% is shown for each bar.

of our adaptive selection algorithm. Consequently, ADMMEs on eNBs and SGWs are fully or mostly occupied by UEs with low mobility. Furthermore, ρ=0.5 pushes selection toward unused and available ADMMEs. As a result, the range of selection or random search is limited to SGWs and a PGW in ADMME selection for UEs with high mobility. Therefore, the number of ADMME switching becomes smaller in a heterogeneous case.

Regarding averaged values shown as bars at the center, Simple and Delta achieve smaller delay than PPGW, PSGW, and PeNB, but Simple suffers from the most frequent ADMME switching among our proposed methods. Although the number of ADMME switching is small with Deterministic, the average delay is the largest. Since our proposed methods all accomplish high fairness, from a viewpoint of balance between delay and C-plane overhead, Delta seems the best.

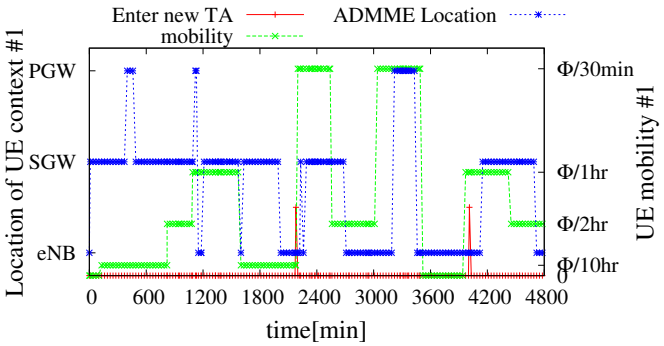

Figure 2.20: Behaviours of ADMME switching for one UE in Scenario 3. $\rho = 0.5$ means both of delay and load are considered in our methods. Scenario 3: dynamic and heterogeneous mobility, T_s changes every 8 hours, and we set ratio of mobile UEs to 100%, 90%, 80%, 70%, 60%. A set of trace data of one UE obtained from one simulation run to have a close look at behaviour of our proposal, Delta.

Finally, we verify our proposal using Scenario 3 where UEs have heterogeneous and dynamic mobility characteristics. Figure 2.20 shows a set of trace data of one UE obtained from one simulation run to have a close look at behaviour of our proposal, Delta and $\rho = 0.5$. The left axis and a blue dotted line show the location of a current ADMME selected for the UE. The right axis and a green dotted line show the mobility of the UE. A red line indicates when the UE moves from one TA to another. In the beginning, selection is under influence of the initial setting, but as time passes, selection becomes consistent with UE's mobility. After about 1200 min, we can see correlation between the location of selected ADMME and the mobility of UE. More specifically, an ADMME at an eNB is selected when the UE is immobile or T_s=10 hrs. When the stay timer changes to 2 hrs, ADMMEs are located at either of an eNB or SGW. In the case of T_s=1, ADMMEs at PGW or SGW is selected. Then when the UE has T_s=0.5 hrs, the PGW is finally selected as a location of a current ADMME.

To evaluate superiority and inferiority of methods in a unified way, we normalize all three measures to fit to the range of [0,1] by Eq. 2.9. For example in Fig. 2.21(a) the normalized delay

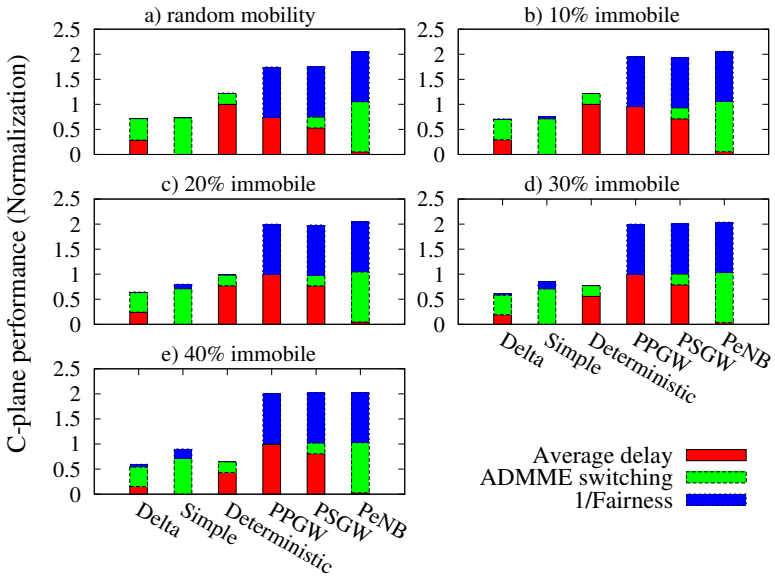

Figure 2.21: C-plane performance in Scenario 3. $\rho = 0.5$ means both of delay and load are considered in our methods. Scenario 3: dynamic and heterogeneous mobility, T_s changes every 8 hours and we set ratio of mobile UEs to 100%, 90%, 80%, 70%, 60%. We normalize all three measures to fit to the range of [0,1] by Eq. 2.9.

of Simple is 0, since it has the smallest delay among all six methods. In the same figure, PPGW, PSGW, and PeNB have the normalized fairness of 1, because their fairness is the lowest. Due to dynamic mobility, we consider averaged values of all UEs independently of their instantaneous mobility.

Independently of the ratio of immobile UEs, PPGW, PSGW, and PeNB have the larger sums, i.e. the worse performance. It is mainly because of the low fairness, but the sum of the normalized delay and number of ADMME switching grows and becomes larger than that of our proposals as the ratio of immobile UEs increases. As we observed from Scenario 1, the average response delay of immobile UEs in our proposals are much lower than PPGW and PSGW. In addition, as for immobile UEs there are no handover or TAU procedures triggered by TA changing, ADMME switching is only carried out in TAU procedures triggered by TAU time out. Therefore, with the increase of the ratio of immobile UEs, the degree of random search and ADMME switching involved are reduced

a lot in our proposals. As a result, normalized relative overhead of PSGW and PeNB increases.

Regarding our proposals, the performance of Deterministic becomes better as the ratio of immobile UEs increases. The reason is as follows. Since there is no ADMME switching for immobile UEs in PSGW, Deterministic does not perform ADMME switching for immobile UEs as well. In addition, since the initial location of ADMMEs is eNBs, the response delay is always the smallest for immobile UEs. On the contrary, because of some degree of freedom of ADMME switching, we do not see such obvious influence in Simple and Delta. However, independently of heterogeneity in mobility of UEs, Delta achieves the smallest and almost stable performance.

Therefore, we can conclude that our proposed autonomous and distributed mobility management schemes, especially Delta, can accomplish autonomous and distributed mobility management leading to smaller delay, higher fairness, and lower C-plane overhead with consideration on dynamic and heterogeneous mobility characteristics of UEs.

2.6 Summary

In this chapter, after stating current research and development activities on the 5G mobile network and distributed management, and the related problems, we present a conceptual architecture of distributed mobility management and an autonomous and adaptive ADMME selection method based on attractor selection model in this distributed architecture. We confirmed that our proposal can autonomously and adaptively allocate an ADMME appropriate for a UE based on attractor selection model according to constantly changing location and mobility of UEs and load status of servers for multiple objectives. Through simulation experiments using three mobility scenarios, we showed that our proposal achieved more than 63% performance improvement comparing to the current method from viewpoints of delay, load balancing, and C-plane overhead under a dynamic mobility scenario. From a viewpoint of delay minimization, the optimal location of ADMME is

the nearest eNB, ie.e PeNB. However it would suffer from frequent changes of a current ADMME and related overhead in the C-Plane, e.g. delay for ADMME switchings. In our proposal, for frequently moving, an ADMME on a PGW is selected. It increases the response delay for its distance but contributes to overhead reduction. Therefore, there is a trade-off between delay and C-plane overhead.

As future work we plan to formulate dependency of our algorithms on UE mobility and C-plane overhead. It enables an ADMME to dynamically and adaptively select an appropriate algorithm and achieve the best performance. In addition, we consider to evaluate the application of ADMME in complex and heterogeneous networks which is close to the real world.

Chapter 3

A Distributed Mobility Management Scheme based on Response Threshold Model in Flat Mobile Architecture

3.1 Introduction

The 3.9G LTE/EPC network adopts the centralized architecture where U-plane an SGW handles the U-Plane, i.e. user traffic, and an MME manages the C-Plane of connected UEs, which leads to traffic congestion in SGW and overload in MME. In addition, excessive bandwidth consumption and large response delay in both of U-plane and C-plane are caused by the physical distances between those management nodes in an EPC network and UEs. Therefore, a lot of research activities to overcome those issues caused by centralized control in the current 3.9G LTE/EPC networks are carried out.

There are some studies focused on the mobility management distribution in C-plane in the 3.9G LTE/EPC hierarchical network architecture, such as DMME [9] and ADMME mentioned in Chapter 2. However, because of the hierarchical architecture of the current 3.9G LTE/EPC networks, C-plane NAS (Non-Access Stratum) messages sent by UEs have to pass along the topology, which sometimes is a long path, for example from an eNB to an SGW, then to a PGW, and finally to the MME at another SGW. It limits the reduction of C-plane response delay and even causes a large C-plane overhead for UE context migrations. Moreover, for accommodating numerous terminal

devices data traffic explodes in PGW/SGW and the cost in U-plane increases greatly.

In order to increase the efficiency and further reduce the cost in a mobile core network, latest researches are oriented from hierarchical to flat architecture like for example DMM [7] [8]. Recent years, partial DMM is a hot topic which adopts a flat architecture and shortens the distance from gateways to UEs by anchoring U-plane traffic closer to user devices. Regarding full DMM, there are only some blueprints for extending the partial DMM to full DMM and they are defective and immature.

In this chapter, we design a novel flat mobile network architecture. More specifically, a mobile network is comprised of a RO (Regional Office) corresponding to a central office of PGW, LOs (Local Offices) serving local areas, and APs (Access Points) corresponding to eNBs. These components of the proposed flat architecture only denote the geographical distribution without any functional or hierarchical relationship between each other. Moreover, MAs (Mobility Anchors) used for U-plane anchor points are deployed in LOs. In a RO an HSS is also deployed for subscription and authentication for all users.

In Chapter 2, we introduce a distributed mobility architecture which is realized on a hierarchical architecture of current 3.9G LTE mobile network. In U-plane, it has a functional and structural hierarchy that consists of a PGW (Packet Data Network Gateway) providing connectivity from UEs (User Equipment) to external networks, SGWs (Serving Gateways) working as a local mobility anchor point for intra-system handovers, and base stations called eNBs. In C-plane the mobility management functions of MME are distributed over a mobile network by the virtualized technology. However, Chapter 3 focuses on the flat architecture which is leaving away from the current standards. Its U-plane adopts a server called MA (mobility anchor) which combines the all or parts of the functionalities of PGW and SGW. All the user data traffic is distributed by the MAs which is closer to the users. Since there is no concentration of user data traffic, the network system

becomes more scalable in flat architecture and it can accommodate larger number of terminal devices. However, ADMME is not flexible enough for this totally different flat architecture and as a result the system performance of autonomous and distributed mobility management in C-plane degrades if we adopt the scheme proposed in Chapter 2.

