

Ph. D. Dissertation

Adaptive MAC Protocols for Massive Devices  
in Cellular Network

Sung-Hyung Lee

Department of Electrical and Computer Engineering

Graduate School

Ajou University

# Adaptive MAC Protocols for Massive Devices in Cellular Network

by

Sung-Hyung Lee

Department of Electrical and Computer Engineering

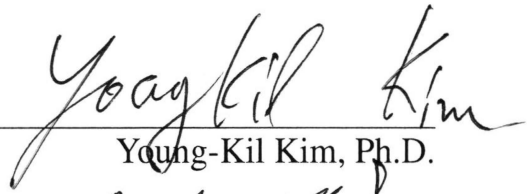
Ajou University

Advisor : Professor Jae-Hyun Kim

A dissertation submitted to the faculty of Ajou University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering. The study was conducted in accordance with Code of Research Ethics.

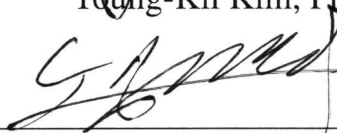
August, 2017.

The doctoral dissertation of Sung-Hyung Lee  
is hereby approved.



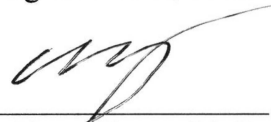
---

Young-Kil Kim, Ph.D.



---

Byeong-hee Roh, Ph.D.



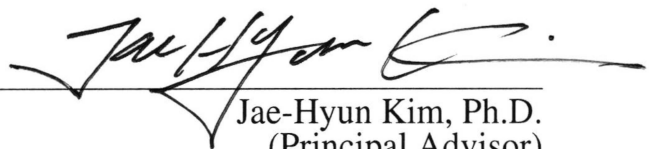
---

Song-Nam Hong, Ph.D.



---

Sunghyun Cho, Ph.D.



---

Jae-Hyun Kim, Ph.D.  
(Principal Advisor)

Graduate School  
Ajou University  
June 7, 2017.

## Acknowledgements

우선, 이 학위논문의 완성까지 저를 지도해주신 김재현 교수님께 감사드립니다. 아무것도 모르고 미숙하게 시작하였던 저를 믿고 지도 해주셔서 제가 학위논문의 완성까지 도달할 수 있었습니다. 교수님께서 가르쳐 주신 철학 및 사회생활에 필요한 덕목을 따라 부끄럽지 않은 제자가 되도록 노력하겠습니다.

부족한 학위 논문을 살펴봐주시고 조언해주신 김영길 교수님, 노병희 교수님, 홍송남 교수님, 조성현 교수님께도 감사의 말씀을 드립니다. 바쁘신 와중에도 논문을 검토해주시고 의견을 제시해 주셔서 논문의 완성도를 높일 수 있었습니다.

무선인터넷연구실에서 함께 지냈던 동료들에게도 감사의 글을 남깁니다. 저에게 열심히 공부하는 모습을 보여주셔서 제가 따라갈 수 있게 해주신 선배 분들인 차재룡, 오성민, 이현진, 이성진, 최호승 형님 및 이주아 누님께 감사드립니다. 열심히 노력하셔서 목표를 이룬 모습을 보여주셔서 힘이 되주었던 오승환, 강신현, 이충희, 추상민, 김지수 형님들께도 감사드립니다. 그리고 오랜 시간동안 연구실에서 함께 했고 학위논문 준비에 조언을 주었던 동기 규환이에게 감사합니다. 저와 함께 연구실 생활을 하다 먼저 목적을 이루고 나간 안두성 형님, 고

광춘, 김신구, 양희인, Nathnael Gebregziabher Weldegiorgis, 정현기, 유승수, 김원경, 최동열, 김종무, Moonmoon Mohanty 에게는 함께 있을 수 있어서 즐거웠다는 말을 남깁니다. 학위 논문 완성 순간을 같이한 강석원 형님, 천혜림 누님, 김진기, 정소이, 김경록, 조준우, 이동학, 오지훈, 이동구, 고준영에게는 마지막까지 도와주고 격려해주어서 감사다는 말을 전합니다.

마지막으로 오랫동안 기다려주시고 격려해주신 아버지, 어머니께 진심으로 감사합니다. 그리고 물심양면으로 도와준 동생 신형이에게도 고맙습니다. 격려해주시고 응원해주었던 그 마음을 이제 갚을 수 있도록 노력하겠습니다.

이 논문이 만들어지기까지 응원해주었던 많은 분들께 이 논문을 바칩니다.

**2017년 6월 28일**

**이성형 올림**

# Abstract

The main theme of this dissertation is the design of adaptive medium access protocol to alleviate the traffic overload and to provide resource efficiency for the massive number of devices. This dissertation focuses on the development of mechanisms to cope with the traffic overload due to the time-variant arrival of devices, and the mismatch between the estimated number of devices and the actual number of devices in cellular networks. This dissertation also focuses on efficient random access (RA) procedure for Internet of things (IoT) devices.

The dynamic allocation of RACH resources (DARR) is one of solutions to alleviate the traffic overload from massive devices when the resource for RA can be increased. This dissertation discusses the challenge of a gap between the theoretical maximum throughput and the actual throughput in DARR. The gap occurs when the BS cannot change the number of preambles for a RACH until multiple numbers of RACHs are completed. In addition, a preamble partition (PP) approach is proposed that uses two groups of preambles to reduce this

gap. The simulation results show that the proposed approach can achieve the throughput which is closer to theoretical maximum throughput than other approaches.

The resources for the machine-to-machine (M2M) communication devices can be limited and shared with other type of devices. In this case, both the DARR and the access class barring (ACB) are required to provide resource efficiency and to alleviate the traffic overload. This dissertation discusses the gap between the theoretical maximum throughput and the actual throughput since the BS cannot change the number of preambles for a RACH until multiple number of RACHs are completed. Based on the discussion, a preamble partition and a stochastic gradient descent approach are proposed to reduce the gap when both of DARR and ACB are used for mobile network system. The simulation results show that the proposed protocol shows the throughput which is close to the throughput with ideal selection.

The data transmission using a RA procedure for human-to-human (H2H) communications requires the overhead for signaling messages. The overhead for the signaling is very high compared with the size of data in small data transmission (SDT). Thus, the recent release of LTE-A standard includes two SDT procedures that reduce the number of exchanges. This dissertation evaluates the performance of conventional

SDT procedures in the viewpoint of resource usage and resource throughput. In addition, this dissertation also proposes an SDT procedure to reduce the resource usage and to increase the resource efficiency. The numerical evaluation shows that the proposed approach decreases the resource usage and increases the resource efficiency.



# Contents

|  |            |
|--|------------|
| <b>List of Figures</b>   | <b>vii</b> |
| <b>List of Tables</b>  | <b>x</b>   |
| <b>1 Introduction</b>  | <b>1</b>   |
| 1.1 Background and motivation . . . . .                                    | 1          |
| 1.2 Problem definition . . . . .   | 4          |
| 1.3 Contribution . . . . .   | 6          |
| 1.4 Overview . . . . .   | 8          |
| <b>2 Background and Related Works</b>                                      | <b>10</b>  |
| 2.1 Data transmission procedure using RACH procedure in<br>LTE-A . . . . . | 10         |
| 2.2 Access class barring and its related works . . . . .                   | 18         |
| 2.3 Dynamic allocation of RACH resource . . . . .                          | 20         |
| <b>3 Preamble Partition based Adaptive DARR Protocol</b>                   | <b>30</b>  |
| 3.1 Introduction . . . . .   | 30         |
| 3.2 System Model . . . . .   | 31         |

|          |  |           |
|----------|--|-----------|
| 3.3      | Problem definition : Decrease of throughput in DARR due to resource allocation update interval . . . . . | 36        |
| 3.4      | The Proposed Preamble Partition Protocol . . . . .   | 43        |
| 3.5      | Performance Evaluation . . . . .   | 50        |
| 3.6      | Summary . . . . .  | 59        |
| <b>4</b> | <b>Preamble Partition based Adaptive DARR and ACB Protocol</b>   | <b>60</b> |
| 4.1      | Introduction . . . . .   | 60        |
| 4.2      | System Model . . . . .   | 61        |
| 4.3      | Problem Definition: Needs of adaptive MAC protocol considering si-periodicity . . . . .                  | 65        |
| 4.3.1    | Background for the adaptive MAC protocol with DARR and ACB . . . . .                                     | 66        |
| 4.3.2    | Throughput degradation in DARR due to the resource allocation update interval . . . . .                  | 67        |
| 4.3.3    | Throughput of ACB with ACB factor update interval  | 68        |
| 4.4      | Proposed dynamic resource allocation and congestion control protocol . . . . .                           | 69        |
| 4.4.1    | Background for device grouping . . . . .   | 69        |
| 4.4.2    | Background for the optimization with stochastic process . . . . .  | 70        |
| 4.4.3    | Preamble partition based adaptive resource allocation and congestion control protocol . . . . .          | 73        |
| 4.5      | Performance evaluation . . . . .   | 77        |

|          |  |            |
|----------|--|------------|
| 4.6      | Summary . . . . .  | 89         |
| <b>5</b> | <b>Optimizing Random Access Procedure for the Data Transmission of MTC Devices</b> | <b>90</b>  |
| 5.1      | Introduction . . . . .   | 90         |
| 5.2      | System Model . . . . .   | 92         |
| 5.3      | Conventional procedures for the data transmission using RA                         | 93         |
| 5.3.1    | Control plane (CP) solution . . . . .  | 93         |
| 5.3.2    | User plane (UP) solution . . . . .   | 97         |
| 5.3.3    | Data in MSG1 . . . . .   | 99         |
| 5.3.4    | Data in MSG3 . . . . .   | 100        |
| 5.4      | Numerical evaluation for the conventional SDT procedures                           | 102        |
| 5.4.1    | CP solution . . . . .  | 104        |
| 5.4.2    | UP solution . . . . .  | 108        |
| 5.4.3    | Data in MSG1 . . . . .   | 110        |
| 5.4.4    | Data in MSG3 . . . . .   | 111        |
| 5.5      | Proposed short data transmission procedure for IoT devices                         | 114        |
| 5.6      | Numerical evaluation for the proposed SDT procedure . .                            | 116        |
| 5.7      | Performance Evaluation . . . . .   | 118        |
| 5.8      | Summary . . . . .  | 124        |
| <b>6</b> | <b>Conclusion</b>  | <b>125</b> |
|          | <b>References</b>  | <b>128</b> |

# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | Network architecture for massive machine type communications . . . . .  | 3  |
| 1.2 | The objectives of this dissertation . . . . .   | 6  |
| 2.1 | Data transmission using RACH procedure in conventional LTE-A . . . . .  | 12 |
| 2.2 | RACH and SIB2 transmission in frame structure of LTE-A  | 16 |
| 2.3 | RACH and SIB2 transmission in frame structure of LTE-A  | 17 |
| 2.4 | Frame structure of FSA . . . . .  | 25 |
| 3.1 | $M_i$ , $R_r$ , and optimum for $R_r$ with $M$ of 100,000 . . . . .   | 41 |
| 3.2 | $M_i$ , $R_r$ , and optimum for $R_r$ with $M$ of 30,000 . . . . .  | 41 |
| 3.3 | $M_i$ , $R_r$ , and optimum for $R_r$ with $M$ of 50,000 . . . . .  | 42 |
| 3.4 | Number of contending devices ( $M_i$ ), number of non-deep backlogged devices [ $B_N(i)$ ], and optimum value for $R_r$ ( $R^*$ ): (a) $M = 50,000$ and (b) $M = 100,000$ . . . . . | 45 |
| 3.5 | The throughput of the cell with proposed preamble partition approach for $M$ of 100,000 . . . . .   | 52 |
| 3.6 | $M_i$ , $R_{m,r}$ , and $R_r$ for $M$ of 100,000 . . . . .  | 52 |

|      |   |    |
|------|---|----|
| 3.7  | The average throughput vs. the number of arrivals per time slot for uniformly distributed arrival . . . . .                                       | 54 |
| 3.8  | The average throughput vs. the number of arrivals per time slot for Beta distributed arrival . . . . .  | 54 |
| 3.9  | Success ratio vs. arrival rate per RACH period for uniformly distributed arrival . . . . .  | 56 |
| 3.10 | Success ratio vs. arrival rate per RACH period for Beta distributed arrival . . . . .   | 56 |
| 3.11 | Average number of preamble transmissions for success vs. arrival rate per RACH period for uniformly distributed arrival . . . . .                 | 58 |
| 3.12 | Average number of preamble transmissions for success vs. arrival rate per RACH period for Beta distributed arrival                                | 58 |
| 4.1  | Average throughput vs. number of devices for the si-period of 32 . . . . .  | 81 |
| 4.2  | (a) Average number of successful preambles per RACH<br>(b) Average number of allocated preambles per RACH. Si-periodicity is equal to 32. . . . . | 83 |
| 4.3  | CDF of throughputs for si-period of 32 with (a) 10,000 devices, (b) 30,000 devices, (c) 50,000 devices . . . . .                                  | 84 |
| 4.4  | Average delay vs. number of devices for si-period of 32 .   | 85 |
| 4.5  | Average delay vs. number of devices for si-period of 32, magnified for delay less than 0.5 s . . . . .  | 85 |
| 4.6  | Average throughput vs. si-periodicity for 10,000 devices .  | 86 |

|      |  |     |
|------|--|-----|
| 4.7  | Average throughput vs. si-periodicity for 20,000 devices .   | 87  |
| 4.8  | Average throughput vs. si-periodicity for 30,000 devices .   | 87  |
| 4.9  | CDF of utilities for different si-periodicity with $M$ of 30,000<br>and si-period of (a) 8 time slots, (b) 16 time slots, (c) 32<br>time slots . . . . . | 88  |
| 5.1  | Initial attach procedure . . . . .   | 94  |
| 5.2  | Data transmission procedure of CP solution . . . . .   | 96  |
| 5.3  | Data transmission procedure of CP solution . . . . .   | 98  |
| 5.4  | RA procedure with the data in MSG1 . . . . .   | 101 |
| 5.5  | RA procedure with the data in MSG3 . . . . .   | 103 |
| 5.6  | The proposed SDT procedure . . . . .   | 117 |
| 5.7  | Average number of allocated resources in uplink vs. $M$ for<br>BPSK . . . . .  | 121 |
| 5.8  | Uplink resource efficiency vs. $M$ for BPSK . . . . .  | 122 |
| 5.9  | Average number of allocated resources in uplink vs. $M$ for<br>QPSK . . . . .  | 123 |
| 5.10 | Uplink resource efficiency vs. $M$ for QPSK . . . . .  | 123 |

# List of Tables

|     |  |     |
|-----|--|-----|
| 2.1 | Summary of ACB and related schemes (1) . . . . .                               | 21  |
| 2.2 | Summary of ACB and related schemes (2) . . . . .                               | 22  |
| 2.3 | Summary of DFSA and ADP (1) . . . . .  | 28  |
| 2.4 | Summary of DFSA and ADP (2) . . . . .  | 29  |
| 3.1 | Summary for conventional DARR policies . . . . .                               | 38  |
| 3.2 | Parameters for performance evaluation of preamble partition protocol . . . . . | 51  |
| 4.1 | Parameters for performance evaluation of preamble partition approach . . . . . | 79  |
| 5.1 | Parameters for evaluation of data in MSG5 . . . . .                            | 120 |
| 5.2 | Assumptions for the size of data, $L_D$ . . . . .                              | 120 |

## List of Abbreviations

|               |                                       |
|---------------|---------------------------------------|
| <b>3GPP</b>   | 3rd generation partnership project    |
| <b>ACB</b>    | access class barring                  |
| <b>ACK</b>    | acknowledgment                        |
| <b>ADP</b>    | adaptive determination of the pool    |
| <b>BO</b>     | backoff                               |
| <b>BPSK</b>   | binary phase shift keying             |
| <b>BS</b>     | base station                          |
| <b>CoAP</b>   | constrained application protocol      |
| <b>CP</b>     | control plane                         |
| <b>DARR</b>   | dynamic allocation for RACH resources |
| <b>DCI</b>    | downlink control information          |
| <b>DFSA</b>   | dynamic frame slotted ALOHA           |
| <b>DRA</b>    | dynamic resource allocation           |
| <b>DTLS</b>   | datagram transport layer security     |
| <b>EAB</b>    | extended access barring               |
| <b>eNodeB</b> | enhanced node b                       |
| <b>FSA</b>    | frame slotted ALOHA                   |
| <b>H2H</b>    | human-to-human                        |



|               |  |
|---------------|--|
| <b>HARQ</b>   | hybrid automatic repeat request                            |
| <b>IoT</b>    | Internet of things   |
| <b>IP</b>     | Internet protocol  |
| <b>LTE-A</b>  | long term evolution-advanced                               |
| <b>M2M</b>    | machine-to-machine   |
| <b>MAC</b>    | medium access control                                      |
| <b>MCSA</b>   | multi-channel slotted aloha                                |
| <b>MSG1</b>   | preamble   |
| <b>MSG2</b>   | random access response                                     |
| <b>MSG3</b>   | RRC connection request                                     |
| <b>MSG4</b>   | contention resolution                                      |
| <b>MSG5</b>   | RRC connection setup                                       |
| <b>MSG6</b>   | RRC connection setup complete                              |
| <b>MTC</b>    | machine type communication                                 |
| <b>NB-IoT</b> | narrow band Internet of things                             |
| <b>PDCCH</b>  | physical downlink control channel                          |
| <b>PDSCH</b>  | physical downlink shared channel                           |
| <b>PRADA</b>  | prioritized random access with dynamic access bar-<br>ring |
| <b>PUCCH</b>  | physical uplink control channel                            |
| <b>PUSCH</b>  | physical uplink shared channel                             |
| <b>QAM</b>    | quadrature amplitude modulation                            |
| <b>QPSK</b>   | quadrature phase shift keying                              |
| <b>RA</b>     | random access  |
| <b>RACH</b>   | random access channel                                      |

|                |   |
|----------------|---|
| <b>RAN</b>     | radio access network                              |
| <b>RAR</b>     | random access response                            |
| <b>RB</b>      | resource block                                    |
| <b>RFID</b>    | radio frequency identification                    |
| <b>RRC</b>     | radio resource control                            |
| <b>SDT</b>     | small data transmission                           |
| <b>SIB</b>     | system information block                          |
| <b>TA</b>      | timing alignment                                  |
| <b>TC-RNTI</b> | Temporary cell radio network temporary identifier |
| <b>UDP</b>     | user datagram protocol                            |
| <b>UE</b>      | user equipment                                    |
| <b>UL</b>      | uplink  |
| <b>UP</b>      | user plane  |

# Chapter 1. Introduction

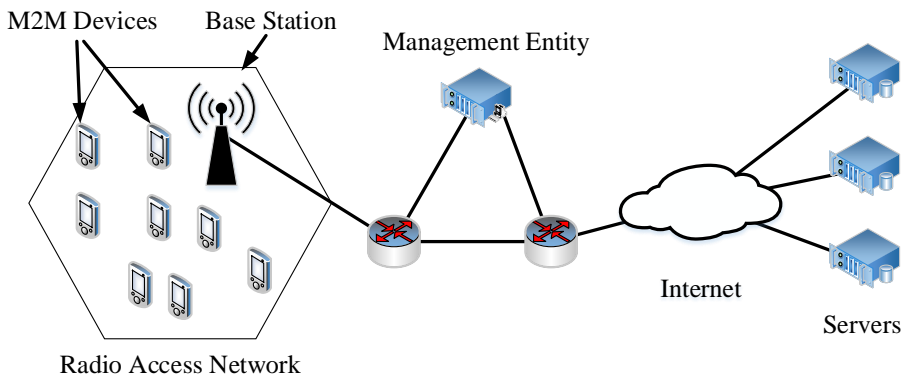
## 1.1. Background and motivation

The Internet of things (IoT) is the inter-networking of devices, where devices can be vehicles, buildings, sensors, or other objects. For the IoT devices, machine-to-machine (M2M) or machine type communication (MTC) technologies have been researched. These technologies enable the IoT devices to distribute their data or to receive any information from other devices. The transmission and reception of data through the network can enable the inter-action between devices, which increases the utility of devices. Thus, M2M or MTC technologies have a wide range of applications, including smart grids, personal communications, controlling traffic flow on road, smart driving, and smart healthcare [1–5]. With the development of cheap devices including communication capability, the number of IoT devices is expected to increase due to their wide range of application area. A report expected that there will be 12.5 billion IoT devices in the world by 2020 [6]. The international telecommunication union (ITU) expects the connection density with 1,000,000 devices per  $km^2$  [7] for international mobile telecommunications-2020 (IMT-2020).

The third generation partnership project (3GPP) has studied the conventional LTE-A system with the 30,000 of devices within a cell [8] and the number of devices in a cell can be increased to more than 100,000 [9].

Figure 1.1 shows the network architecture for massive machine type communications [10]. As represented in Figure 1.1, the major applications with IoT devices require the communication between the devices and the servers, where the devices and the servers are placed in different network domain [4]. The increase of devices can make the traffic overload at any link between the devices and the servers. The traffic overload of IoT devices is one of the challenges for radio access network (RAN) because too many devices can attempt to access a base station (BS) simultaneously in a short period of time. In the cellular network, an IoT device performs a random access (RA) procedure for connection through a BS to other network domain [11]. In addition, the devices will be entered in an idle state or a temporarily disconnected state to reduce their battery usage after the connection since the data can be generated with long interval. In this case, these devices should transmit their data using RA procedure, since they lost their synchronization with the BS. Therefore, the traffic overload at the radio access network is expected in random access channel (RACH) of cellular networks.

The time and frequency resources are not dedicated for the IoT devices in mobile networks. The communication services for IoT devices can be provided by the bandwidth which is shared with human-to-human (H2H) communications. In addition, the IoT devices generally has low



**Fig. 1.1.** Network architecture for massive machine type communications

priority compared with that of H2H communications. Due to these characteristics of IoT services, the BS should allocate the resources for IoT devices without the waste of bandwidth. Therefore, the efficient uses of bandwidth is another challenges in the cellular networks [12].

In the recent days, the most commonly used cellular network system is long term evolution-advanced (LTE-A) system. A device in the LTE-A network requires a RA procedure, called by RACH procedure. In RACH procedure, the device randomly chooses and transmits a preamble from a pool of preambles, or “pool” in this dissertation, using the specific sub-carriers allocated during a special subframe. The set of sub-carriers is referred as RACH. The BS can obtain a request if only one device has selected this preamble in a RA slot. Thus, RACH procedure is similar to the frame slotted ALOHA (FSA) or the multi-channel ALOHA, where the BS can estimate the number of MTC devices that send preambles in a RA slot for congestion control [12].

The random access in cellular network with massive devices has several problems when a large number of M2M devices try to access [4, 13]. The cellular networks are initially designed for human-to-human (H2H) communications. H2H communications generally require small number of devices but high data rate. Thus, the H2H devices require the small resource for the RA and the large resource for the data transmission. However, the M2M communications generally require the large number of devices with low data rates. The interval between two data generated from a device can be few hours to more than a day [14]. Thus, M2M communications will require large resource for the RA. The difference between H2H and IoT communications will cause several problems in M2M communications. The first problem is the low efficiency. Cellular network requires large overhead for random access regardless of the size of data. The second problem is the congestion because of the large number of devices, the insufficient number of resources for random access, and the insufficient bandwidth.

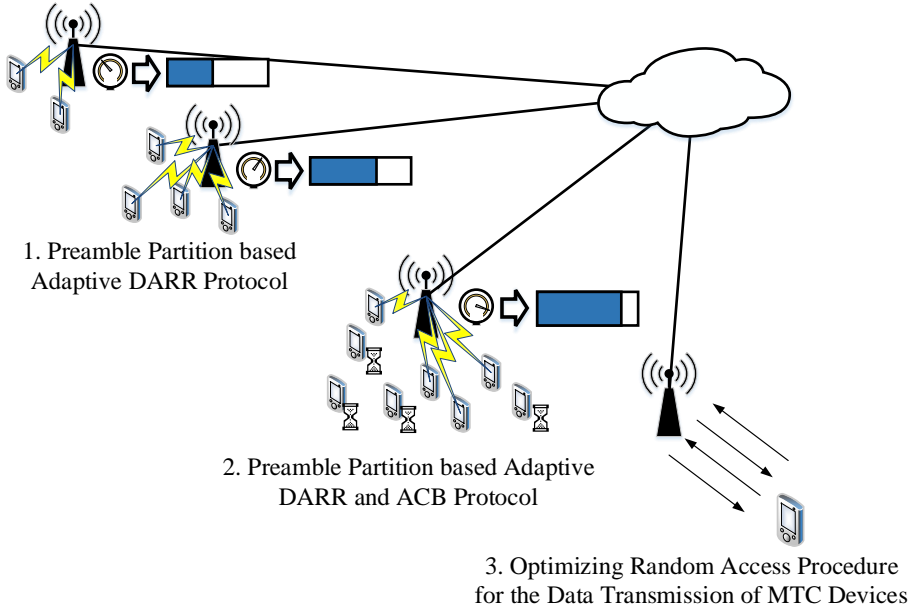
## **1.2. Problem definition**

This dissertation focuses to alleviate the RAN congestion in cellular networks. In case of an emergency, the data from all devices needs to be collected as soon as possible [4]. Thus, the communication system requires to minimize the total amount of the time to gather data packets from all activated devices. The dynamic allocation for RACH resource (DARR) scheme is one of many solutions to minimize the time, where

the DARR adaptively changes the size of pool and other resources [8, 15]. The problems in DARR are similar to those in the studies in dynamic frame slotted aloha (DFSA). However, the schemes in DFSA cannot be applied directly to the LTE-A although they are similar. The DFSA assumes that the size of pool can be updated in every RA slot. However, the parameters related to RACH procedure can be updated with periodicity in cellular network, where the multiple RA slots are allocated during the period. The BS cannot directly apply the schemes in DFSA due to the periodicity, since the studies for DFSA achieve their objectives without the periodicity.