For the sake of more flexible and dynamic distribution and allocation of C-plane management tasks, we separate C-plane mobility management tasks of the LTE/EPC MME into two parts, i.e. mobility management functions and storage of context information of UEs. Furthermore, they are virtualized as VDMMEs (virtualized and distributed mobility management entities) and DMDs (distributed mobility databases), and distributed over a mobile core network with a help of virtualization technologies. Then, an appropriate pair of those virtualized servers is allocated or launched at APs, LOs, or a RO for a UE taking into account its mobility characteristics such as mobility, location, and communication frequency [33], response delay, and load status of servers, and C-plane overhead. For dynamic and autonomous allocations of VDMMEs and DMDs in accordance with UE characteristics and network status in a flat mobile network, we adopt the response threshold model which is a mathematical model derived from self-organized division of labor in social insects [13]. We verify the effectiveness of our proposed method by comparing to a partial distributed mobility management scheme proposed in other literature.

The rest of this chapter is organized as follow. First, Section 3.2 introduces the related work about distributed mobility management in mobile network. In Section 3.3 we will explain our proposed flat mobile core network system and in Section 3.4 we introduce the self-organized VDMME and DMD selection mechanism and the selection algorithm based on the response threshold model in our proposal. Then, we evaluate simulation results comparing to a partial DMM combining with the existing research DMME scheme in Section 3.5. Finally, we summarize this chapter and list the future work in Section 3.6.

3.2 Related Work

As mentioned in Section 3.1, problems with the centralized architecture in current 3.9G LTE/EPC networks have drawn a lot of attentions in industry, academia, and government, e.g. METIS (Mobile and wireless communications Enablers for Twenty-twenty Information Society) of Europe and 5G forum of Korea [5] [6]. A direction toward distributed mobile network architecture is a hot topic, for example DMM (Distributed Mobility Management) approaches which adopt flat mobile network architecture being investigated by IETF (Internet Engineering Task Force) [7] [15], and a software distributed architecture of DMME (Distributed Mobility Management Entity) [9]. In this section, we discuss prior activities on distributed mobility management.

In the centralized approaches, both of U-Plane and C-Plane managements are done by a single entity per UE, e.g. the centralized mobility anchor of HA (Home Agent) and LMA (Local Mobility Anchor) in MIPv6 (Mobile IPv6) and PMIPv6 (Proxy Mobile IPv6). In contrast, in DMM solutions, C-plane mobility management of UEs can be carried out in a decentralized way and mobile data traffic can also be distributed by topologically distributing mobility anchors [7] [8]. There is a consensus that existing IETF IP mobility protocols are used for development of DMM protocols as it can ease migration from the current hierarchical mobile network to the future flat mobile network.

MIPv6 is a host-based IP mobility protocol and PMIPv6 is a network-based IP mobility protocol. Basing on these protocols, mobility management can be distributed at multiple levels of a mobile network, i.e. distribution in a mobile core network, in an access network, and in host level. Mobility anchoring functions duplicated and distributed in different parts of a mobile network can eliminate negative impacts, such as a single point of failure, brought by the centralized mobility anchoring architecture. Corresponding to the different distribution scenarios, the scale of the specific geographical area being covered by each mobility anchor is different and mobility anchors are confined down to nodes closer to UEs resulting in a more flat network architecture.

DMM approaches can be classified to partially distributed and fully distributed. A partial DMM approach considers separation of C-plane and U-plane and / or dynamic mobility management [8] [19] [20] [21]. As C-plane signalling traffic is much lower than data traffic in U-plane, separation and distribution of U-plane enables effective data traffic distribution and takes advantages of the centralized access and control in C-plane. Recently, IETF DMM Working group has been carrying out researches mainly on such partial DMM approaches, for example Distributed Mobility Anchoring (DMA) as an extension of PMIPv6 [22] and Dynamic Tunneling for DMM as a network-based DMM [23]. Speaking of full DMM approach in which DMM are applied on both C-plane and U-plane, however, there are just some blueprints for extension of partial DMM to full DMM [16] and summarizations of its requirements [15]. In other words, researches on full DMM approaches are in an early phase as the number of studies on network-based fully DMM approaches is not enough yet. For example, there is a proposal to adopt P2P techniques or broadcast / multicast for data delivery, but they are not flexible or efficient because of difficulty in sharing consistent mobility information among all entities in dynamic and distributed environment.

On the other hand, ADMME [34] and DMME [9] consider distribution of an MME, that is a mobility management entity in C-plane. ADMME is a network-based partial DMM approach where virtualized MME servers are deployed at a PGW, SGWs, and eNBs, and one appropriate server is dynamically and adaptively allocated to a UE according to its mobility characteristics and a management policy. Regarding to DMME, first the tasks of MME is split into two parts, i.e. mobility management functions to process C-plane messages and storage of context information of UEs. Then, mobility management functions are duplicated and replicas are distributed over a mobile core network. DMME selects a replica appropriate for a UE to mitigate latency in message exchanges, but selection takes into account mobility characteristics only. As such, some server locations would be overloaded and response delay becomes large. In addition, DMME cannot

handle replica switching for UEs in ECM_CONNECTED state. Furthermore, databases used for storing UE contexts are centralized in DMME scheme. It is to avoid C-plane overhead in user context migration among distributed databases, but it suffers from C-plane message concentration. Both of ADMME and DMME are designed for current 3.9G LTE/EPC networks. Therefore, the degree of improvement is limited as mentioned in the previous section.

Considering the above, we propose a network-based full DMM approach, which distribute both of C-plane and U-plane over a flat mobile core network.

3.3 Architecture

In this section, we introduce our novel flat mobile network architecture, which can handle explosion of mobile data traffic, accommodate numerous of terminal devices, and deal with heterogeneous applications.

Figure 3.1 illustrates our conceptual flat architecture of distributed mobility management. In our architecture, a mobile core network consists of APs, LOs and a RO. They correspond to eNB, a local server site covering a specific local service area, and a central office of PGW in a 3.9G LTE/EPC network, respectively. Each AP is associated with an LO covering the area where the AP resides. For U-plane traffic management, an AP is connected with an MA in the corresponding LO. When an error occurs at an MA, another MA in the same LO takes over its management tasks.

An HSS is located in the RO as shown in Fig. 3.1. Regarding C-plane management, we deploy two kinds of virtual servers, i.e. VDMME(virtualized and distributed mobility management entity) and DMD(distributed mobility databases). VDMMEs is responsible for processing the C-plane message related with UE mobility management without storing UE mobility information. On the contrary, DMDs just handle the storage and management of context information of UEs. In our architecture, any of APs and LOs can serve one or more VDMMEs. Similarly, DMDs can be

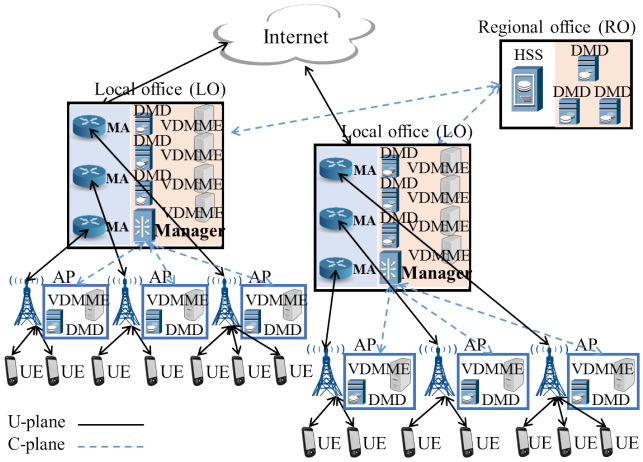

Figure 3.1: Architecture of flat mobile network. A mobile core network consists of APs, LOs and a RO. Each AP is associated with an LO covering the area where the AP resides.

located at APs, LOs, and a RO. In order to deliver the arriving requests to all the servers in LO as soon as possible, we deploy a manager in each LO which can duplicate all requests and deliver the copies to all VDMMEs and DMDs in the same LO.

Distribution of mobility management entities and databases enables mitigation of delay and overhead in the C-plane by their dynamic and flexible allocation to UEs. Generally speaking, a VDMME and a DMD at the nearest location, i.e. an AP of a resident cell, are optimum from a viewpoint of delay minimization. However, when a UE frequently moves between cells and TAs, such allocation causes considerable C-plane overhead in migration of UE context information from one DMD to another. Switching between VDMMEs incurs additional message exchanges. Therefore, for a UE with high mobility, a DMD at a RO or LO and a VDMME at an LO should be selected. In addition, it should be avoided to allocate an overloaded server because it increases the response delay. In the next section, we propose a mechanism and an algorithm to accomplish dynamic and adaptive allocation of a VDMME and a DMD appropriate for a UE taking into account a variety of factors, such as mobility and location of a UE and load status of servers.

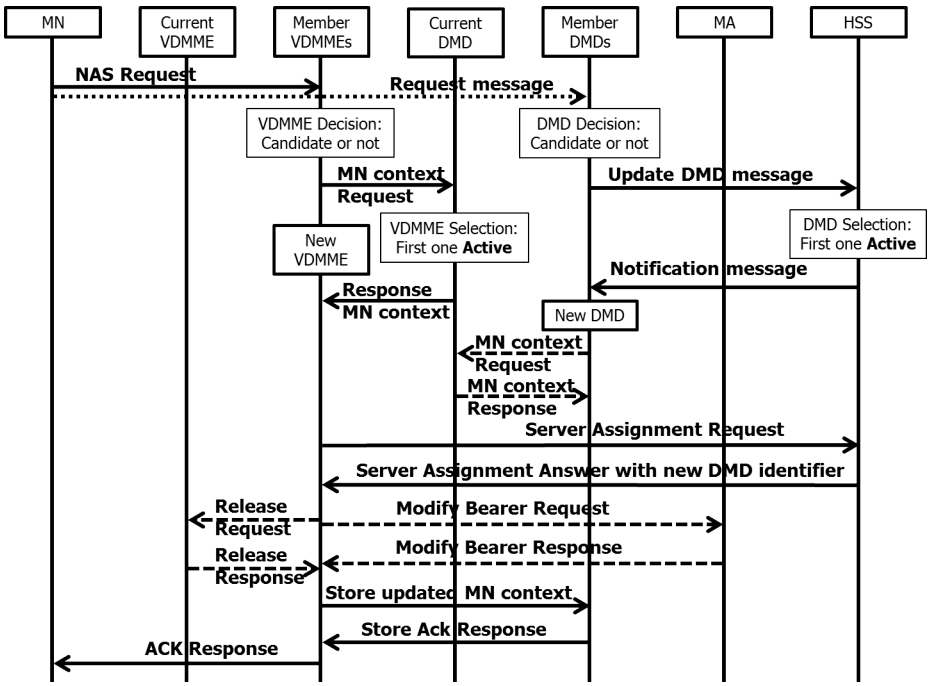

Figure 3.2: C-plane message processing procedure. A C-plane message processing procedure in case that both of a current VDMME and a current DMD delegate mobility management tasks to a new VDMME and a current DMD, respectively.