The cellular networks generally have the limits on bandwidth. The access class barring (ACB) is one of solutions to control traffic overload when the resources for the random access is limited. Since the resources for IoT devices should not be wasted although the traffic load of IoT devices is low, the combination of DARR and ACB is required to achieve both resource efficiency and congestion control. Thus, the adaptive MAC protocol for massive devices in cellular network should consider both DARR and ACB to efficiently use the radio resources and to control traffic overload.

The IoT devices transmit the packet with very small size of application payload, where their inter-arrival time can be very long. In this case, the traffic will be generated when a device is in the idle state. The basic data transmission procedure in LTE-A requires initial attach, authentication, and association procedure for every data transmission. To reduce the overhead, the 3GPP and other researchers studied and proposed their



**Fig. 1.2.** The objectives of this dissertation

small data transmission (SDT) procedures. The cellular network cannot allocate resources with the unit of bits, but with the unit of a bunch of bits. For example, the BS in LTE-A can allocate time-frequency resources in the unit of resource blocks (RBs). Thus, the resource usage and resource efficiency need to be evaluated considering the unit for resource allocation. In addition, the efficient SDT procedure is required to reduce the resource usage to improve resource efficiency.

### 1.3. Contribution

The goal of this dissertation is the design of adaptive medium access protocol to alleviate the traffic overload and to provide resource efficiency for



the massive number of devices in cellular network system. As mentioned above, the adaptive medium access protocol needs to consider an interval to change the size of pool and/or the ACB factor. In addition, the adaptive medium access protocol requires a small data procedure which efficiently uses the time and bandwidth resources. Figure 1.2 shows the objectives and the contributions of this dissertation, where the details are as follows:

- A preamble partition based adaptive DARR protocol is presented. The problem of the throughput degradation of DARR in LTE-A is discussed, which is due to an interval to update the amount of resources used for RA procedure. To mitigate the throughput degradation in DARR, a DARR protocol is proposed which separates a resource pool into two pools to select the size of pool. The proposed protocol is evaluated by a simulation. Simulation results show that the proposed protocol can achieve performance that is close to the optimal throughput of FSA than the throughput without the proposed approach.
- A preamble partition and stochastic gradient descent based DARR and ACB protocol is presented. The throughput with an interval to update the pool size and the ACB factor is discussed. To increase the throughput for DARR and ACB, the preamble partition approach is proposed to obtain better information to select the amount of resource. In addition, the stochastic gradient descent method is proposed to find optimal amount of resource. The

proposed approach also adaptively separates or merges the preamble pools to provide optimized throughput considering both DARR and ACB. Simulation results show that our proposed approach can achieve performance that is close to the performance when BS can always select best parameters for DARR and ACB.

- An efficient SDT procedure is presented. This dissertation evaluates the performance of conventional SDT procedures in terms of the resource usage and resource utility. Based on the evaluations, an SDT procedure which omits the radio resource control (RRC) connection to reduce resource usage is proposed to improve resource efficiency. The numerical evaluation results show that the proposed procedure can reduce the resource usage comparing with other conventional SDT procedures.

## 1.4. Overview

The rest of the dissertation is organized as follows. In Chapter 2, the RACH procedure in cellular network is summarized, and the related works for the adaptive resource allocation and the congestion control are reviewed. In Chapter 3, the preamble partition based adaptive DARR protocol is proposed and evaluated. In Chapter 4, the preamble partition and the stochastic gradient descent based adaptive DARR and ACB protocol is proposed and evaluated. In Chapter 5, the conventional SDT procedures are summarized and evaluated in terms of resource usage and resource throughput. A new SDT procedure is also proposed and evalu-

ated in same chapter. The conclusion for the dissertation is presented in Chapter 6.

## Chapter 2. Background and Related Works

### 2.1. Data transmission procedure using RACH procedure in LTE-A

Recently, the LTE-A is the emerging standard for the mobile network market share [16]. In addition, the 5G network is expected to include the random access mechanism which is similar to that in LTE-A [17]. Thus, this dissertation reviews the random access procedure of LTE-A for the background of research.

An M2M device must start the RACH procedure to the BS in following situations in LTE-A [2]:

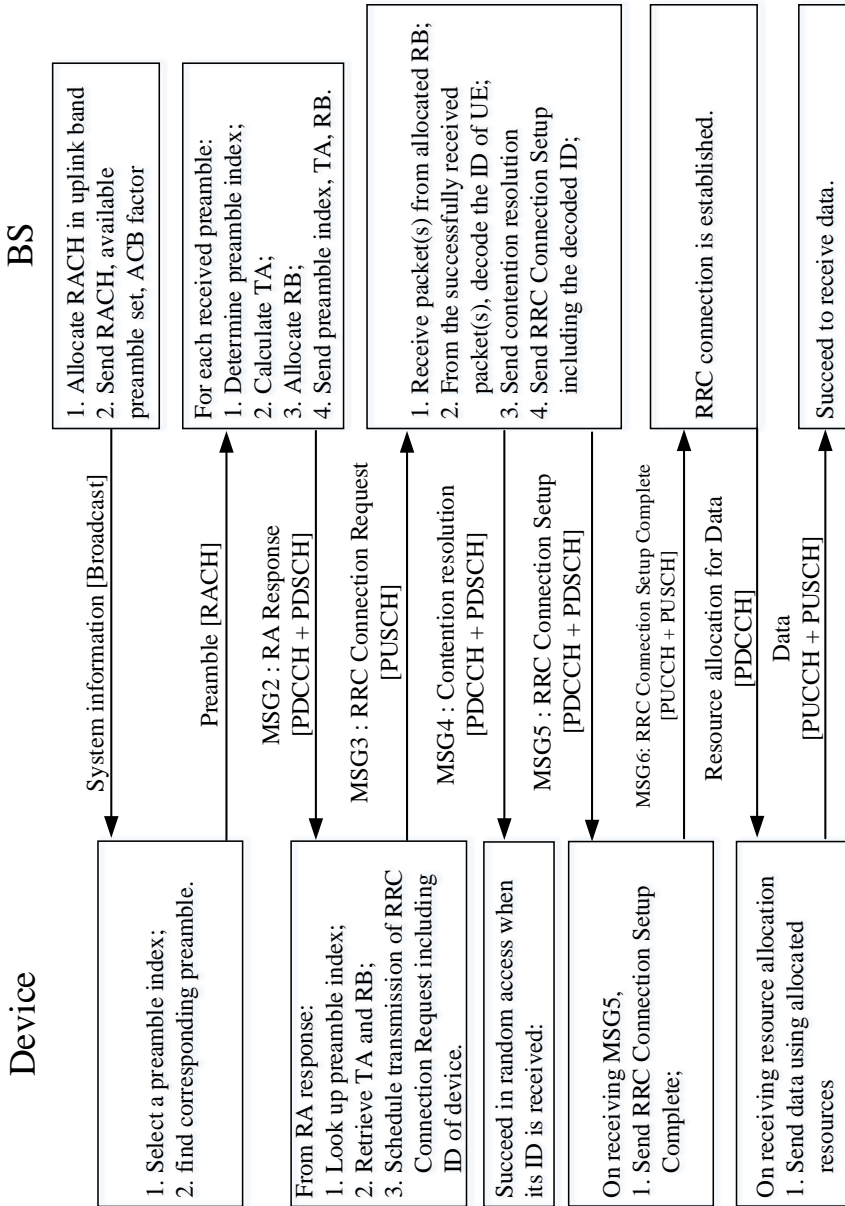
- Initial access to the network.
- When device requires to receive or transmit new data but the device is not synchronized.
- The transmission of new data when the uplink control channel (PUCCH) resources for scheduling request is not available.
- Handover.

- When the re-establishment of the connection is required after a radio link failure.

The M2M device must proceed the RACH procedure for the initial access. The device enters idle state if the data is not generated in a short time after the initial access. In the idle state, the device cannot maintain the synchronization with the BS [2]. Thus, the M2M device will use the RACH procedure for the transmission of data.

Figure 2.1 shows the radio resource control (RRC) connection establishment and data transmission using RACH procedure in conventional LTE-A [4, 8, 18–20]. An M2M device establishes the RRC connection between the device and a BS using RACH procedure, then the device transmits its data to the BS [21]. The data will be a message from non-access stratum (NAS) layer in the device when the device performs the RACH procedure for initial attach. The data will be an packet which includes upper layer headers and payload. The RACH procedure includes a four-message exchanges between the M2M device and the BS. After the RACH procedure, the BS can allocate resources for the transmission of data. The followings are the procedure for the uplink data transmission, which includes the RACH procedure.

**Step 1. Allocation of RACH:** Before start of RACH procedure, the BS periodically announces the time-frequency position of RACH, the available indexes of preambles, and related information for RACH procedure. The announcement is done by broadcasting system information block-2 (SIB2) message.



**Fig. 2.1.** Data transmission using RACH procedure in conventional LTE-A

**Step 2. Preamble transmission:** An M2M device desired to transmit its data first selects the next available RACH to try an access request. The device then selects a preamble index randomly from the set of preamble indexes. The device then transmits a preamble corresponding to the selected preamble index to the BS through RACH in uplink channel. The number of preamble indexes in LTE-A is 64. Each preamble index is corresponding to a preamble, where each preamble is orthogonal to each other. The preamble is the pseudo-random sequence generated from Zadoff-Chu sequence [22, 23]. If two or more devices transmit the same preamble in the same RACH, a collision occurs in the later step. However, the detection of collision is not available in this step. The preamble is sometimes referred as MSG1.

**Step 3. Random access response (RAR):** When the BS detects the preamble in RACH, the BS transmits a RA response (RAR) message using the physical downlink control channel (PDCCH) and the physical downlink shared channel (PDSCH). The device transmitted a preamble continuously observes the downlink control information (DCI) [24] in PDCCH until the RAR response window timer expires [11]. The DCI includes random access radio network temporary identifier (RA-RNTI) which is corresponding to the time and frequency of RACH that was used to transmit preamble. If the device can obtain the DCI with RA-RNTI corresponding to the transmitted preamble, the device searches corresponding PDSCH message which is pointed by the DCI. The PDSCH message includes the RAR message. RAR message includes the preamble index of detected preamble, time alignment instruction to synchronize,

uplink grant (UL grant) which is the uplink resource allocation to transmit third message, temporary cell radio network temporary identifier (TC-RNTI), and backoff indicator. The RAR is referred as MSG2.

**Step 4. RRC connection request:** When the device received the RAR corresponding to transmitted preamble, the device transmits the RRC connection request message using RBs in PUSCH. The time-frequency position for the RB can be found in UL grant. The RRC connection request message is referred as MSG3. If only one device transmits a preamble in step 2, MSG3 is transmitted to BS without collision. If two or more devices transmit the same preamble in step 2, the multiple devices transmit their MSG3 using same RBs. Thus, MSG3s experience collision [25]. If the BS can decode any MSG3, the BS transmits contention resolution in next step as a acknowledgment for MSG3. Otherwise, no contention resolution will be transmitted. The hybrid automatic retransmission request (HARQ) is used for the reliability of transmission for MSG3.

**Step 5. Contention resolution:** When the BS decodes a MSG3, it responses using a contention resolution to response to MSG3. The RRC connection setup message can be also transmitted with contention resolution. The transmission of contention resolution and the RRC connection setup message is referred as MSG4. If a device does not receive within contention resolution window after first transmission of MSG3, the device returns to step 2. Each device counts the transmission of preambles, and declare the failure of random access if the counter reaches the maximum value allowed by BS. The exchange of message from step 2 to step 5 is



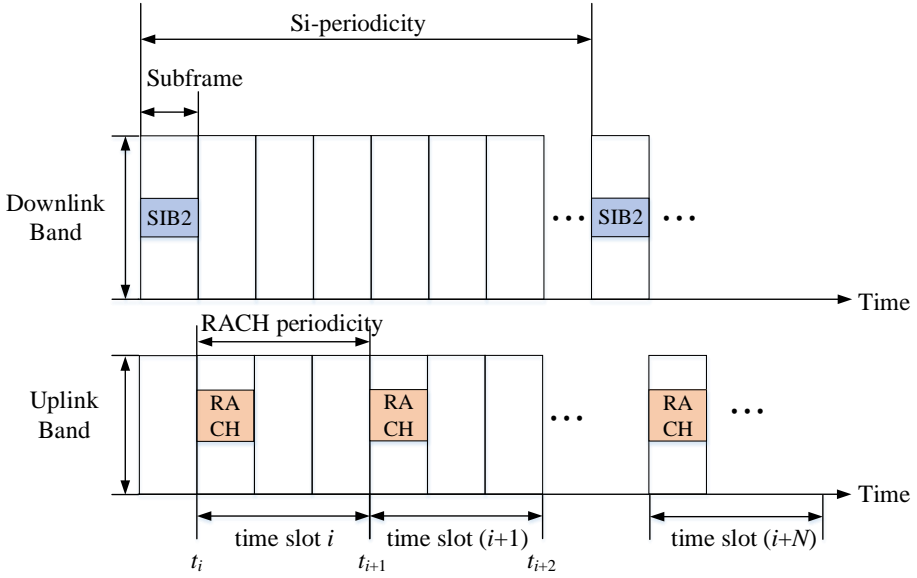
the RACH procedure. As in [8], the HARQ also can be applied for the transmission of MSG4.

**Step 6. RRC connection setup complete:** To finish the RRC connection between the device and the BS, the device receiving MSG4 transmits RRC connection setup complete to the BS. The RRC connection setup complete can be referred as MSG5. Since the MSG4 does not include the resource allocation for the transmission of MSG5, the additional resource request and allocation procedure between the transmission of MSG4 and MSG5 may be required.

**Step 7. Resource allocation for data:** For the transmission of data, the device requires the RBs in PUSCH. The device can request or the BS can allocate the RBs for the transmission of data after step 6. The BS transmits UL grant to the device to notify the time-frequency position of RBs.

**Step 8. Transmission of data:** The device received UL grant can transmit the data using the allocated resource. The HARQ can be used for the reliability of transmission.

Figure 2.2 shows about the allocation of RACH and SIB2 transmission in wideband LTE-A. The RACH is allocated at specific subframes [22] given as the configuration in system information block-2 (SIB2) [26]. The BS periodically transmits SIB2 in downlink channel to notify the position of RACH, where the periodicity is given as “si-periodicity”. The other parameters for RACH are also transmitted using SIB2, where the parameters includes the information for the pool of preambles and access class barring factor (ACB factor), and others. The periodicity for SIB2

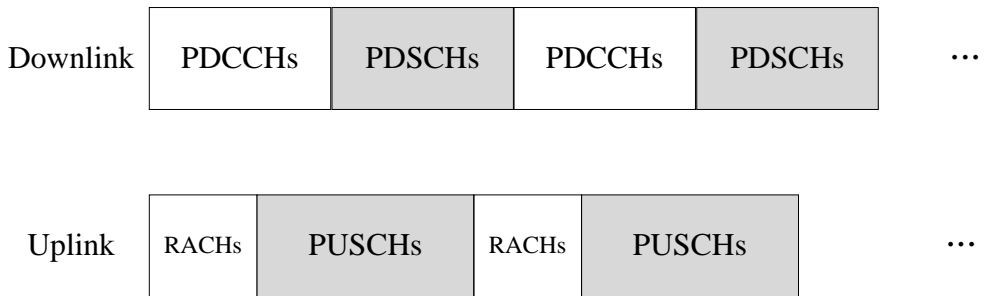


**Fig. 2.2.** RACH and SIB2 transmission in frame structure of LTE-A

is generally longer than the RACH periodicity, where RACH periodicity is the interval between two RACH [26]. The other spaces in downlink band can be used for PDCCH and PDSCH, and that in uplink band can be used for and PUSCH.

The frame structure for new radio technology will have different number of symbols per subframe [27–29]. Currently, the periodicity for RACH and the periodicity for SIB2 is not explicitly determined. Since the number of symbols per subframe increases, both of the periodicity for RACH and the periodicity for SIB2 can be decreased. Still, the periodicity for RACH can be different to the the periodicity for SIB2, where the difference can result the similar structure as in Fig. 2.2.

The frame structure for narrow band IoT (NB-IoT) is similar to the



**Fig. 2.3.** RACH and SIB2 transmission in frame structure of LTE-A

frame structure of wideband LTE-A, but the bandwidth becomes narrow (180 kHz). Figure 2.3 shows the frame structure for NB-IoT. The PDCCH and PDSCH are repeatedly allocated in the downlink band, and the RACH and PUSCH are repeatedly allocated in the uplink band. In the NB-IoT, RACH occupies all uplink bandwidths and its duration is equal to 8 ms, i.e. 8 subframes. Although the time-frequency positions for each channel are different, the SIB2 is also transmitted with the periodicity of si-periodicity. The transmission of SIB2 requires a PDCCH and some of resource blocks in PDSCH.

The default RACH periodicity can be from 1 to 20 subframes in wideband LTE-A. Si-periodicity can be one of following values: 8, 16, 32, 64, 128, 256, 512 frames. Since 1 frame is equal to 10 subframes in LTE-A, the number of RACHs in a si-period can be from 4 to 5120. For example, if the RACH periodicity is 5 subframes as in [8] and if si-periodicity is 16 frames, then the number of RACHs in a si-period is equal to 32. The si-periodicity for NB-IoT can be from 640 subframes to 40960 subframes where RACH periodicity can be 40 subframes to 2560

subframes [26].

## 2.2. Access class barring and its related works

The 3GPP proposed several expected solutions to overcome the traffic overload by M2M devices in RAN [8, 30]. The solutions include the access class barring (ACB) scheme, separation of RACH resources for M2M devices, the dynamic allocation of RACH resources (DARR), MTC specific backoff, slotted access, and pull based scheme. The most researches for congestion control in LTE-A focuses on ACB, especially for the user equipment (UE) individual ACB scaling, due to the limit on the number of preambles and resource blocks. In ACB, the number of contended devices is controlled by announcing an ACB factor, where ACB factor is a value representing a probability. If ACB factor is announced as  $p$ , then the devices can transmit preamble with probability  $p$ , or defer its preamble transmission for one RA period by probability  $(1 - p)$  [2, 31].

Duan and others presented that the optimal ACB factor is the ratio of the number of preambles to the number of activated devices in network [32]. In addition, the heuristic algorithm to update ACB factor in BS is given in [32] where this algorithm assumes that the BS knows the number of users in the cell. In [4], which is later study of Duan et al., they proposed dynamic ACB with fixed resource allocation (D-ACB with FRA) for the fixed number of preambles. D-ACB with FRA estimates the activated number of devices and derives ACB factor for given number of preambles. He and others [33] presented that the number of activated

devices can be estimated by the ratio of the number of contended devices in a RACH to the ACB factor used in the RACH. In addition, they proposed the dynamic ACB factor control algorithm based on the traffic arrival model in [8]. Tavana and others [34] derived the ACB factor by predicting the number of activations in next RACH. The prediction based algorithm is improved by applying Kalman filter. Moon and Lim proposed the adaptive ACB when the distribution of device arrival is known [35]. Koseoglu also proposed the adaptive ACB scheme which is the pricing based load control when the arrival distribution is known [36].

The extended access barring (EAB) can be used which blocks random access of low priority devices during congestion [37]. The EAB broadcasts the barring information for access classes, and blocks the access of barred access classes.

The class based access barring is also studied to ensure the human-to-human communication when the M2M devices share the network with human devices. The prioritized random access with dynamic access barring (PRADA) is proposed in [38]. In PRADA, the BS virtually allocate the RACH for multiple number of classes of traffic. For example, a RACH in a specific time can be used by devices if and only if the device has high priority. In addition, the BS can delay the random access of low priority devices if the network shows high traffic load.

The backoff scheme can be applied, which changes the backoff duration for devices according to the congestion state of the network. In [39], the duration of backoff is dynamically determined according to the estimated number of devices. The optimal parameters used for RACH

procedure in LTE-A are also investigated. In [40], authors argue that the optimal backoff window size for conventional RACH procedure is 1 and the maximum number of trials for RACH procedure should be limited to 2 or 3 when the collision probability is less than 5%. In [41], the performance analysis with the exponential backoff window is performed where the backoff in standard LTE-A has fixed backoff window.

However, the basic of ACB, EAB, or backoff scheme is to postpone the access of devices to later RA slots. Thus, these methods increase the average delay for all devices proportionally to the number of devices in the network. Although the applications of M2M communications has loose delay requirement, the requirement may be dissatisfied with intensive device arrivals by the massive number of devices.

The ACB and its related works are summarized in Table 2.1 and Table 2.2.

### **2.3. Dynamic allocation of RACH resource**

In dynamic allocation of RACH resource (DARR), the BS dynamically changes the number of preambles and corresponding number of resource blocks. The concept of DARR is proposed by 3GPP in [8] without detail. Lo and others proposed the self-optimizing overload control scheme which adjusts the number of RACHs per second based on the observed collision probability [42]. Hwang and others proposed the dynamic RACH resource separation scheme which adaptively controls the ratio of the preambles for M2M devices to the total number of preambles [43], when

**Table 2.1.** Summary of ACB and related schemes (I)

| Paper      | Description  |
|------------|--|
| [8]        | Presented the concept of ACB   |
| [32]       | ACB : Suggest optimal $p$ for ACB<br>The heuristic algorithm is given (assume that BS knows the number of devices in cell)   |
| [33]       | ACB : Suggest optimal $p$ for ACB using previous $p$<br>Propose a ACB factor control for known traffic arrival model         |
| [34]       | ACB : Predict the number of arrivals for next RACH to derive ACB factor<br>Improve algorithm by Kalman filter                |
| [4]        | ACB : ACB for fixed and variable number of preambles   |
| [35], [36] | ACBs for known traffic model   |
| [37]       | EAB : Introduction of the concept of EAB and performance evaluation  |
| [38]       | Class based access : Class based RACH allocation and the delaying of access<br>for low priority devices in high traffic load |

**Table 2.2.** Summary of ACB and related schemes (2)

| Paper | Description   |
|-------|---|
| [39]  | Backoff : The performance evaluation of simulation in [8] is performed, then dynamic backoff scheme is proposed |
| [40]  | Backoff : Optimal parameters for RACH procedure is studied  |
| [41]  | Backoff : Exponential backoff is studied  |



the M2M and H2H devices share the preambles. Li and others did similar study but they found their preamble allocation by solving an optimization problem [44]. Li and others' work is expanded in [45]. Later, Choi proposed the adaptive determination of the number of preambles (ADP) in [12]. ADP not uses the backoff as the fast retrial algorithm [46] since the backoff is not useful for the congestion control. In addition, ADP adaptively changes the number of preambles. In [12], the optimal number of preambles is given by stochastic gradient ascending method, which becomes

$$N = \hat{L}_i + \alpha(\hat{M}_i - \hat{L}_i), \quad (2.1)$$

where  $N$  is the selected number of preambles,  $\hat{L}_i$  is the estimated optimal number of preambles from last observation,  $\hat{M}_i$  is the estimated number of devices, and  $\alpha$  is a step size. The number of preambles then broadcasted to devices in before each start of RACH.

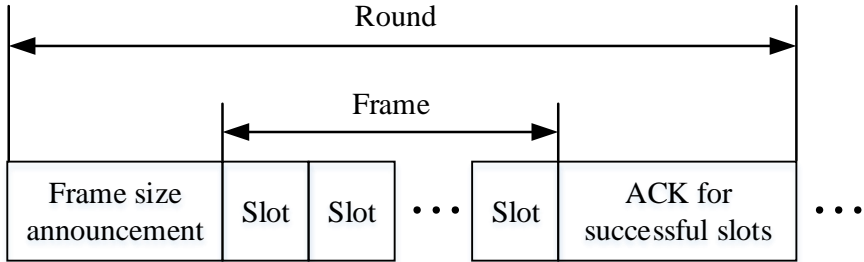
Duan and others proposed dynamic ACB with dynamic resource allocation (D-ACB with DRA) in [4] for variable number of preambles, which is the combination ACB and DARR. D-ACB with DRA first estimates the activated number of devices. Based on the estimated number, D-ACB with DRA selects the number of preambles using a scaling factor,  $b$ , and ACB factor. The selected number of preambles and ACB factor are changed and announced in before every start of RACH.