3.4 Autonomous Allocation of VDMME and DMD

In this section, we first explain a mechanism of dynamic allocation of a VDMME and a DMD. Then we give details of an algorithm to select and allocate an appropriate pair of virtual servers to a UE. Finally, we explain the new functions required in mobile networks.

3.4.1 Mechanism

Allocation of servers is initiated by a request message of attaching, handover, or TAU sent from a UE. First when an AP of a cell where a UE resides receives a request message from the UE, it forwards the request to a VDMME and a DMD which are allocated to the requesting UE. We call them a current VDMME and a current DMD, respectively. In the meanwhile, the AP forwards the request to a VDMME and a DMD at itself, a manager in the nearest LO which will duplicate the

request and send the copies to all VDMMEs and DMDs in the LO, and a DMD in a RO serving the corresponding TA. We call all those VDMMEs and DMDs, including current ones, which receive the request or the copies, member VDMMEs and member DMDs, respectively.

Figure 3.2 shows a C-plane message processing procedure in case that both of a current VDMME and a current DMD delegate mobility management tasks to a new VDMME and a current DMD, respectively. We call such delegation VDMME and / or DMD switching. On receiving a request sent by a UE, each of member VDMMEs and DMDs autonomously determines whether to serve the requesting UE or not by using our algorithm explained in the next section. Those VDMMEs that decide to become a new current VDMME, which we call candidate VDMMEs, notify the current DMD of their decision. One candidate VDMME is selected by the DMD if it satisfies the conditions, such as priority of the VDMMEs deployed at associating AP and LO, the first one of a notification received by the current DMD etc. Similarly, candidate DMDs send notifications to an HSS and an HSS selects one among them as a new current DMD. Results of selection are only notified to selected candidates.

A selected VDMME receives a notification message with UE context consisting of entity identity, UE state, bearer context etc., from the current DMD. It takes over the role of mobility management from a current VDMME and becomes an active VDMME. At the same time, a selected DMD sends a UE context request to an old current DMD and becomes a new current DMD after receiving all information about the UE. Then the new current VDMME sends server assignment request to an HSS server and then obtains the identifier of the new current DMD. It also sends a release request to the old current VDMME. The new current VDMME additionally sends a control message to MAs for setup or modification of bearers if the requesting UE moved between cells for example. After that, the new current VDMME updates the UE context and sends it to the new current DMD.

3.4.2 Basic Behaviors of VDMMEs and DMDs

When a C-plane procedure is triggered, a UE sends request messages to all the member servers as we have introduced above. A tuple (i, h, $S_i(h)$, $s_i(h)$, DMD Switching frequency, current DMD identifier) is embedded in the head of request messages, where i is the UE identifier and h is a sequence number of the procedure. $S_i(h)$ and $s_i(h)$ is the serving demand intensity of UE i for VDMME and DMD, respectively. In addition, the normalization delays calculated by history estimated delays of past W procedures depending on the server locations are also embedded in the request messages.

The flowchart shown in Fig. 3.3 illustrates the behaviors of a member VDMME on receiving a C-plane request sent by a UE. On receiving a request, member VDMME j first evaluates the decision making algorithm explained in section 3.4.3. If member VDMME j decides to serve UE i as a candidate, it sends a UE context request to the current DMD. Next, if candidate VDMME j satisfies the conditions for being selected as a new current VDMME among candidates by the current DMD, it will receive the response message with the UE context from the current DMD. Then it sends an assignment request to HSS and after receiving response from HSS VDMME j sends the release request to the current VDMME. From that, the VDMME j turns to active state for serving UE i which means VDMME j becomes a new current VDMME and takes over the mobility management for UE i. On the contrary, if VDMME j decides not to become a candidate or it is not selected by the current DMD, VDMME j that is not the current VDMME or receives a release request from a new VDMME goes back to idle state.

Similarly, when a member DMD receives a request message from a UE, it exhibits the behaviors as the flow chart shown in Fig. 3.4. If member DMD k is decided to serve for UE i as a candidate by evaluation of the decision-making algorithm, it sends an update DMD message to HSS. Next, if candidate DMD k is selected by the HSS as it satisfies the selection conditions, it will receive

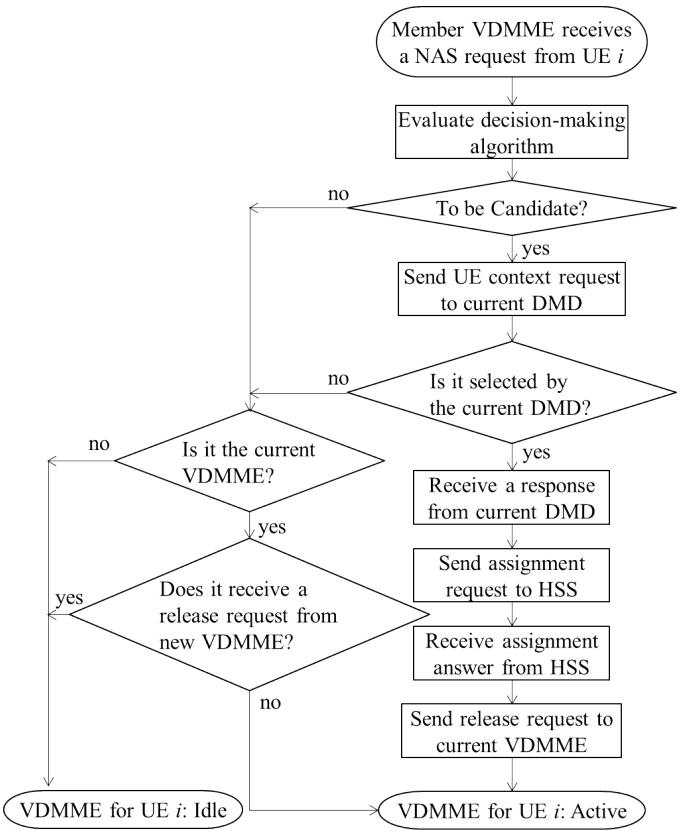

Figure 3.3: Behaviors of member VDMMEs on receiving request messages. On receiving a request, member VDMME j evaluates decision-making algorithm. VDMME j: $X_{ij} = true$, sends a context request to current DMD. VDMME j: a candidate selected as a new current one, receives the response with UE context from the current DMD. It sends an assignment request to HSS, and a release request to the current one. VDMME j: active state.

a notification message from the HSS server. Then, if DMD k is the current DMD, it maintains active state, otherwise it needs to send a request message to the current DMD to acquire the UE context. From that DMD k turns to active state and takes over the storage task of UE i. Conversely, if DMD k is not a candidate or selected by HSS, DMD k which is the current DMD and receive a UE context request from new DMD, sends UE context to the new DMD and turns to idle state, otherwise, it maintains active state. When the current DMD turns to idle state and migrates the context information to the new DMD, it does not instantly delete the UE context but maintains for

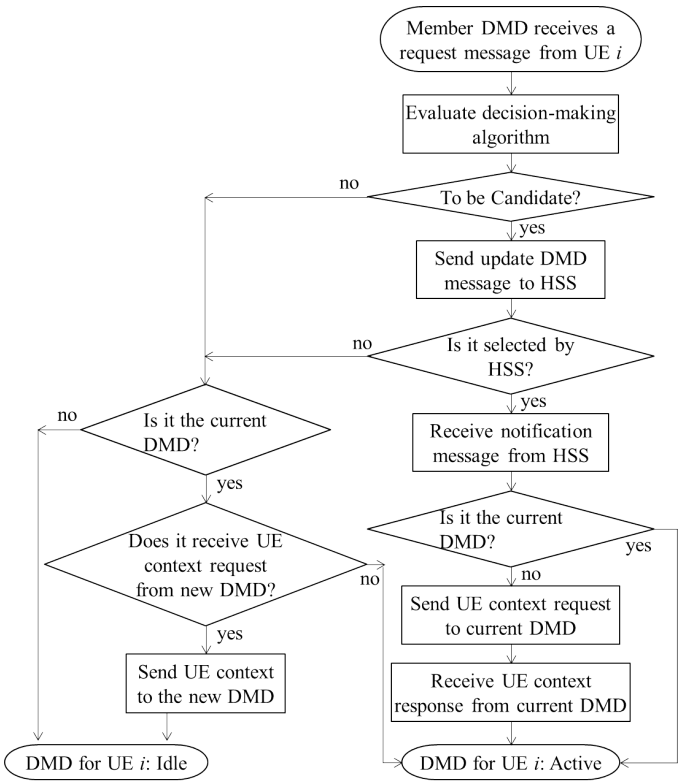

Figure 3.4: Behaviors of member DMDs on receiving request messages. Member DMD k: becomes a candidate by the decision-making algorithm, and sends an update DMD message to HSS. Candidate DMD k: selected by HSS, will receive a notification from HSS. DMD k: if active, maintains the state, otherwise asks for the UE context from the current one.

a short valid time interval to ensure seamless mobility management for UEs.

In our proposal, each UE maintains two tables of estimated delay as Table 3.1 for VDMMEs and DMDs respectively, which are the history records of estimated response delay. The estimated response delay is an estimation of the sum of the transmission delay and processing delay. Transmission delay is sum of propagation delay between nodes during a C-plane procedure based on the mechanism we explained in section 3.4.1. Regarding the processing delay, it is calculated based on utilization of the server by using a queuing model. Additionally, a set $\vec{S}$ of tuples (i, h, θ_{ij}, A_{ij}) are maintained in each server node which are updated by each corresponding request. h is the sequence of a request message. $s_i(h)$ is the demand intensity of UE i calculated at the

Table 3.1: An example of history records of estimated delay. The estimated response delay is an estimation of the sum of the transmission delay and processing delay. Transmission delay is sum of propagation delay between nodes during a C-plane procedure based on the mechanism.

Time	estimated delay of member VDMMEs		
	VDMME at AP	VDMME at LO	current VDMME
h	delay 1	delay 2	delay 1
h-1	delay 1	delay 2	delay 3
h-2	delay 1	delay 2	delay 2
...	...	...	...
h-W+1	delay 1	delay 2	delay 3

beginning of the h-th procedure, which represents a degree that a server wants to perform the mobility management of UE i. Moreover, each server maintains a $\vec{X}$ of X_{ij}, a set $\vec{\theta}$ of θ_{ij}, and O_i=$active$, $idle$. i is UE identifier and j is server number. If a server j is decided to serve for UE i as an active one, X_{ij}=$true$, if not $X_{ij} = false$. $\theta_{ij}(0 \leq \theta_{ij} \leq \theta_{max})$ is a threshold for hesitation of server j to serve for UE i. O_i=$active$, $idle$ is an operational mode which control the server node j to serve for UE i or not. If a member server does not receive any request message from a UE after its valid timer times out, it will delete the tuple information of the UE.

3.4.3 Decision-making Algorithm

As we introduced in Section 3.3, in this chapter we focus on the distributed mobility management in flat architecture which has much different functional and structural hierarchy from the current 3.9G LTE/EPC network. There are more servers and most of them are deployed closer to UEs than hierarchical architecture. Sometimes multiple distributed servers are appropriate for serving one UE in flat architecture. Accordingly, the algorithm is required to accomplish autonomous, adaptive, and consistent decision-making by each individual server for serving UEs and realize

multi-objectives optimization as a whole in our flat architecture. If we adopt other algorithms of distributed task allocation such as analyzed in [11] for VDMME and DMD allocation, one member server received a request from a UE will determine to serve for the UE or transfer it to another member server based on the effort level for optimizing the whole system performance. As the algorithm become more complicated due to the large number of servers and tasks, the overhead of computation and communication is considerable for solving the problem of optimization of task allocation to VDMMEs and DMDs. In contrast, bio-inspired stochastic algorithms can achieve potential optimization with high computationally efficient. Therefore, we also adopt a bio-inspired algorithm for autonomous and adaptive allocation of appropriate VDMMEs and DMDs for UEs, which has high adaptability and flexibility.

Chapter 2 adopted the attractor selection model for autonomously and adaptively allocating appropriate ADMMEs for UEs in the current hierarchical architecture of 3.9G LTE/EPC networks. If the current ADMME is not appropriate, by using the attractor selection, a new appropriate ADMME cannot be selected by the current server instantly except after several random walks, while these random walks can be tolerated. If we use a response threshold model for ADMME selection, the convergence time is a little faster than attractor selection. However, its C-plane overhead of computation for the decision-making of each member server and communication between multiple selected servers and the current ADMMEs increase a lot due to long distances between servers in hierarchical architecture. Conversely, as there are much better solutions of allocation of VDMMEs and DMDs due to a large combination space in our flat architecture comparing to the allocation of appropriate ADMMEs, random walks before finally allocating an appropriate pair of those virtualized servers based on the attractor selection model leads to considerable convergence time and overhead. When we apply the response threshold model to implement dynamic and autonomous allocations of VDMMEs and DMDs, the appropriate servers are easier to be selected which

contributes to accomplish a higher and faster convergence comparing to the attractor selection model.

Therefore, to implement autonomous and adaptive decision-making on whether to become a candidate or not by member VDMMEs and member DMDs, we adopt a biologically-inspired mathematical model of division of labors in insect societies [13]. Each individual in a colony are autonomously engaged in a task or a number of tasks as a worker without any deterministic policy or centralized control. As a consequence of autonomous decision making, the number of workers can be well adjusted to satisfy the demand of each task simply by mutual and local interactions among individuals. In the model, the probability with which an idle individual i engages in a task as a worker is given by $s^2/(s^2 + \theta_i^2)$, while a worker stops performing a task with probability q. The scalar s is called stimulus, corresponding to the demand that a task is performed. s increases with the certain constant rate if it is not satisfied while decreases proportionally to the number of workers. This is because, the task is not performed by enough individuals, or at high enough rate. θ_i is a threshold of individual i, meaning the degree of hesitation in performing a task. With the increase of number of individuals engaging in the task, the stimulus s decreases so that non-workers are difficult to become workers with a low probability.

We regard VDMMEs and DMDs as individuals and mobility management of a UE as a task to apply the response threshold model to our problem. Furthermore, we extend the model to incorporate multiple factors. In the following details of autonomous decision making based on the response threshold model are given. First, in sending the h-th request, UE calculates two stimulus values $S_V(h)$ and $S_D(h)$ for VDMME selection and DMD selection, respectively. The stimulus $S_V(h)$ is derived as,

$$S_V(h) = S_V(h-1) + \delta_V - \alpha_V(h), \tag{3.1}$$

$$\alpha_V(h) = D_V(h)\{1 + N_V(h-1)\}, \tag{3.2}$$

where δ_V ($\delta_V \geq 0$) is an increasing rate of demand intensity. $D_V(h)$ and $1 + N_V(h-1)$ express the goodness of the current VDMME regarding the response delay and the number of VDMME candidates, respectively. $D_V(h)$ is given as,

$$D_V(h) = \frac{\sum_{k=1}^{\min(h,W)} \rho\{d_{Vmax}(h-k) - d(h-k)\}}{\sum_{k=1}^{\min(h,W)} \rho\{d_{Vmax}(h-k) - d_{Vmin}(h-k)\}}, \tag{3.3}$$

where W is a window size. $d_V(h)$ is observed response delay at the h-th request. It is derived as the duration from emission of the h-th request to reception of a response from the current VDMME and DMD. Thus, the observed response delay includes both the sum of propagation delay reflecting the distance between nodes and processing delay at servers reflecting load status. $d_{Vmax}(h-k)$ and $d_{Vmin}(h-k)$ are the maximum and minimum estimated response delay of member VDMMEs in the nearest LO or associated AP in the $(h-k)$-th procedure, respectively. $N_V(h-1)$ is the number of candidate VDMMEs in the $h-1$-th procedure.

The stimulus $S_D(h)$ of DMD selection is derived in the similar way with a difference in derivation of $\alpha_D(h)$.

$$S_D(h) = S_D(h) + \delta_D - \alpha_D(h), \tag{3.4}$$

$$\alpha_D(h) = D_D(h)\{1 + N_D(h-1)\}, \tag{3.5}$$

where $N_D(h-1)$ is the number of DMD switching that a UE observed in the past W requests.

For autonomous decision-making based on the response threshold model, each server maintains a set of three variables X_i, θ_i, and O_i for UE i. When a server is a candidate, $X_i = true$. Otherwise $X_i = false$. θ_i is a threshold value ($0 \leq \theta_i \leq \theta_{max}$) indicating the degree of hesitation in serving UE i. O_i expresses the operational mode. When server i is serving UE i as a current server, $O_i = active$. Otherwise $O_i = idle$. Therefore, for UE i, only one pair of VDMME and DMD have

$X_i = true$ and $O_i = active$ in a network.

When a server whose $X_i(h-1)$ is $false$ receives the h-th request from UE i, it becomes a candidate with probability given below.

$$P(X_i(h) = false \rightarrow X_i(h) = true) = \frac{S_i(h)^2}{S_i(h)^2 + \theta_i(h)^2 A_i(h)}. \tag{3.6}$$

$S_i(h)$ is either of $S_V(h)$ at a VDMME or $S_D(h)$ at a DMD. $A_i(h)$ biases the threshold. A VDMME derives $A_i(h)$ as,

$$A_i(h) = w_d d_i(h) + w_u U(h), \tag{3.7}$$

where w_d and w_u are weighting parameters. $d_i(h)$ is normalized response delay derived by $d_i(h) = 1 - D(h)$, in which $D(h)$ is calculated similarly to $D_V(h)$ but using the estimated delay that the UE receives the response from the VDMME instead of the current VDMME. $U(h)$ is the utilization of the VDMME, derived by $U(h) = \lambda/\mu$ according the queueing theory [35], where λ is the request arrival rate defined as the number of the requests arriving at the VDMME as a current server for UEs per millisecond and $1/\mu$ is the average processing time of a request at the server.

On the contrary, a DMD calculates $A_i(h)$ as,

$$A_i(h) = w_f F_i(h) + w_d d_i(h) + w_u U(h), \tag{3.8}$$

where $F_i(h) = 1 - e^{-(b-H)^2/W}$ is related to the mobility of a requesting UE. H is the number of UE migrations between cells in the past W requests. Parameter b is defined depending on a location of a DMD, i.e. 0 for AP, 3 for LO, and 8 for RO. When H is close to b, $F_i(h)$ is small and the probability to become a candidate is high. Thus, for example, a DMD in an AP becomes a candidate with high probability for an immobile UE.

On receiving a request, a server whose $X_i(h-1) = true$ changes to $X_i(h) = false$ at a constant quitting probability q_i ($0 \leq q_i \leq 1$). For a candidate server with $O_i(h) = idle$, when the activity α of current server which can be derived by the information embedded in a request is larger than the parameter δ, its quitting probability decreases by,

$$q \leftarrow \begin{cases} q * e^{(c(D(h)-\beta)^d)}, & \text{if } D(h) \geq \beta, \\ 0, & \text{if } q \leq 0, \end{cases} \tag{3.9}$$

where $c = 6$, $d = 1/4$ for VDMME and $1/2$ for DMD and $\beta = 0.4$. It contributes to task rotation among servers. If the appropriate candidate with high performance has been selected, the other candidates become easier to turn back to $X_i = false$ state by a larger quitting probability, which can avoid the redundant switchings between candidates to reduce the C-plane overhead of server switchings and improve the system performance. It is because, as the appropriateness of candidates constantly changes corresponding to the dynamic changes of UE mobility and network status, the inappropriate candidates will become to a state of $X_i = false$ with a high quitting probability and not be selected accidentally. We also adopt a reinforcement mechanism of the response threshold model. More specifically, a server adjusts the threshold as,

$$\theta_i(h+1) = \begin{cases} \theta_i(h) - \xi, & \text{if} X_i(h) = true \\ \theta_i(h) + \phi, & \text{if } X_i(h) = false \end{cases} \tag{3.10}$$

where ξ and ϕ are parameters of the speed of differentiation.

3.4.4 New Functions Equipped in Mobile Networks

The following new functions are required to implement proposal keeping consistency with 3GPP specifications as much as possible.

VDMME and DMD distribution function: Each UE must be able to be associated with any VDMME and DMD deployed in a mobile core network. That is, association must be flexible and

dynamic. For this purpose, APs periodically update a mapping table for GUMMEIs (Globally Unique MME Identity) , i.e. DMM Identifiers, and IP addresses, for identification of allocated VDMME and DMD. Details are similar to the mapping method in [34].

VDMME and DMD switching functions: considering 3GPP specifications [25], VDMME switching is carried out in attach procedure for UE registration, TAU (Tracking Area Update) procedure for location update, and inter-APs handover procedure. On the contrary, UE contexts are migrated between DMDs only in Attach and TAU procedures.

Information collection function: Information required in decision making of VDMME and DMD switching include load status of servers, inter-node delay, and UE characteristics. Load status of servers is expressed in the form of utilization of a server, derived as the ratio of the request arrival rate to the processing rate based on the queueing theory [35]. The request arrival rate is the number of arriving requests per ms which can be easily counted by server itself. Inter-node delay means the delay between pair of network entities, i.e. UE, AP, MA, VDMME, and DMD, which is derived from the timestamps embedded in request messages by the current DMD and HSS. UE characteristics refer to location, mobility and communication frequency of UEs. Besides of current DMD, the C-plane requests are also forwarded to the servers at the locations of connecting AP, LO, and a DMD in RO. Additionally, UE mobility and communication frequency are represented by the number of Handover requests in the observation window W, as Handover is triggered when a connecting UE moves from the coverage are of one cell to the area of another. Accordingly, with the increase of UE mobility and communication frequency, number of Handover procedures in past W procedures becomes large. Number of Handovers for each UE are updated at UEs and also embedded in request messages. The settings of collection interval time and size of observation window also influence VDMME and DMD allocation.

3.5 Evaluation

In this section, we evaluate our proposal from viewpoints of response delay including the transmission delay and the processing delay, and the C-plane overhead caused by the bandwidth consumption.