The random access procedure in LTE-A is similar to the multi-channel slotted ALOHA (MCSA). Thus, the applications of researches in MCSA is studied. In [47], the modeling and estimation for the MCSA was performed considering the environment in [8]. In this research, the ratio of

the number of successful access to the number of contended devices in an access cycle is maximized when the number of random access opportunities in an access cycle is equal to the number of contended devices. Note that the access cycle is corresponding to RACH and the random access opportunity is corresponding to the preamble.

DARR dynamically adjusts the number of preambles, where this approach can be seen in the studies for dynamic frame slotted ALOHA (DFSA). Figure 2.4 shows the generalized frame structure of the frame slotted aloha (FSA) [48–50]. In the FSA, a BS constructs a frame which includes the fixed number of multiple time slots. Each time slot can be used for the data transmission. The number of time slots in a frame is sometimes represented as the frame size of the frame. The frame size is broadcasted to devices before the start of new frame by a frame size announcement message in Figure 2.4. A device randomly selects a time slot in frame and transmits its data to the BS using the time slot. The BS can detect and decode the data without collision if only one device has selected the time slot. The collision occurs when two or more devices transmit their data in a time slot.

Reducing collision or increasing the number of data transmission without collision is important for DFSA. The devices in radio frequency identification (RFID) is very simple, thus the devices can always transmit their data to the BS in every frame although the data is already delivered to the BS [51]. To reduce the collision, the BS can mute devices that succeeded their transmission of data. The frame slotted ALOHA has an optimal number of devices being allowed to access the BS in a frame.



**Fig. 2.4.** Frame structure of FSA

Thus, the BS can control the number of contending devices per a frame, where the number of contending devices represents that the number of devices transmitted their data in a frame. The BS can mute some of devices although these devices not succeeded their transmission of data.

If the frame size can be variable, the BS can change the number of slots in a frame to change the optimal number of devices for the frame. This approach is referred as the DFSA. Vogt proposed the update algorithm of frame size for low identification delay, which updates the frame size using binary exponential algorithm [50]. Zhen and others proposed 1.4 times of the estimated number of devices to achieve low level of collisions [48]. Cha and others update the frame size equal to the estimated number of devices to maximize throughput, where throughput is the ratio of the number of successful slots in frame to the frame size [52, 53]. Another study also optimized the throughput [54]. Khandelwal and others set the frame size to 1.943 times of the estimated number of devices to reduce total time for collecting all data [49]. Prodanoff proposed the

frame size equal to  $M/(\ln 2)$  where  $M$  is the total number of devices [55]. Lee et al. proposed the frame based on the Lambert's Omega function to minimize data collection time when the durations of idle, success, collided slots are different [56]. Dhakal et al. proposed the frame size of  $3.49c_2$  where  $c_2$  is the number of collided slots, and proposed the partitioning of devices for fast data collection [57]. Kim proposed a frame size update algorithm which is equal to

$$\arg \min_{N \in A_K} |N - M \ln 2|, \quad (2.2)$$

to maximize success probability, where  $N$  is a candidate for frame size in set  $A_K$ , and  $M$  is the estimated number of devices in previous frame [58]. The  $Q$  protocol is also proposed [59].  $Q$  is an integer between zero and eight, and the frame size becomes equal to  $2Q$ . After finishing a frame, the BS increases or decreases  $Q$  by a constant  $c$ , for each collision or idle slots, respectively. The final value of  $Q$  is then used to determine the frame size for next frame.

The estimation of the number of contending devices in RA slot can be required for the BS. The estimation is similar to the studies in DFSA although the details need to be studied. In DFSA, the BS estimates the number of devices using the number of slots filled with zero ( $c_0$ ), one ( $c_1$ ), and multiple packets ( $c_2$ ) in frame. Vogt [50] presented the minimum number of devices involved in the contention in a frame, where the number is  $c_1 + 2c_2$ . Vogt also presented the estimation method based on Chebyshev's inequality and based on the difference between actual results, and the theoretically computed results. Later, Zhen and others

presents the estimation using the number of collided devices [48]. Chen and others presented the estimation with the idle slots and successful slots [54]. Cha and others presented the estimation with the ratio of the number of collided slots to the number of slots in the frame. In addition, Cha and others also presents the alternative estimation, which is equal to  $2.39c_2$  [52, 53]. Khandelwal and others estimates the number of devices using the unused slots in the frame [49]. Let the estimated number of devices is equal to  $\hat{M}$ . Khandelwal and others' suggestion is equal to

$$\hat{M} = \begin{cases} \frac{\log(c_0/N)}{\log(1-1/N)} & ; c_0 > 0, \\ c_1 + 2c_2 & ; c_0 = 0, \end{cases} \quad (2.3)$$

where  $N$  is the frame size and  $c_0$  is positive.

The fixed number of preambles per cell limits the number of successful access per RACH, thus the studies to increase the effective number of preambles are also performed. In [60], the effective number of preambles is increased by the combination of preamble arrival results in two RACHs. For example, the arrival of a preamble in first RACH and the idle of the preamble in second RACH can be regarded as a new type preamble. In [61], the BS divides the cell into multiple areas to spatially reuse the preambles. The sparse coding multiple access (SCMA) or other non-orthogonal multiple access can expand the random access resource that can be used for each devices [62]. These studies can enable the expansion of the preamble pools per cell without expansion of preamble sequences.

The studies for DARR or DFSA are summarized in Table 2.3 and Table 2.4.

**Table 2.3.** Summary of DFSA and ADP (1)

| Paper      | Description   |
|------------|---|
| [42]       | <b>DARR:</b> Observe the collision probability and adjust the number of RACHs per second  |
| [43], [44] | <b>DARR:</b> The ratio of the resources for M2M devices to the total resources is selected  |
| [12]       | <b>DARR:</b> stochastic gradient method   |
| [4]        | <b>DARR:</b> Decide the number of preambles based on the number of activated devices and a scaling factor   |
| [47]       | <b>Multi-channel slotted ALOHA:</b> Modeling the random access of M2M devices as the MCSA   |
| [50]       | <b>Estimation:</b> Present minimum number of contended devices,<br>Present the estimation method based on Chebyshev's inequality,<br>Present the estimation method based on the difference between actual results and the theoretically computed results. <b>DFSA:</b> Binary exponential algorithm |
| [48]       | <b>Estimation:</b> Estimation based on the number of collided slots<br><b>DFSA:</b> 1.4 of the estimated value (reduce collision)   |

Table 2.4. Summary of DFSA and ADP (2)

| Paper      | Description  |
|------------|--|
| [52], [53] | <p><b>Estimation:</b> Estimation based on the number of collided slots</p> <p><b>DFSA:</b> Equal to the estimated value to maximize throughput</p> |
| [49]       | <p><b>Estimation:</b> Estimation based on the number of idle slots</p> <p><b>DFSA:</b> 1.943 of the estimated value to reduce collision</p>        |
| [59]       | <p><b>DFSA:</b> Frame size is <math>2^Q</math> where <math>Q</math> can be increased or decreased by a step size, <math>c</math></p>               |
| [55]       | <p><b>DFSA:</b> Propose frame size to maximize success probability.</p>  |
| [58]       | <p><b>DFSA:</b> Present the optimum size of frame when the frame size can be <math>2^n</math>, <math>n = 1, 2, \dots</math></p>                    |
| [60]       | <p><b>Increase RACH resource:</b> Propose the code expanded RA to increase effective number of preambles in cell</p>                               |
| [61]       | <p><b>Increase RACH resource:</b> Propose the spatial reuse of preambles</p>   |

# Chapter 3. Preamble Partition based Adaptive DARR Protocol

## 3.1. Introduction

IoT service includes the security and public safety networking services. In case of an emergency, the data from all IoT devices needs to be collected as soon as possible [4]. Thus, the communication system requires to minimize the total amount of the time to gather data packets from all activated devices. The dynamic allocation for RACH resource (DARR) scheme is one of many solutions to minimize the time, where the DARR adaptively changes the size of pool and other resources [8, 15]. The problems in DARR are similar to those in the studies in dynamic frame slotted aloha (DFSA). The DFSA assumes that the size of pool can be updated in every RA slot. However, the parameters related to RACH procedure can be updated with periodicity in mobile network, where the multiple RA slots are allocated during the period. The BS cannot directly apply the schemes in DFSA due to the periodicity, since the studies for DFSA achieve their objectives without the periodicity.

In this chapter, the discussion about the throughput degradation of



DARR in LTE-A is presented, which is resulted from the periodicity of updating the size of the pool. In addition, a preamble partition approach based DARR protocol is presented to resolve throughput degradation problem in DARR. The proposed approach separates a pool into two pools to select the size of the pools to obtain information for the decision of the pool size. The performance evaluation using simulations is represented where the results show that the proposed approach can achieve performance that is closer to the optimal throughput of FSA than the throughput without the proposed approach <sup>1</sup>.

### 3.2. System Model

Suppose that a cell consists of a BS and  $M$  MTC devices. These devices are activated not simultaneously but with in a short period of time,  $T_a$ , which is referred as the "activation time" in this chapter [4]. Each devices are activated at time  $t$  with a probability density function of  $g(t)$ . Let  $I_a$  be the number of RACHs within the activation time. The activation time is divided into  $I_a$  discrete slots, where each slot has longer duration than a RACH. The  $i$ -th slot begins with  $i$ -th RACH, where  $i = 1, 2, \dots, I_a, \dots$ . The length of each time slot is equal to the interval between two consecutive RACHs, where this interval is sometimes referred as RACH period in this chapter. Thus, the time slot and RACH period is changeably used in this chapter. Let  $T_{RAREP}$  be the length of each time slot. The  $i$ -th

---

<sup>1</sup>Works in this chapter was accepted and will be published on the ETRI Journal [63]

time slot begins at time  $t_{i-1}$  and ends at time  $t_i$ . Let  $\lambda_i$  be the number of activations during time slot  $i$ , which is equal to

$$\lambda_i = \left\lceil M \int_{t_{i-1}}^{t_i} g(t) dt \right\rceil, i = 1, 2, \dots, \quad (3.1)$$

where  $\lceil x \rceil$  is the smallest integer equal or larger than  $x$ . Note that [8] recommended the uniform distribution or Beta distribution for  $g(t)$ .

Let  $K$  be the number of time slots during a si-period. The  $r$ -th si-period will include time slots from  $((r-1)K+1)$ -th time slot to  $(rK)$ -th time slot. The  $i$ -th time slot ( $i=1,2,\dots$ ) is included in the  $r$ -th si-period ( $r=1,2,\dots$ ) with a relation given as

$$r = 1 + \lfloor (i-1)/K \rfloor, \quad (3.2)$$

where  $\lfloor x \rfloor$  is the largest integer equal or smaller than  $x$ .

Let the SIB2 for  $r$ -th period be changed and broadcasted before the start of  $(r-1)K+1$ -th time slot. Let  $T_{UPDATE}$  be the si-periodicity shown in Figure 2.1, which is the length of each si-period.  $K$  becomes

$$K = \lfloor T_{UPDATE}/T_{RAREP} \rfloor. \quad (3.3)$$

The SIB2 includes information for preamble pool. Suppose that there are  $R_r$  preambles in the pool for the  $r$ -th si-period, where the pool is given as  $\mathbf{C}_r = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{R_r}\}$ . The minimum size of pool is given as  $R_{min}$ , i.e.,  $R_r \geq R_{min}$ . An activated device transmits a preamble through a RACH, where the preamble is randomly chosen from the preamble pool. In this chapter, this dissertation assumes that the size of the preamble pool can be variable without upper limit. Note that the pool size

can be increased by increasing the number of preamble sequences, or by allocating the multiple number of RACHs per  $T_{RAREP}$ .

An MTC device can transmit preambles up to  $N_{\max}$  times [8]. This dissertation also assumes that the preamble is always detectable in the BS [4, 12]. A collision always occurs when two or more devices select the same preamble in the same RACH period [4]. The transmissions of MSG3 and MSG4 are assumed to always successful, since hybrid automatic repeat request (HARQ) provides very high transmission probability for MSG3 and MSG4 [8]. The performance of the RACH procedure with this system model thus depends on the collision probability.