3.5.1 Simulation Setting

A mobile core network consists of one RO for 3 TAs. There are 7 LOs per TA and 7 cells per LO. As in many other papers, each AP covers a hexagonal cell and 7 cells organizes a hexagonal TA [31] [30], in total, there are 21 LOs and 147 APs. We consider torus topology and each TA shares borders with all the other TAs. There are 3 DMDs in a RO, each of which corresponds to a TA. There are 3 DMDs and 5 VDMMEs in each LO, and one VDMME and one DMD at each AP, respectively.

The number of UEs is 14700 and they are equally distributed over cells at the beginning of simulations. All UEs are always attached but only 30% of them are randomly selected to be connected at every time step, which is 10 minutes long. To simulate mobility, each UE has a stay timer. When a stay timer expires, a UE moves to a random neighbour cell and sets the timer again. A timer interval is randomly set following the Gaussian distribution with average of T_s hrs and variance of 1. To avoid synchronization, we set the initial value of stay timer of a UE randomly from 0 to its T_s. In this paper, we consider three mobility scenarios with different T_s setting. Scenario 1 is a static and homogeneous mobility scenario, in which all UEs use the same mobility: T_s =0.5, 2, or 10 hrs and immobile. In Scenario 2, UEs are divided into four mobility groups of equal size: immobile, low mobility (T_s=10 hrs), medium mobility (2 hrs), and high mobility (0.5 hrs). Scenario 3 is a dynamic mobility scenario where all UEs change their mobility randomly among four types of mobility levels every 8 hrs. A TAU timer is set at 30 mins and initial values

are randomized. Initially, all UEs are allocated servers at the nearest AP as current servers.

Basing on these mobility management settings, we can easily know the average number of requests per time step (10 mins) for all UEs from the simulation results of Scenario 1, as $N_{request}$=7250, 5340, 4980, and 4900 requests per time step for UEs with T_s=0.5 hrs, 2 hrs, 10 hrs and immobile, respectively. In practice, utilization of MME servers in busy hours is about 40% (maximum utilization is 80%) and the MME processing delay is 15 ms (including UE context retrieval of 10 ms) [36] [37]. We assume that for Scenario 1 with high mobility UEs (T_s=0.5 hrs) is similar to the busy-hours situation and mobility management tasks for all UEs are allocated to VDMMEs in all LOs and DMDs in both of LOs and RO. However, because of the simplification of topology setting, the request arrival rate λ' in our simulation is much less than the realistic arrival rate λ. Here, we assume λ' is basically proportional to the actual value as λ=$a\lambda'$. Basing on these assumptions, we can set the mean service rate of VDMMEs by the C-plane request arrival rate λ, for VDMME μ_V=λ_V/ρ=0.00012a/40%=0.00029a ms^{-1}. Additionally, each server can be modelled as an M/G/1 queuing system and the service time of server follows the negative exponential distribution with mean of $1/\mu$ and variance of σ^2. The corresponding response delay can be derived by,

$$E(t) = \frac{1}{\mu} + \frac{\lambda(\sigma^2 + 1/\mu^2)}{2(1-\rho)}, \tag{3.11}$$

where ρ=λ/μ is the utilization of the server, which is set to 40% for VDMMEs and service time of VDMME μ_V=λ_V+1/$E(t)$. In order to set the average response delay of VDMME in busy hours close to 5 ms, we set the value of a to 1150 in our simulation and the mean service time of VDMME as μ_V=0.331 ms^{-1}. Furthermore, by assuming the average response delay of DMD to 10 ms and the request arrival rate to DMD for high mobility UEs as λ_D=0.211, the mean service rate for DMD is μ_D=λ_D+1/10=0.311 and the utilization of DMD in busy hours is about 68%.

Table 3.2: Parameter setting

Parameter	Description	Value
θ	threshold	0.5
ξ	threshold adaptation parameter	0.01
ϕ	threshold adaptation parameter	0.1
δ_V	increasing rate of demand for VDMME	0.5
δ_D	increasing rate of demand for DMD	0.6
q_i	quitting probability	0.01
W	history window	5
w	weighting values in Eq.(8) & (9)	10

For calculating the transmission delay, we set propagation delays between a UE and an AP, between an AP and an LO, and between an LO and a RO at 2 ms, 5 ms, and 15 ms, respectively, basing on the 3GPP specification [36]. Furthermore, delays between neighbouring APs and between neighbouring LOs are set at 4 ms and 10 ms, respectively, taking into account their physical distances.

In the section 3.5.3, we show results of response delay, i.e. delay from emission of a request to reception of a response including processing delay at VDMMEs and DMDs but except for time required for authorization by HSS, the ratio of DMD switching,the number of candidates as an indicator of C-plane overhead, and the average number of signalling messages multiplied by the average number of hops and candidates which a C-plane procedure takes per UE for mobility management, measured at the end of a simulation run of 80 simulation hrs averaged over 100 runs. Parameters used in our proposal are summarized in Table 3.2.

3.5.2 Comparative Method

As a comparative method, we imitate a full DMM scheme with U-plane distribution combining with a DMME (Distributed Mobility Management Entity) method [9] [22], called DMME-CMD. This comparative method is based on the same topology with our proposal. MA servers are deployed at LOs for U-plane distribution. Regarding C-plane, DMME-CMD always use CMDs (Centralized Management Databases) in a RO while DMMEs are deployed at APs, LOs and a RO. When a C-plane mobility management procedure is triggered, a UE sends directly request messages to a DMME at the nearest AP, managers in the nearest LO and the RO which will forward the request message to one of DMMEs with minimum utilization at each location. Those DMMEs at different locations receive the request and become candidates, which check the central database and acquire the UE context from it. If the DMME received the UE context satisfies the conditions such as the number of Handover related to UE mobility, the DMME will preempt the mobility management by updating the central database and become active DMME. UE mobility can be expressed by parameter N_{new} which is the number of C-plane mobility management procedures triggered by a UE entering a new AP in the past W requests. For example, for immobile or low mobility UEs ($N_{new} \leq 1$) a candidate DMME at nearest AP preempts, for middle mobility UEs ($1 < N_{new} \leq 3$) a candidate in LO preempts, and for high mobility UEs ($3 < N_{new}$) a DMME in RO preempts. Then the CMD responds with UE context to the preempting DMME and it sends a release message to the current DMME and an authorizing request to the HSS server. However, DMME is designed not to switch during Handover procedures due to the many subtle MME requirements. In Handover procedure, new DMME only forwards the request to the current DMME and the current DMME sends a modify bearer request to MAs. Finally, after the current DMME received a modify bearer acknowledge message, it sends a response to the new DMME, and then new DMME will take over the mobility management function by updating the UE context in CMD in TAU procedure.

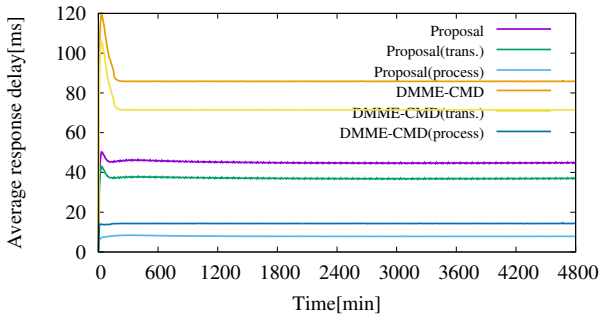

Figure 3.5: Average of total response delay in Scenario 1. Temporal variations of delay averaged over all UEs and simulation runs in Scenario 1 with Ts=0.5 hrs.

For evaluation of DMME-CMD, response delay also includes the processing time in both of a DMME and a CMD. For fair comparison, we deploy the same number of DMMEs at APs and LOs with our proposal. In RO, we deploy 35 DMMEs for each TA. Each DMME replica can be also modelled as a M/G/1 system so that its mean service time is shown as follows,

$$E(t) = \frac{1}{\mu - \lambda}, \tag{3.12}$$

for high mobility UEs DMMEs in RO provides the services for all UEs and its service time of DMME is also assumed to be about 5 ms in busy-hours in reality. For the sake of simplicity, each CMD has the same capacity with DMDs and the retrieve time is also set as 10 ms in busy-hours close to the reality, additionally in case of UEs with high mobility all CMDs receives 7250 requests per 10 minutes in total. Therefore, based on the above equation of queuing model we could know that 22 CMDs should be deployed for serving one TA in the RO for satisfying the requirements.

3.5.3 Results and Discussion

First, in Fig. 3.5, we show ”Proposal” and ”DMME-CMD” corresponds to the total of transmission delay indicated by ”(trans.)” and processing delay indicated by ”(process)” in our proposal and DMME-CMD, respectively. As shown in the figure, there is sharp increases at the beginning of

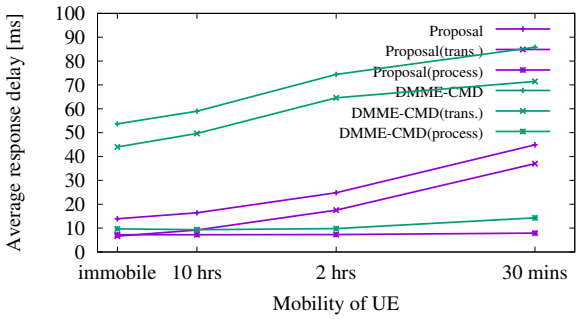

Figure 3.6: Average of total response delay in Scenario 1. All of total response, transmission, and processing delays of our proposal are lower than those of DMME-CMD.

simulation runs, because most of the tasks are not preferred to be allocated to the VDMMEs and DMDs at AP due to the initial setting regardless of UE mobility. However, after certain period of time, average values become stable. It justifies our evaluation based on averaged values at the end of simulation as explained in 3.5.1.

In Fig. 3.6, we compare our proposal with DMME-CMD from a viewpoint of average delay depending on UE mobility in Scenario 1. All of total response, transmission, and processing delays of our proposal are lower than those of DMME-CMD. A reason why DMME-CMD suffers from a large response delay is that UE contexts are always stored in CMDs in the RO which is distant from UEs.

On the contrary, the increase in average transmission delay in our proposal is mainly caused by UE mobility and the location of VDMMEs and DMDs. With the increase of UE mobility, the appropriateness of candidate servers changes constantly according to dynamic changes of UE location and communication frequency. Accordingly, the appropriate servers become inappropriate soon which cause a large transmission delay in our proposal before new appropriate servers are selected. In addition, for a UE with low mobility, a VDMME and a DMDs at the nearest AP are more likely to become candidates and be selected as new current servers. On the contrary, when a

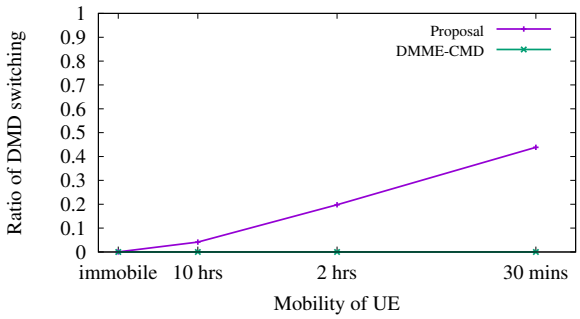

Figure 3.7: Ratio of DMD switching in Scenario 1. Ratio of DMD switching is the rate of the number of DMD switching to the number of requests per time step. It also increases as the UE mobility increases.

UE with high mobility frequently moves from the coverage area of one cell to the area of another, even entering into coverage area of a new LO or TA, a VDMME in the nearest LO and a DMD in the nearest LO or the RO are preferred considering both the total response delay and the C-plane overhead of DMD switching. Long distances between the selected VDMME and the DMD and the UE in this situation also lead to the increase in average transmission delay in our proposal.

Fig. 3.7 shows that the ratio of DMD switching also increases as the UE mobility increases in our proposal, whereas there is no DMD switching in DMME-CMD because of centralized databases. When a UE resides a cell for 30 mins on average and moves to another cell, it incurs DMD switching on about every other movement. It is because, for this group of high mobility, VDMMEs in the nearest LO and DMDs in the nearest LO and the RO are the adaptive selections for UEs, and these servers soon become inappropriate as UEs entering into the covering area of a new LO. In order to reduce the transmission delay, new appropriate servers are selected frequently although we already consider to reduce the C-plane overhead caused by DMD switching. If a DMD in a LO is allocated, movement between cells of its area does not involve DMD switching. The reason for this is that a UE can always associate with the DMD in the nearest LO with a low transmission delay and need not to allocate another DMD for it when the UE moves within the

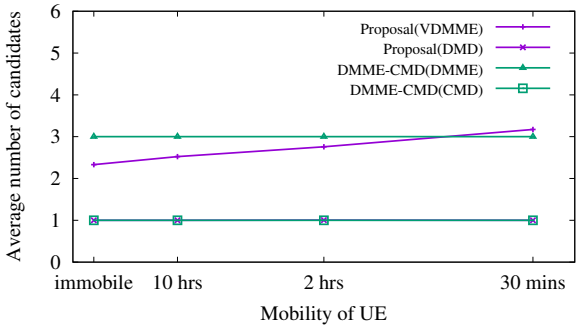

Figure 3.8: Average of candidates in Scenario 1. It is one of indexes of C-plane overhead. The average number of candidate VDMMEs does not change independently of UE mobility. It is as small as that of DMME-CMD.

coverage area of the LO.

In Fig. 3.8, as one of indexes of C-plane overhead, the average number of candidates is shown. Since candidate servers send notifications to a current DMD or an HSS and bandwidth is consumed, the number of candidates should be kept small while leaving room for selection. Figure 3.8 shows that the average number of candidate VDMMEs does not change independently of UE mobility in our proposal and it is as small as that of DMME-CMD. In DMME-CMD, there are always three candidate DMMEs, i.e. a DMME at the nearest AP, a DMME with minimum utilization at an LO, and a DMME with minimum utilization at an RO. On the contrary, in our proposal for a low-mobility UE, a VDMME at the nearest AP is selected as a candidate, while one of VDMMEs in the nearest LO may also become a candidate for serving UEs. It is because, although the probability of VDMMEs in LO is lower than the VDMME at the nearest AP, each LO has more VDMMEs than AP, consequently, one or more VDMMEs in LO will also become candidate accidentally.

For further investigation on C-plane overhead, we additionally show the average number of control messages in Fig. 3.9. It is the average number of all signalling messages between two nodes per UE per procedure for C-plane mobility management including all messages from a UE to the members, from all candidates to the current DMD and HSS, the notifications to new current

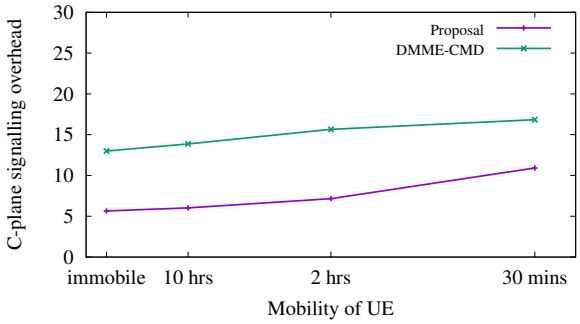

Figure 3.9: C-plane signalling overhead in Scenario 1. It is the average number of all signalling messages between two nodes per UE per procedure for C-plane mobility management including all messages from a UE to the members, from all candidates to the current DMD and HSS, the notifications to new current servers, and signalling executed following the MME procedure.

servers, and signalling executed following the MME procedure. The superiority of our proposal in C-plane overhead becomes more significant independently of UE mobility. It is because our proposal is more flexible than DMME-CMD that we consider the cooperation between VDMMEs and DMDs, however, in DMME-CMD the databases are deployed in a centralized location, i.e. RO. Specifically, for a low mobility UE, the distance from the DMME at the nearest AP to the CMD in RO in DMME-CMD is farther than from the VDMME at the nearest AP to the DMD at the same AP. Moreover, for a UE with high mobility, DMME-CMD select a DMME and a CMD in the RO which is distant to the MA comparing to our proposal that selects one of VDMMEs in the nearest LO and one of DMDs in same LO or the RO.

Therefore, our proposal achieved better performance than DMME-CMD. Additionally, since we extend the probability equation, with which member VDMMEs and DMDs determine whether to become a candidate, by parameter A related to the appropriateness of servers including server utilization, performance of load balancing in our proposal is also better than DMME-CMD. We introduce Jains fairness index to express the goodness of load balancing of servers. Figure 3.10 shows that in our proposal the fairness of VDMMEs maintains a high value, while the fairness of

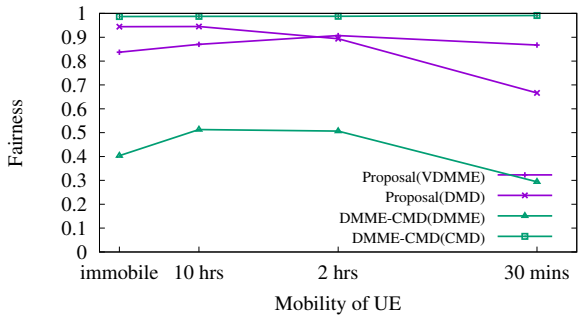

Figure 3.10: Fairness of the VDMME and DMD in Scenario 1. It expresses the goodness of load balancing of servers by Jains fairness index. In our proposal the fairness of VDMMEs maintains a high value.

DMDs decreases with the increase of UE mobility as more contexts are preferred to be stored in DMDs of RO. Although the fairness of CMDs in DMME-CMD is high, all the CMDs are deployed in RO which leads to excessive concentration of load in the RO.

Regarding Scenario 2, results are summarized in Fig. 3.11. For comparison purposes, there are four sets of bars in each figure, each of which is comprised of results of Proposal in Scenario 1 (S1), results of Proposal in Scenario 2 (S2), results of DMME-CMD in S1, and results of DMME-CMD in S2. From the left to the right, each set corresponds to different UE mobility, i.e. immobile, T_s=10 hrs, 2 hrs, and 0.5 hrs, respectively. In the middle figure, since there is no DMD switching for immobile UEs and DMME-CMD, there are only three sets with three bars.

As shown in the top figure, it is obvious that heterogeneous in UE mobility does not affect the average delay in both of our proposal and DMME-CMD as results of S1 and S2 are very similar with each other. In addition, UEs in our proposal receive response messages much faster than DMME-CMD. The superiority comes from distribution of databases and dynamic allocation of appropriate VDMMEs and DMDs, but it incurs overhead in UE context migration as discussed in Scenario 1. A reason why the ratio of DMD switching is smaller in Scenario 2 than Scenario 1 is that appropriate VDMMEs and DMDs are more easily and faster to be selected with a high probability

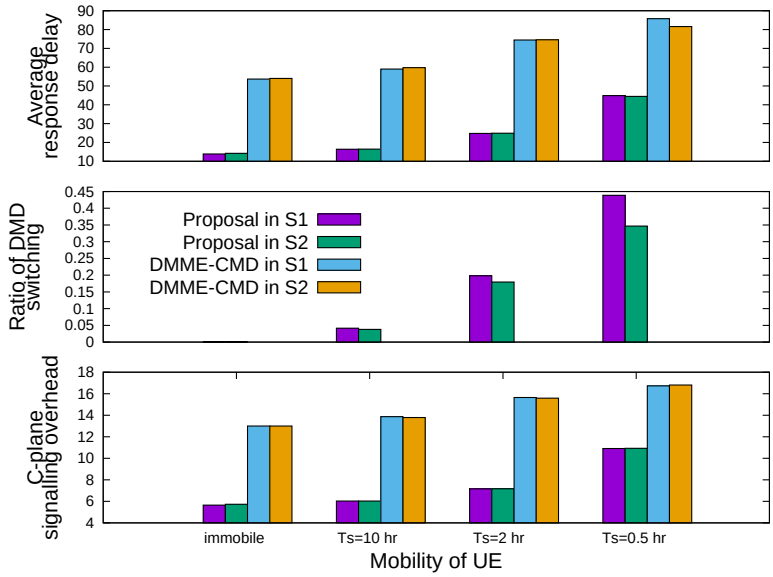

Figure 3.11: Results in Scenario 2 v.s. Scenario 1. Heterogeneous mobility, four mobility levels of equal size (25%): immobile, T_s=0.5, 2, 10 hrs. There are four sets of bars in each figure, each of which is comprised of results of Proposal in Scenario 1 (S1), results of Proposal in Scenario 2 (S2), results of DMME-CMD in S1, and results of DMME-CMD in S2.

in Eq. 3.6, in which parameter A becomes mainly dominated by the normalized response delay. In Scenario 1, because of the homogeneous mobility level, the mobility management tasks prefer to concentrate in some locations considering the response delay and the utilization of appropriate servers is higher than the inappropriate servers so that utilization has different impact on different servers. Conversely, the fairness of all servers in Scenario 2 is higher than Scenario 1, about 0.90, which means the utilization of all servers are very close to each other due to four mobility-levels of UEs in heterogeneous mobility scenario, i.e. utilization has nearly the same influence on probability of different servers.

Despite DMD switching, the C-plane signalling overhead including the signalling for context migrations in our proposal is smaller than that in DMME-CMD as shown in the bottom figure. Therefore, we can conclude that our proposal, by distribution and dynamic allocation of VDMME and DMD, achieves better performance in both of delay and overhead when UEs with different

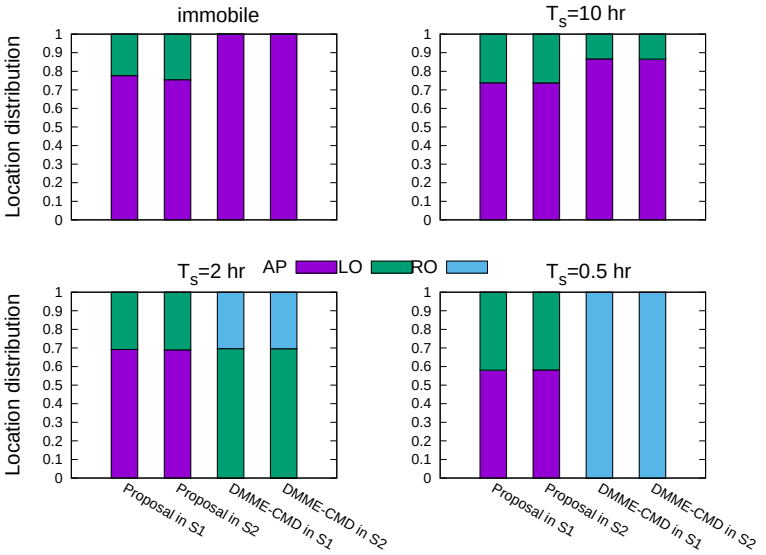

Figure 3.12: VDMME/DMME location distribution in Scenario 2 v.s. Scenario 1. Ratios of locations of current VDMMEs in our proposal and DMMEs in DMME-CMD are shown for both scenarios and different UE mobility.

mobility coexist.