Assume that  $M_i$  devices transmit preambles randomly chosen from  $\mathbf{C}_r$  in  $i$ -th RACH period. In a practical system, a BS cannot know  $M_i$  but can estimate it. Let  $\hat{M}_i$  be the estimate for  $M_i$ . Let  $\mathbf{O}_i = [O_{i,0}, O_{i,1}, O_{i,2}]$  be the observation vector for the preambles allocated for the  $i$ -th RACH period, where  $O_{i,0}$ ,  $O_{i,1}$ , and  $O_{i,2}$  are the observed number of unused, successful, and collided preambles, respectively. The BS can use an estimation function of  $\mathbf{O}_i$  and  $R_r$ ,  $f(\mathbf{O}_i, R_r)$ , to obtain  $\hat{M}_i$  after completing RA in the  $i$ -th RACH period, which can be represented as

$$\hat{M}_i = f(\mathbf{O}_i, R_r). \quad (3.4)$$

In this dissertation, the BS attempts to increase the throughput. Let  $S_i$  be the number of devices that successfully transmit their preamble in the  $i$ -th RACH period. Let  $U_i$  be the throughput for each  $i$ -th RACH period. The throughput is the ratio of the number of successful preambles in a RACH period to the allocated number of preambles in the RACH period,

i.e.  $U_i = S_i/R_r$ .

Let  $T_{RAR}$  be the waiting time before the start of the random access response window, let  $W_{RAR}$  be the size of the RAR window, and let  $W_{BO}$  be the size of the backoff window. The devices failed the random access in from  $(i - k_2)$ -th RACH period to  $(i - k_1)$ -th RACH period will backoff to  $i$ -th RACH period, where  $k_1$  and  $k_2$  are obtained by

$$k_1 = \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RAREP}} \right\rceil, \quad (3.5)$$

$$k_2 = \left\lceil \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RAREP}} \right\rceil. \quad (3.6)$$

Let  $B_{j \rightarrow i}$  be the number of devices backoffed from  $j$ -th RACH period to  $i$ -th RACH period. From the definition of  $k_1$  and  $k_2$ , we can get

$$\sum_{j=k_1}^{k_2} B_{i \rightarrow (i+j)} = M_i - S_i. \quad (3.7)$$

From (3.7),  $M_i$  can be represented as

$$M_i = \lambda_i + \sum_{j=k_1}^{k_2} B_{j \rightarrow i}. \quad (3.8)$$

Let  $P_{S,i}$  be the success probability for the preamble transmission in the  $i$ -th RACH period. Since other devices should select other preambles given that a device transmitted a preamble,  $P_{S,i}$  becomes [12]

$$P_{S,i} = \left(1 - \frac{1}{R_r}\right)^{M_i-1}. \quad (3.9)$$

The conditional mean of  $S_i$  for given  $M_i$  and  $R_r$  becomes

$$\mathbb{E}[S_i | M_i, R_r] = M_i P_{S,i} = M_i \left(1 - \frac{1}{R_r}\right)^{M_i-1}, \quad (3.10)$$

where  $\mathbb{E}[\cdot]$  denotes the statistical expectation. For a given  $M_i$ , the expectation of throughput becomes

$$\mathbb{E}[U_i|M_i, R_r] = \mathbb{E}[S_i|M_i, R_r]/R_r = \frac{M_i}{R_r} \left(1 - \frac{1}{R_r}\right)^{M_i-1}. \quad (3.11)$$

From  $\partial E[U_i|M_i, R_r]/\partial R_r = 0$ , we can obtain that the throughput is maximized when  $R_r = M_i$  with expected throughput of  $E[U_i|M_i, R_r] \simeq e^{-1}$  [12, 52, 53].

Let  $\Gamma$  be the time from the transmission of preambles to the end of RAR window. Let  $\Psi$  be the time from the transmission of preambles to the end of backoff window. They are given as

$$\Gamma = T_{RAR} + W_{RAR}, \quad (3.12)$$

$$\Psi = T_{RAR} + W_{RAR} + W_{BO}. \quad (3.13)$$

Let  $\alpha_{j \rightarrow i}$  be the ratio of the devices backoffing to  $i$ -th time slot to the devices required to backoff from  $j$ -th time slot.  $\alpha_{j \rightarrow i}$  is equal to

$$\alpha_{j \rightarrow i} = \begin{cases} \alpha_a = ([\Gamma/T_{RAREP}]T_{RAREP} - \Gamma)/W_{BO} & ; j = i - k_1, \\ \alpha_d = (\Psi - T_{RAREP}[\Psi/T_{RAREP}])/W_{BO} & ; j = i - k_2, \\ \alpha_{bc} = T_{RAREP}/W_{BO} & ; \text{otherwise.} \end{cases} \quad (3.14)$$

The expectation of  $M_i$  is derived in [64], which is the analysis model for [8]. Let  $M_i[n]$  be the number of devices that transmits its preamble in  $i$ -th time slot, where the number of the preamble transmission is equal to  $n$ . (i.e.  $n=1$  for new arrival and  $n > 1$  for backoff). According to [64],

the expectation for  $M_i$ ,  $E[M_i]$ , is equal to

$$\begin{aligned}\mathbb{E}[M_i] &= \mathbb{E}[M_i[1]] + \sum_{n=2}^{N_{\max}} \mathbb{E}[M_i[n]] \\ &= \mathbb{E}[M_i[1]] + \sum_{n=2}^{N_{\max}} \sum_{j=i-k_2}^{i-k_1} \alpha_{j \rightarrow i} (1 - P_{S,j}) \mathbb{E}[M_j[n-1]].\end{aligned}\tag{3.15}$$

### 3.3. Problem definition : Decrease of throughput in DARR due to resource allocation update interval

In this section, this dissertation discusses about the throughput degradation of the DARR. The throughput degradation is caused by the difference of the assumptions in previous studies for DARR and that in actual mobile network. More precisely, the previous studies assume that si-period is equal to RACH period, but si-period is larger than RACH period in LTE-A. For the description, this paper first summarizes about the throughput optimization in DFSA or in conventional studies for DARR in LTE-A [12]. In addition, this paper discusses the degradation of throughput in the DFSA due to the si-periodicity, where the si-periodicity is corresponding to the periodicity of SIB2.

The BS cannot know  $M_i$  in the network without explicit signaling. However,  $M_i$  is important to select the pool size since the throughput is maximized when  $R_r = M_i$ . If the  $M_i$  is similar to  $M_{i-1}$ , the BS can select  $R_r = M_{i-1}$  or  $R_r = \hat{M}_{i-1}$  for the random access in  $i$ -th time slot. Fortunately,  $M_i$  includes the number of devices backoffed from previous time slots, and  $\lambda_i$  and  $\lambda_{i-1}$  are correlated according to the arrival distribution for the activation of devices [8, 14]. Based on the correlation between

$M_i$  and  $M_{i-1}$ , several studies propose their schemes to find optimal pool size for their objectives. In these studies, the si-periodicity in LTE-A is assumed same as that in DFSA, which means that the BS can update the parameters at start of every RACH period, i.e.  $T_{UPDATE} = T_{RAREP}$  or  $K = 1$  [4, 12]. In this case, the BS will update the size of pool as

$$R_{r+1} = \max([\hat{M}_i], R_{\min}). \quad (3.16)$$

However, the mobile network generally has  $K > 1$ . For example, the LTE-A has  $K > 1$  due to si-periodicity, therefore the RA procedure in LTE-A operates with the condition of  $T_{UPDATE} > T_{RAREP}$ . In this case, the BS needs to determine the pool size using multiple  $\hat{M}_i$  because there are multiple time slots in a si-period. In the previous studies for LTE-A, no approaches were presented for determining the size of the preamble pool from multiple observations.

The selection methods in previous studies can be considered as the "most recent" value based methods since they decide the preamble pool using most recent  $\hat{M}_i$ , which can be represented as

$$R_{r+1} = \max\left(\left[\hat{M}_{(rK)}\right], R_{\min}\right). \quad (3.17)$$

The one of most known method to obtain a representative value from multiple observation is the averaging of observations [65]. Thus, the sample mean of  $\hat{M}_i$  during the most recent si-period can be used to obtain  $R_{r+1}$ . Let "mean" based method be this approach, which can be represented as

$$R_{r+1} = \max\left(\left[\frac{1}{K} \sum_{i=(r-1)K+1}^{rK} \hat{M}_i\right], R_{\min}\right). \quad (3.18)$$

**Table 3.1.** Summary for conventional DARR policies

| Policy        | Mathematical representation  |
|---------------|--|
| Most Recent   | $R_{r+1} = \max \left( \left[ \hat{M}_{(rK)} \right], R_{\min} \right)$  |
| Mean          | $R_{r+1} = \max \left( \left[ \frac{1}{K} \sum_{i=(r-1)K+1}^{r \cdot K} \hat{M}_i \right], R_{\min} \right)$   |
| Weighted Mean | $R_{r+1} = \max \left( \left[ \frac{1}{K} \frac{\sum_{i=(r-1)K+1}^{r \cdot K} w_i \hat{M}_i}{\sum_{i=(r-1)K+1}^{r \cdot K} w_i} \right], R_{\min} \right)$ |
| Max           | $R_{r+1} = \max \left( \left[ \max_{r(K-1)+1 \leq i \leq rK} \hat{M}_i \right], R_{\min} \right)$  |

Since the recent value is important than older values, “weighted sample mean policy” can be used for the determination of pool size. In the weighted sample mean policy, weight  $w_i$  is multiplied by  $\hat{M}_i$  in (8), which can be represented as

$$R_{r+1} = \max \left( \left[ \frac{1}{K} \sum_{i=(r-1)K+1}^{r \cdot K} w_i \hat{M}_i \right], R_{\min} \right). \quad (3.19)$$

In addition to these policies, the “max policy” can be used, which selects the maximum value from  $K$  observations for  $R_{r+1}$ , which can be represented as

$$R_{r+1} = \max \left( \left[ \max_{r(K-1)+1 \leq i \leq rK} \hat{M}_i \right], R_{\min} \right). \quad (3.20)$$

Table 3.1 summarizes the conventional approaches to decide the size of pool.

Unfortunately, the existence of the si-periodicity and the mis-selection of initial preamble pool can cause rapid changes of  $M_i$ , since the BS



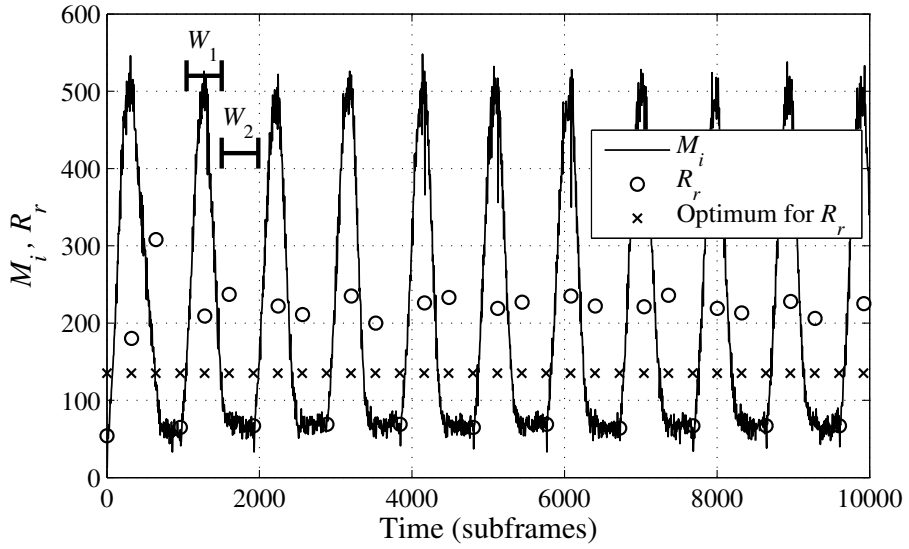
cannot adjust the parameters for the RACH procedure during the si-periodicity. For example, let devices arrive with uniformly distributed arrival with  $\mathbb{E}[\lambda_i] = \bar{\lambda}$ . Let  $T_{RAR} = W_{RAR} = W_{BO} = 0$ . In this case,  $M_i = \lambda_i + M_{i-1} - S_{i-1}$ . In addition,  $\mathbb{E}[S_i|M_i, R_r]$  needs to be equal to  $\bar{\lambda}$  to maximize throughput. Let  $R_r^*$  be the optimal size of preamble pool. If  $R_1$  is selected as  $R_1 < R_1^*$ ,  $\mathbb{E}[S_i|M_i, R_r]$  becomes smaller than  $\bar{\lambda}$ , and  $M_i$  will increase in first si-period. Conventional methods will select  $R_2$  as  $R_2 > R_2^*$ . In second si-period,  $\mathbb{E}[S_i|M_i, p_r, R_r]$  becomes larger than  $\bar{\lambda}$ , thus  $M_i$  will continuously decrease. Conventional methods will select  $R_3$  as  $R_3 < R_3^*$ , which returns to the situation for first si-period. Thus,  $M_i$  will change with continuous fluctuation with peaks and troughs until the activation of devices stops.