To have deeper insight into server allocation, we show distribution of locations of current servers in Figs. 3.12 and 3.13. In Fig. 3.12, the ratios of locations of current VDMMEs in our proposal and DMMEs in DMME-CMD are shown for both scenarios and different UE mobility. For example, in the top left figure, about 77.6% of immobile UEs are allocated VDMMEs at APs and the remaining UEs are allocated VDMMEs at LOs in our proposal in Scenario 1. As shown in the figure, as UE mobility increases, the ratio of LO increases to reduce the VDMME switching in our proposal. A reason why VDMMEs in RO are not allocated is that for processing the C-plane events VDMMEs often communicate with eNBs and MAs, consequently, VDMMEs deployed in APs and MAs closer to UEs can reduce response delay and mitigate bandwidth consumption. On the contrary, because of the deterministic allocation rule explained in 3.5.2, all immobile UEs are allocated DMMEs at APs and UEs with highest mobility are allocated DMMEs in the RO in DMME-CMD. Although it is natural that there is no difference between scenarios in DMME-CMD, our proposal also performs

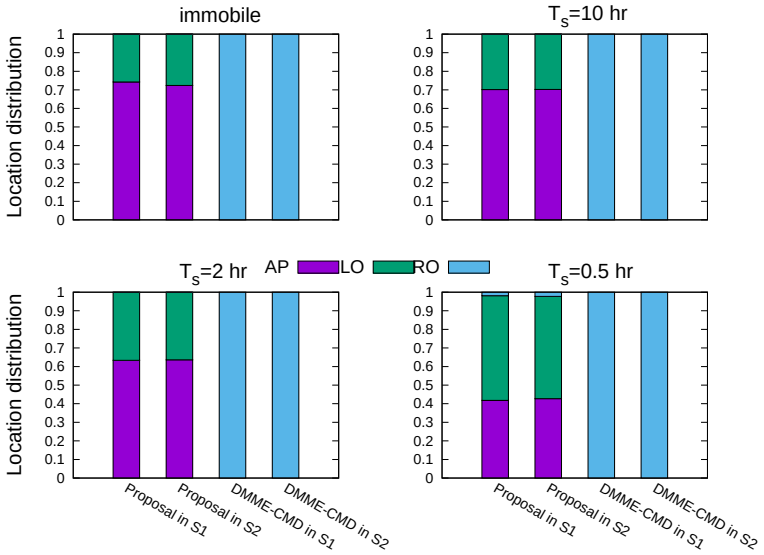

Figure 3.13: DMD/CMD location distribution in Scenario 2 v.s. Scenario 1. Ratios of locations of current DMDs in our proposal and CMDs in DMME-CMD are shown for both scenarios and different UE mobility.

very similar allocation for homogeneous and heterogeneous cases. It is because each VDMME stochastically determines whether to a candidate to serve for a UE with a probability, additionally, deployment of VDMMEs as mentioned in 3.5.1, i.e. one at each AP and 5 VDMMEs per LO, also has an influence on the allocation. Basing on these, although the probability for becoming a candidate changes on a certain extent due to the mobility scenario, the influence on the location distribution is not significant as this probability is only for becoming a candidate not become a new current VDMME.

The similar tendency can be observed in DMD allocation in Fig. 3.13. There is only centralized databases at the RO in DMME-CMD, but our proposal allocated one of distributed databases, namely DMD, according to UE mobility characteristics. In comparison with Fig. 3.12, DMDs at slightly higher location than that of selected VDMMEs for UEs with same mobility level are allocated. The reason for this is that since the context migrations caused by DMD switching consume a lot of bandwidth and cannot be ignored, DMDs at higher locations are preferred to

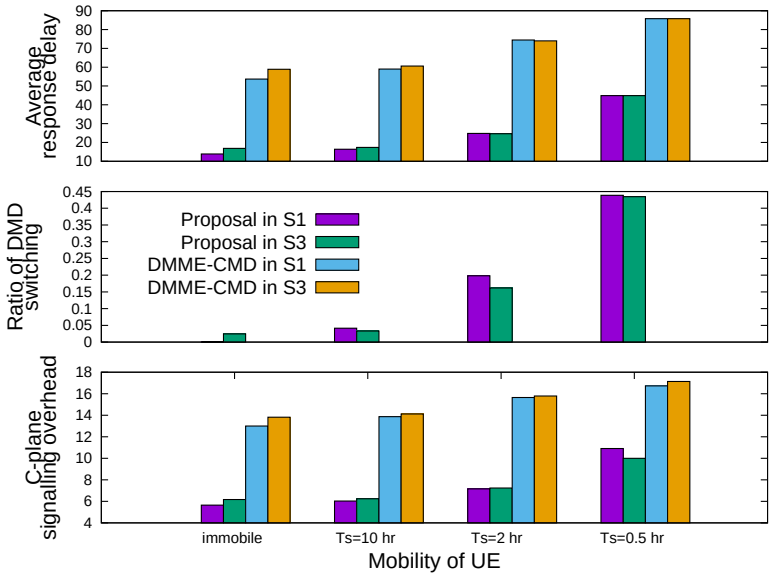

Figure 3.14: Results in Scenario 3 v.s. Scenario 1. Scenario 3: dynamic and heterogeneous mobility, T_s changes every 8 hours among four mobility levels: immobile, T_s=0.5, 2, 10 hrs. Our proposal is superior to DMME-CMD at the sacrifice of UE context migration.

be selected considering the C-plane overhead caused by DMD switching in our proposal. On the contrary, since the release of current VDMME is carried out in parallel and the size of signalling message is small, C-plane overhead for VDMME switching can be ignored so that we don't care for VDMME switching so much.

Finally in Scenario 3, UEs have heterogeneous and dynamic mobility to simulate more practical situations. Results are summarized in Fig. 3.14 for comparison with results of Scenario 1. As can be seen, our proposal is superior to DMME-CMD as shown in Fig. 3.11 at the sacrifice of UE context migration. A reason why there is small difference between Figs. 3.11 and 3.14 is that the number or ratio of UEs with different mobility is not identical in Scenario 3. For example, the slight increase in response delay for immobile UEs in Scenario 3 implies that there are more immobile UEs putting more load on DMMEs at APs. On the contrary, heterogeneous and dynamic mobility of UEs has a larger influence on DMME-CMD. Obviously, the C-plane signaling overhead increases

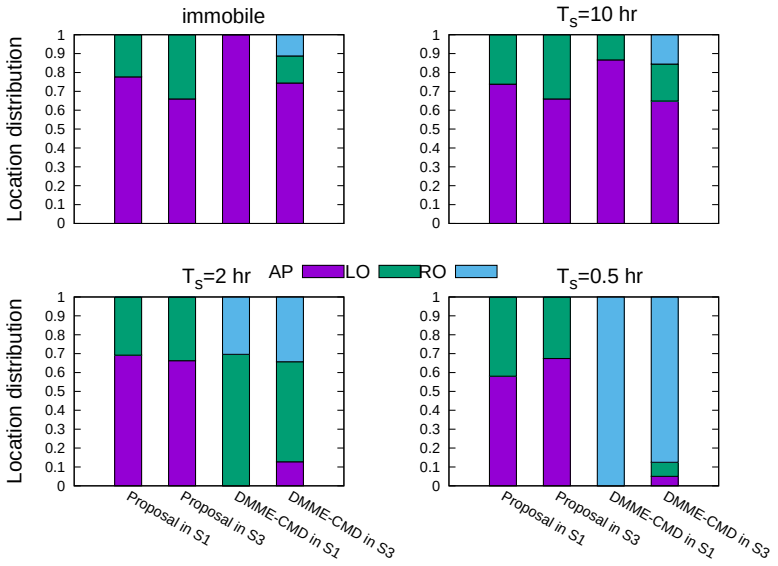

Figure 3.15: VDMME/DMME location distribution in Scenario 3 v.s. Scenario 1. Ratios of locations of current VDMMEs in our proposal and DMMEs in DMME-CMD are shown for both scenarios and different UE mobility.

from Scenario 1 independently of UE mobility with DMME-CMD as shown in the bottom figure of Fig. 3.14. In Figs. 3.15 and 3.16, we summarize the distribution of locations of servers allocated to UEs. By comparing Fig. 3.15 and Fig. 3.12, we notice more irregular DMME allocation in DMME-CMD. For an immobile UE, a DMME at the nearest AP is the optimal from viewpoints of delay and overhead. However, because of dynamic mobility, the instantaneous number of handover, i.e. N_{new}, occasionally larger than one for an immobile UE. Consequently, such a UE will be allocated a DMME at an LO or an RO and thus delay and overhead increases. Since the algorithm takes into account the number of UE migrations between cells as well, our proposal suffers from the dynamic mobility. For example, in Fig. 3.16, VDMMEs at LOs are allocated more to UEs than in Fig. 3.14. However, our proposal still successfully avoids allocating VDMMEs at an RO which leads to larger delay and higher overhead. Speaking of the fairness, the fairness indexes are 0.92 and 0.88 for VDMMEs and DMDs respectively and high enough.

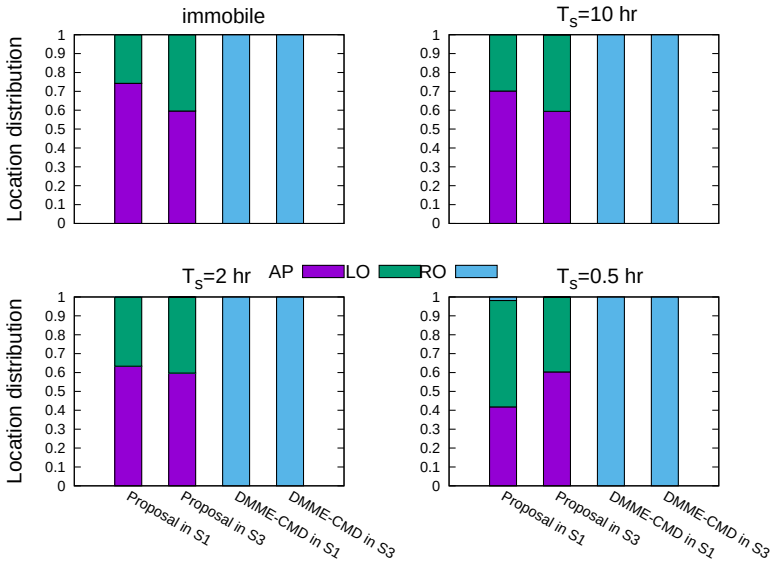

Figure 3.16: DMD/CMD location distribution in Scenario 3 v.s. Scenario 1. Ratios of locations of current DMDs in our proposal and CMDs in DMME-CMD are shown for both scenarios and different UE mobility.

From the above results, we can conclude that our distributed mobility management scheme can mitigate both of delay and overhead in C-plane independently of mobility characteristics of UEs, by autonomous and adaptive allocation of servers based on the response threshold model.

3.6 Summary

In this chapter, after statement of the problems related to the current research of distributed mobility management in mobile network which contributes to tackle the issues of the centralized architecture in current 3.9G LTE/EPC networks, we present a novel flat mobile core network architecture based on the DMM technology and we introduce an autonomous and adaptive VDMME and DMD selection scheme based on the response threshold model. Basically scheme can achieve a high performance, by achieving low latency in the control-plane, high fairness of server load and low C-plane overhead. The system can allocate the tasks of the UE mobility management in C-pane to appropriate VDMMEs and DMDs according to the UEs' characteristic, server load and DMD

switching rate, which reduce much more costs than the existing research of DMME scheme in the flat architecture. However, the number of DMD switching is not small enough. We plan to investigate and improve our algorithm to reduce the C-plane overhead caused by DMD switching. We also need to scale up our evaluation of allocation of VDMMEs and DMDs in a more complex network to verify that our scheme can be applied in the real world.

Chapter 4

Conclusion

After analyzing the emerging challenges and requirements in the next generation of mobile and wireless communication network systems, this thesis states current research and development activities on distributed mobility management and distributed architecture in mobile networks and its related problems. As distribution of mobility management is an effective approach to overcome the issues caused by centralized control in 3.9G LTE/EPC networks, this thesis focused on the distributed mobility management in C-plane in mobile core networks.

First, in Chapter 2, we presented a novel conceptual architecture of autonomous and distributed mobility management in mobile networks based on the current 3.9G LTE/TPC network architecture. For flexible and adaptive distribution of mobility management tasks, we introduce ADMME instead of the LTE/EPC MME, which can be deployed at any nodes in a mobile core network with a help of virtualization technology. Dynamic allocation of appropriate ADMMEs for UEs adaptive to the constantly changing of UE mobility and network status can be accomplished, by a biologically-inspired adaptation algorithm named attractor selection. Through the simulation results in this study, it is verified that the performance of the current mobile network can be improved by taking balance between delay, load status of nodes and C-plane overhead of ADMME relocations, independently of mobility scenarios.

By accomplishing the studies mentioned above, we found that there are notable limitations of the system performance on mitigation of delay and reduction of C-plane overhead. Then our thesis moved to a flat architecture in Chapter 3, which adopts the DMM techniques to anchor the data traffic in U-plane closer to UEs among distributed MAs, for further flexible and adaptive mobility management leaving away from the current standards. For more flexible and dynamic distribution and allocation of C-plane management tasks, splting of the C-plane mobility management tasks of MME similar to DMME scheme is considered. Our main idea is to adopt bio-inspired response threshold model to accomplish autonomous decision-making by individual servers and mitigate response delay and C-plane overhead. We confirmed that our proposal can mitigate delay to the half or even one-fifth and C-plane overhead by more than 41.7% from a partial distributed mobility management scheme proposed in other literature through simulation experiments under three mobility scenarios.

Based on obtained results we confirmed superiority of our proposals to other existing methods, which open up new perspectives on architecture design of a sustainable mobile core network. We can expect that either in the current architecture or in the future flat architecture the distributed mobility management techniques are profitable to accommodate considerable terminal devices with low communication and management cost.

Distributed mobility management is the promising technology of future wireless and mobile communication networks. Our work and study in this thesis are just a first attempt to improve the performance of mobile network systems by reducing the management and communication cost while accommodating considerable number of terminal devices. The number and heterogeneity of user devices are considered to increase and become more diverse. The type of communication and consequently traffic flows would change very much. For example, the amount of peer-to-peer, device-to-device, or terminal-to-terminal type of communication within a mobile core network

would increase. In addition, with proliferation of ICN/CCN and edge computing technology, the substantial proportion of data traffic would stay within a mobile core network. In such a network, distribution and adaptive allocation of mobility anchors and mobility management entities become a more serious problem. Because of autonomy, adaptiveness, and optimality, our bio-inspired algorithms and mechanisms, not necessarily the identical ones, can become base technologies to develop a new mobile core network to satisfy ever-increasing diverse and dynamically changing requirements from a considerably large number of heterogeneous and mobile user terminals.

Bibliography

[1] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Communications*, vol. 21(2), pp. 106–112, 2014.

[2] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52(2), pp. 74–80, May 2014.

[3] A. Daj and C. Samoil, "Digital marketing and regulatory challenges of Machine-to-Machine (M2M) communications," in *Proceedings of 2012 9th International Conference on Remote Engineering and Virtual Instrumentation (REV)*, July 2012.

[4] "LTE; general packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access," TS 23.401 V10.5.0, 3GPP, Oct. 2011.

[5] A. Osseiran, F. Boccardi, V. Braun, and K. Kusume, "Scenarios for the 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52(5), pp. 26–35, May 2014.

[6] Y. Park, "5G vision and requirements of 5G forum, korea," tech. rep., ITU, Feb. 2014.

[7] H. Chan, D. Liu, P. Seite, H. Yokota, and J. Korhonen, "RFC 7333: Requirements for distributed mobility management," Aug. 2014.

[8] F. Giust, A. D. la Oliva, and C. Bernardos, "Mobility management in next generation mobile networks," in *Proceedings of 2013 IEEE 14th International Symposium on A World of Wireless, Mobile and Multimedia Networks(WoWMoM)*, pp. 1–3, June 2013.

[9] X. An and F. Pianese, "DMME: A distributed LTE mobility management entity," *Bell Labs Technical Journal*, vol. 17(2), pp. 97–120, Feb. 2012.

[10] A. Basta and W. Kellerer, "Applying NFV and SDN to LTE mobile core gateways, the functions placement problem," in *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, & Challenges*, (Chicago, USA), pp. 33–38, Aug. 2014.

[11] S. Kraus and T. Plotkin, "Algorithms of distributed task allocation for cooperative agents," *ElSEVIER Theoretical Computer Science*, vol. 242(1), pp. 1–27, 2000.

[12] A. Kashiwagi and I. Urabe, "Adaptive response of a gene network to environmental changes by fitness-induced attractor selection," *PLoS ONE*, vol. 1(1):e49, pp. 1–10, Dec. 2006.

[13] E. Bonabeau, A. Sobkowski, G. Theraulaz, and J. L. Deneubourg, "Adaptive task allocation inspired by a model of division of labor in social insects," in *Proceedings of Biocomputing and Emergent Computation*, pp. 36–45, 1997.

[14] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019." White Paper, Feb. 2015.

[15] D. Liu, J. Zuniga, P. Seite, H. Chan, and C. Bernardos, "RFC 7429: Distributed mobility management: Current practices and gap analysis," Jan. 2015.

[16] C. Bernardos, A. de la Oliva, and F. Giust, "A PMIPv6-based solution for distributed mobility management," Mar. 2015.

[17] J. Lei and X. Fu, "Evaluating the benefits of introducing PMIPv6 for localized mobility management," in *Proceedings of 2008 International Wireless Communications and Mobile Computing Conference (IWCMC)*, (Crete Island, Greece), pp. 74–80, Aug. 2008.

[18] J. Laganier, T.Higuchi, and K. Nishida, "Mobility management for all-IP core network," *NTT Docomo Technical Journal*, vol. 11(3), pp. 34–39, Dec. 2009.

[19] J. Lee and S. Pack, "Mobility management in future wireless networks: Past, present, and future." IEEE WCNC Tutorial, Apr. 2013.

[20] F. Guist, C. Bernarods, and A. Oliva, "Analytic evaluation and experimental validation of a network-based IPv6 distributed mobility management solution," *IEEE Transactions on Mobile Computing*, vol. 13(11), pp. 2484–2497, Nov. 2014.

[21] R. Costa and T. Melia, "When mobile networks meet content delivery networks: challenges and opportunities," in *Proceedings of the seventh ACM international workshop on Mobility in the evolving internet architecture*, (Istanbul, Turkey), pp. 11–16, 2012.

[22] H. A. Chan, H. Yokota, and J. Xie, "Distributed and dynamic mobility management in mobile internet: Current approaches and issues," *Journal of Communications*, vol. 6(1), pp. 4–15, Feb. 2011.

[23] J. L. Z. Yan, "Dynamic tunneling for network-based distributed mobility management coexisting with PMIPv6," in *Proceedings of IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, (London, United Kingdom), pp. 2995–3000, Sept. 2013.

[24] T. Condeixa and S. Sargento, "Dynamic mobile ip anchoring," in *Proceedings of 2013 IEEE International Conference on Communications (ICC)*, (Budapest, Hungary), pp. 3607–3612, June 2013.

[25] 3GPP, "General packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access," Sept. 2014.

[26] 3GPP, "3rd generation partnership project; technical specification group core network and terminals; numbering, addressing and identification (release 9)," June 2010.

[27] 3GPP, "3rd generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access network (E-UTRAN); architecture description (release 10)," Sept. 2011.

[28] 3GPP, "General packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access," Tech. Rep. TS 23.401 V13.3.0, June 2015.

[29] K. Leibnitz and M. Murata, "Attractor selection and perturbation for robust networks in fluctuating environments," *IEEE Network*, vol. 24(3), pp. 14–18, May 2010.

[30] J. Li, H. Kameda, and K. Li, "Optimal dynamic mobility management for PCS networks," *IEEE/ACM Transactions on Networking*, vol. 8(3), pp. 319–327, June 2000.

[31] R. Langar, N. Bouabdallah, and R. Boutaba, "A comprehensive analysis of mobility management in MPLS-based wireless access networks," *IEEE/ACM Transactions on Networking*, vol. 16(4), pp. 918–931, Aug. 2008.

[32] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *ACM Computing Research Repository*, vol. cs.NI/9809, 1998.

[33] "3GPP; service requirements for Machine-Type Communications (MTC); stage 1," TS 22.368 V12.3.0, 3GPP, Dec. 2013.

[34] H. Yang, N. Wakamiya, M. Murata, T. Iwai, and S. Yamano, "Autonomous and distributed mobility management scheme in mobile core networks," *Springer Wireless Networks*, pp. 1–18, Apr. 2016.

[35] A. Zanella, "Queueing theory," in *Principles of Communications Networks and Systems* (N. Benvenuto and M. Zorzi, eds.), John Wiley & Sons, Ltd, Chichester, UK, Sept. 2011.

[36] "3GPP; technical specification group radio access network; feasibility study for evolved universal terrestrial radio access (UTRA) and universal terrestrial radio access network (UTRAN)," TR 25.912 V13.0.0, 3GPP, Dec. 2015.

[37] "ASR5x00 MME overload protection features," Tech. Rep. 119002, CISCO, June 2015.