Although  $T_{RAR}$ ,  $W_{RAR}$ , and  $W_{BO}$  are non-zero, fluctuation can be observed. To maximize the expected throughput,  $M_i$  needs to be equal to  $R_r$  without regarding to  $i$ . This can be possible if  $P_{S,i} = e^{-1}$  and  $M_i[1] = c \forall i$  where  $c$  is an arbitrary constant. However,  $M_i$  or  $M_i[1]$  are random variables so  $P_{S,i}$  can be different to  $e^{-1}$ . Because  $R_r$  is fixed during a si-periodicity and (3.15) is recursive,  $M_i$  is expected to increase during si-periodicity if  $P_{S,i}$  becomes smaller than  $e^{-1}$ , and is expected to decrease during si-periodicity if  $P_{S,i}$  is larger than  $e^{-1}$ .  $\hat{M}_i$  is also affected by the change. Therefore, the selection of  $R_{r+1}$  using the sample mean, the weighted sample mean, or the latest value of  $\hat{M}_i$  cannot be effective because of the changes of  $M_i$  during the si-periodicity.

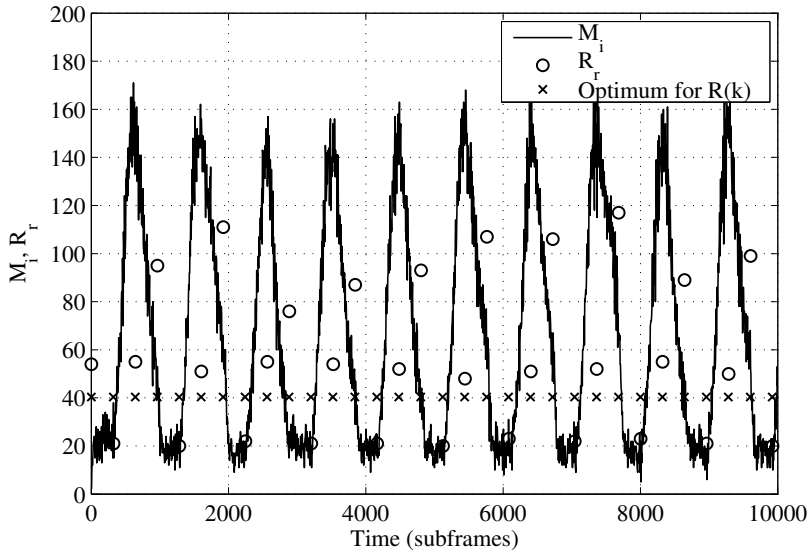
The simulations for the selection of  $R_{r+1}$  using the sample mean shows the ineffectiveness in throughput. Figure 3.1 shows the values of  $M_i$ ,  $R_r$ ,

and the optimum value for  $R_r$ , which is defined as  $R^*$ , in a simulation when  $R_{r+1}$  is updated using (3.18), the devices arrive with uniform distribution, and the  $M$  of 100,000. The detailed description for the parameters used for the figure will be presented in Section 3.5. The difference between the determined parameters ( $R_r$ ) and the optimum parameters ( $R^*$ ) can be observed in Figure 3.1.  $R^*$  is about 135 in this example.  $R_1$  is equal to 54, which is too small to support the active devices in the time slots. Thus,  $P_{S,i}$  is lower than  $e^{-1}$ . The collided devices in a time slot backoff to later time slots, thus the number of contending devices in a time slot increases,  $P_{S,i}$  also decreases, and repeats during first si-period. Therefore,  $M_i$  increases exponentially in first si-period. Thus,  $R_2$  is selected as a larger value than  $R^*$ , however it is still small to resolve the congestion. Therefore,  $R_3$  is selected as a very large value. In third si-period,  $P_{S,i}$  increases larger than  $e^{-1}$  and  $M_i$  becomes lower than 135. Thus,  $R_4$  becomes too small than the optimum value. Again,  $P_{S,i}$  decreases lower than  $e^{-1}$  and  $M_i$  rises exponentially, and repeats. This "wave problem" arises due to the si-periodicity longer  $T_{RAREP}$ .

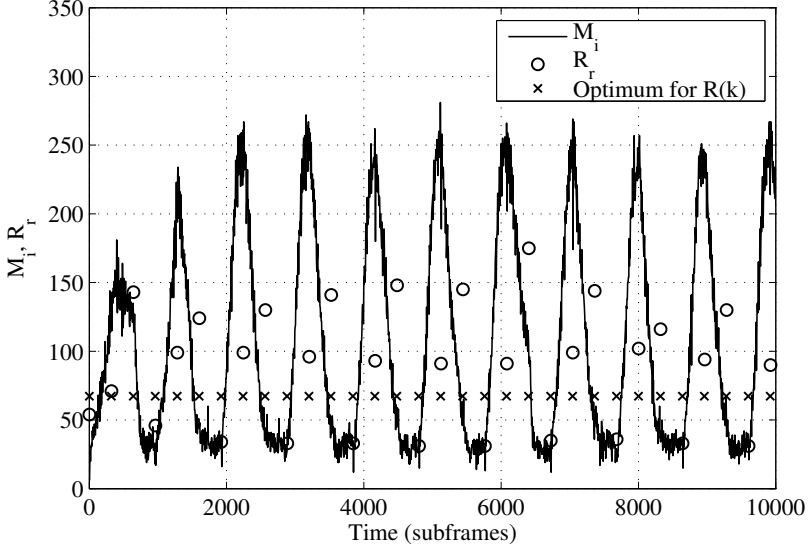
Figure 3.2 shows the simulation with  $M$  of 30,000 devices. Figure 3.3 shows the simulation with  $M$  of 50,000 devices. The optimum parameter ( $R^*$ ) for 30,000 devices is about 40, and that for 50,000 devices is about 67. In both cases, the wave problem is occurred thus the difference between  $M_i$  and  $R_r$  is occurred. Although the size of preamble can be very similar to the optimum value, the accumulated number of collided users or the random arrivals change  $M_i$  to the values far from the optimum value as shown in Figure 3.3.



**Fig. 3.1.**  $M_i$ ,  $R_r$ , and optimum for  $R_r$  with  $M$  of 100,000



**Fig. 3.2.**  $M_i$ ,  $R_r$ , and optimum for  $R_r$  with  $M$  of 30,000



**Fig. 3.3.**  $M_i$ ,  $R_r$ , and optimum for  $R_r$  with  $M$  of 50,000

Although the demonstration is given sample mean, the determination using the most recent, weighted mean, max policies also show same problem. This “fluctuation problem” arises because the si-periodicity is longer than RACH periodicity. The si-periodicity reduces the throughput of the system because the fluctuation problem makes difference between  $M_i$  and  $R_r$ . To resolve the fluctuation problem, the BS needs to determine proper  $R_1$ . Because changes in  $M_i[1]$  can also cause the problem, the BS should predict  $M_i[1]$  to determine  $R_1$ , or control  $M_i[1]$  to decrease the difference between  $M_i$  and  $R_1$ . However, it is difficult to decide these values in practical systems. Therefore, this dissertation proposes the preamble partition protocol using collision information of devices.

### 3.4. The Proposed Preamble Partition Protocol

The analysis in the previous section shows that fluctuation problem decreases the throughput. The fluctuation problem occurs when the BS selects parameters using  $\hat{M}_i$  which can differ to  $R^*$ . The analysis in previous section implies that the BS requires a value for the size of pool which is similar to  $R^*$  in every si-period to stabilize  $M_i$  and  $\hat{M}_i$ .

When the devices are grouped using  $n$ , the value for the better size of pool can be obtained. For a given threshold,  $N_{th}$ , let the deep backlogged devices be the devices that experienced backlogs larger than  $N_{th}$  times. Let  $B_N(i)$  be the number of non-deep backlogged devices, and  $B_D(i)$  be that of deep backlogged devices. They are equal to

$$B_N(i) = \sum_{n=1}^{N_{th}} M_i[n], \quad (3.21)$$

$$B_D(i) = \sum_{n=N_{th}+1}^{N_{max}} M_i[n]. \quad (3.22)$$

Let  $B_N^*(i)$  and  $B_D^*(i)$  be  $B_N(i)$  and  $B_D(i)$  in the optimum condition, i.e.  $P_{S,i} = e^{-1}, \forall i$ , respectively. They are equal to

$$B_N^*(i) = \sum_{n=1}^{N_{th}} \bar{\lambda} T_{RAREP} (1 - e^{-1})^{n-1}, \quad (3.23)$$

$$B_D^*(i) = \sum_{n=N_{th}+1}^{N_{max}} \bar{\lambda} T_{RAREP} (1 - e^{-1})^{n-1}, \quad (3.24)$$

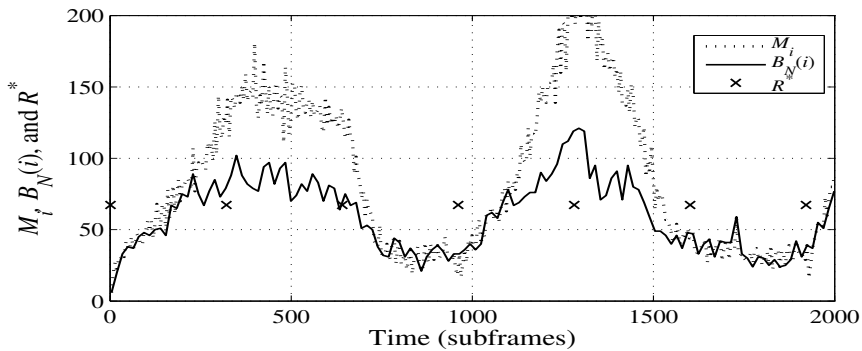
where  $\bar{\lambda}$  is the mean arrival rate of devices. In (3.23) and (3.24),  $(1 - e^{-1})^{n-1}$  decreases as  $n$  increases. Thus, we can expect that  $B_N^*(i) + B_D^*(i)$

can be close to  $B_N^*(i)$  for a sufficiently large  $N_{th}$ . As  $|P_{S,i} - e^{-1}|$  increases, both  $|B_N(i) - B_N^*(i)|$  and  $|B_D(i) - B_D^*(i)|$  increase where  $|x|$  is the absolute value of  $x$ . However, the speed of increase of  $|B_N(i) - B_N^*(i)|$  can be slower than that of  $|B_D(i) - B_D^*(i)|$  with a sufficiently small  $N_{th}$ . From the two properties, we can expect that if  $N_{th}$  is correctly selected,  $B_N(i)$  can be close to  $R^*$ , where  $R^* = B_N^*(i) + B_D^*(i)$ . Figure 3.4 shows  $M_i$ ,  $B_N(i)$ , and  $R^*$  for  $M$  of 50,000 and 100,000 with  $N_{th} = 4$ .  $B_D(i)$  is excluded in Figure 3.4 but can be obtained from the difference between  $M_i$  and  $B_N(i)$ . As shown in Figure 3.4,  $B_N(i)$  changes around  $R^*$  compared to  $M_i$ . Thus, if the BS can estimate and use  $B_N(i)$  to determine the size of the pool, the throughput can be increased because the BS can obtain and use the size of the pool close to  $R^*$ . Since  $P_{S,i}$  can differ from  $e^{-1}$  in a real system yet  $P_{S,i}$  is generally unknown, we selected  $N_{th}$  to minimize the square error between  $R^*$  and  $\sum_{n=1}^{N_{th}} \bar{\lambda} T_{RAREP} (1-p)^{n-1}$  for the arbitrary success probability  $p$  in the range of probability. The selection can be represented as

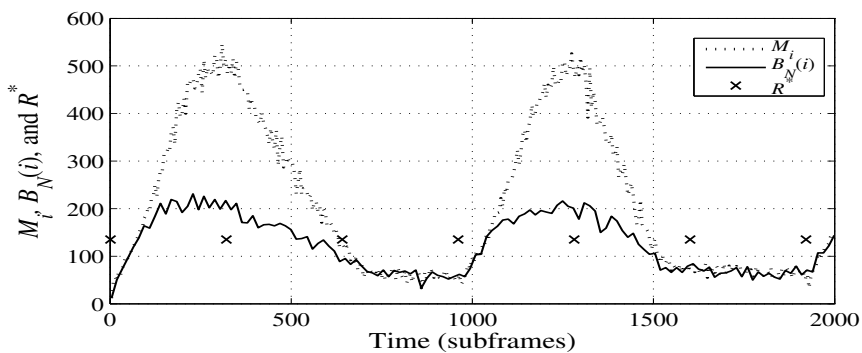
$$\arg \min_{N_{th}} \int_0^1 \bar{\lambda} T_{RAREP} \left\{ \sum_{n=1}^{N_{max}} (1 - e^{-1})^{n-1} - \sum_{n=1}^{N_{th}} (1 - p^{-1})^{n-1} \right\}^2 dp. \quad (3.25)$$

Note that  $\bar{\lambda} T_{RAREP}$  can be replaced by 1 since it can be regarded as a constant.

In conventional DARR, the BS cannot estimate  $B_N(i)$  because the BS cannot distinguish whether or not a preamble in RACH is sent by non-deep backlogged devices. To estimate and use  $B_N(i)$  in the BS, this dissertation proposes the preamble partition protocol, in which a single preamble pool is divided into two preamble pools. In the preamble par-



(a)



(b)

**Fig. 3.4.** Number of contending devices ( $M_i$ ), number of non-deep backlogged devices [ $B_N(i)$ ], and optimum value for  $R_r$  ( $R^*$ ): (a)  $M = 50,000$  and (b)  $M = 100,000$ .

tition protocol, the non-deep backlogged devices use one of the preamble pools, and the deep backlogged devices use the other preamble pool. Therefore, the BS can distinguish whether or not a preamble is sent by non-deep backlogged devices, and can estimate the number of devices in each group. The proposed protocol does not require an additional message or complex computation compared to the conventional DARR. The proposed preamble partition protocol requires additional bits in SIB2 as well as algorithms in the devices and BS.

**Preamble group selection in the devices:** In the proposed preamble partition, the device periodically receives a broadcasted message as conventional LTE-A, such as SIB2, which contains information about the preambles in the pool for group 1, those for group 2, and  $N_{th}$ . Let  $\mathbf{C}_m$  be the pool for group  $m$  ( $m=1$  or  $2$ ). Let  $R_{m,r}$  be the size of the pool in the  $r$ -th si-period for group  $m$ . Suppose that the preamble transmission of a device occurred in the  $i$ -th RACH period, and the transmission will be the  $n$ -th preamble transmission. If  $R_{2,r} = 0$ , or  $R_{2,r} > 0$  and  $n \leq N_{th}$ , the device selects a preamble in  $\mathbf{C}_1$ . Otherwise (i.e.  $R_{2,r} > 0$  and  $n > N_{th}$ ), the device selects a preamble in  $\mathbf{C}_2$ . Since the devices in which  $n > N_{th}$  are excluded in group 1 in most cases, the estimation for the number of contended devices in group 1 becomes the estimation of  $B_N(i)$ . The pseudo code for the proposed protocol used for devices is represented in Algorithm 3.1. As shown in line 2 in Algorithm 3.1, additional bits for  $\mathbf{C}_2$  and  $N_{th}$  in SIB2 are required. However, the complexity of the algorithm for the device is the same as that for the conventional LTE-A, since similar operations, except lines 8 to 12 in Algorithm 3.1, are also



required in the conventional system.

**Decision for the size of pools and notification in the BS:** The preambles in the two pools are determined and notified by the BS. In initial stage, the BS starts RA with arbitrary  $R_{1,1}$ , for example, 54 as in conventional LTE-A. The BS cannot expect the necessity of group 2, thus the BS starts RA without group 2, i.e.  $R_{2,1} = 0$ . The BS determines and broadcasts  $R_{1,r}$  and  $R_{2,r}$  in every  $T_{UPDATE}$  subframes for DARR in LTE-A. Let the number of contended devices at time slot  $i$  in group  $m$  be  $M_{m,i}$ , and its estimate be  $\hat{M}_{m,i}$ . As in DFSA, the BS can count the number of unused, succeeded, and collided preambles in each RACH period and for each group. Let the number of unused preambles for group  $m$  in  $i$ -th time slot be  $\hat{M}_{m,i}$ .  $\hat{M}_{m,i}$  is equal to

$$\hat{M}_{m,i} = f(O_{m,i}, R_{m,r}). \quad (3.26)$$

Let the sample mean of  $\hat{M}_{m,i}$  in  $r$ -th si-period be  $\bar{M}_{m,r}$ , which is equal to

$$\bar{M}_{m,r} = \frac{1}{K} \sum_{i=(r-1)K+1}^{r \cdot K} \hat{M}_{m,i}. \quad (3.27)$$

At every  $T_{UPDATE}$ , the BS determines  $R_{1,r+1}$  as

$$R_{1,r+1} = \max(\lceil \bar{M}_{1,r} \rceil, R_{\min}). \quad (3.28)$$

The BS can set as  $\lceil \bar{M}_{2,r} \rceil$  if  $R_{2,r} > 0$ . However, if  $R_{2,r} = 0$ , the BS cannot obtain  $\bar{M}_{2,r}$ . In this case, the BS needs to determine  $R_{2,r+1}$  using  $\bar{M}_{1,r}$ . Let  $\gamma$  be the expected ratio of devices contended in group 1 to that in group 2 in the system with optimum success probability ( $P_{S,i} = e^{-1}$ ),

---

**Algorithm 3.1** Algorithm for the devices with the preamble partition protocol

---

- 1: On receiving SIB2 from BS:
- 2:   obtain and update  $\mathbf{C}_1$ ,  $\mathbf{C}_2$  and  $N_{th}$ .
- 3: On receiving the request for RA procedure from upper layer:
- 4:   if  $\mathbf{C}_1$ ,  $\mathbf{C}_2$  and  $N_{th}$  are not obtained
- 5:     wait until they are obtained.
- 6:   end if
- 7:   for  $n = 1$  to  $N_{max}$
- 8:     if  $n \leq N_{th}$
- 9:       randomly selects a preamble in  $\mathbf{C}_1$ .
- 10:     else
- 11:       randomly selects a preamble in  $\mathbf{C}_2$ .
- 12:     end if
- 13:     transmit the selected preamble to BS in upcoming time slot.
- 14:     activate timer and wait for MSG2.
- 15:     if(MSG2 is arrived before timer expiration)
- 16:       perform remaining procedures including data transmission.
- 17:       if(data is successfully delivered to BS)
- 18:         return as success.
- 19:       end if
- 20:     end if
- 21:   end for

---

which can be obtained by

$$\gamma = \frac{\sum_{n=N_{th}+1}^{N_{\max}} (1 - e^{-1})^{n-1}}{\sum_{n=1}^{N_{th}} (1 - e^{-1})^{n-1}}. \quad (3.29)$$

Therefore, the BS determines the BS determines  $R_{2,r+1}$  as

$$R_{2,r+1} = \begin{cases} \lceil \gamma \bar{M}_{1,r} \rceil & ; R_{2,r} = 0, \\ \lceil \bar{M}_{2,r} \rceil & ; R_{2,r} > 0. \end{cases} \quad (3.30)$$

The algorithm for BS is represented in Algorithm 3.2

---

**Algorithm 3.2** Algorithm for the BS with the preamble partition protocol

---

- 1: On completing time slot  $i$ :
  - 2: obtain  $\hat{M}_{m,i}$  using (3.26).
  - 3: On completing  $r$ -th si-period:
  - 4: obtain  $\bar{M}_{1,r}$  and  $\bar{M}_{2,r}$  using (3.27).
  - 5:  $R_{1,r+1} = \max(\lceil \bar{M}_{1,r} \rceil, R_{min})$
  - 6: if  $R_{2,r} = 0$
  - 7:      $R_{2,r+1} = \lceil \gamma \bar{M}_{1,r} \rceil$
  - 8: else
  - 9:      $R_{2,r+1} = \lceil \bar{M}_{2,r} \rceil$
  - 10: end if
  - 11: construct the pools for each group,  $\mathbf{C}_1$  and  $\mathbf{C}_2$ .
  - 12: transmit SIB2 including information for  $\mathbf{C}_1$ ,  $\mathbf{C}_2$ , and  $N_{th}$ .
-

### 3.5. Performance Evaluation

In this section, we present simulation results of the throughput in time slot  $i$  over time and the average throughput with respect to  $M$ . The values for parameters used in simulation are summarized in Table 3.2. In the evaluation, the arrival distribution for devices is set as uniform or Beta ( $\alpha = 3$ ,  $\beta = 4$ ) distribution.  $T_a$  and  $N_{max}$  are set as 10 seconds and 10, respectively [8]. In the simulation, 1 subframe is equal to 1 milliseconds.  $T_{RAR}$ ,  $W_{RAR}$ ,  $W_{BO}$  are set as 3, 5, and 20 subframes, respectively [8].  $T_{RAREP}$  is 5 subframes.  $T_{UPDATE}$  is 160 subframes, which is an available periodicity of SIB2 [26, 37]. We use  $R_{min}$  of 10 [12], and  $R_1$ , and  $R_{1,1}$  of 54 [8]. The  $N_{th}$  is equal to 4 with the parameters. The BS estimates the number of contended devices in a RACH using the estimation function from the study by Khandelwal et al. [49] for the simulation which is equal to

$$f(\mathbf{O}_i, R_r) = \begin{cases} \min \left\{ \frac{\log\left(\frac{O_{i,0}}{R_r}\right)}{\log\left(1 - \frac{1}{R_r}\right)}, O_{i,1} + 2O_{i,2} \right\} & ; O_{i,0} > 0, \\ O_{i,1} + 2O_{i,2} & ; O_{i,0} = 0. \end{cases} \quad (3.31)$$

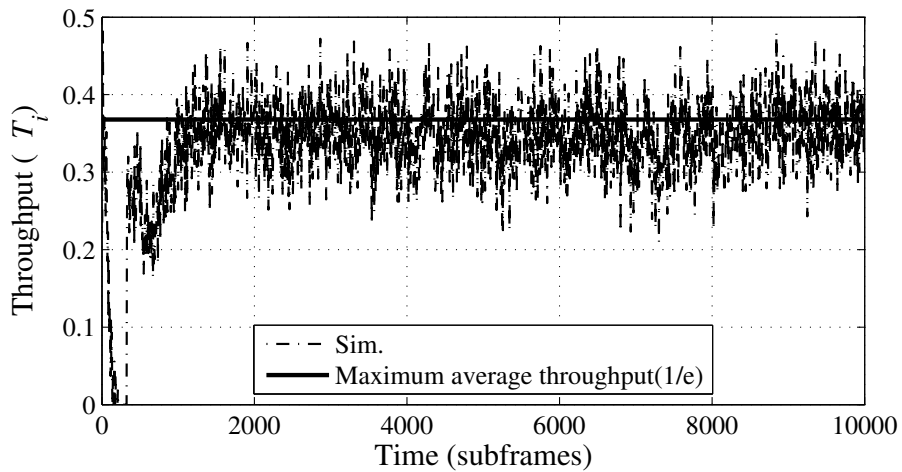
The Riverbed Modeler (known as OPNET Modeler) is used for the simulation. We have implemented the state machines and the operations in each state for the devices and the BS, which simulate RACH procedure with given system model and the parameters for evaluation. In the simulation, each device selects an arrival time according to the arrival distribution and  $I_a$ .  $M_i[1]$  and  $M_i$  changes randomly during simulation due to the arrival distribution and the probability to select a preamble

**Table 3.2.** Parameters for performance evaluation of preamble partition protocol

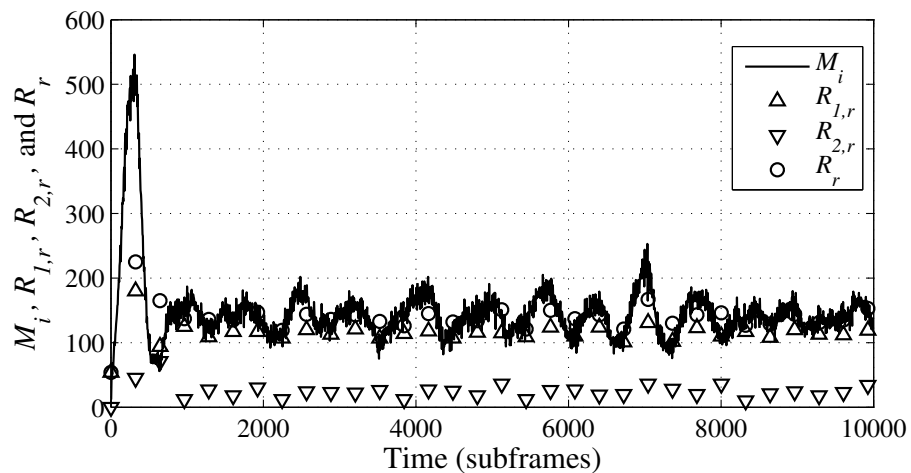
| Parameters           | Value                                       |
|----------------------|---|
| Arrival distribution | Uniform,<br>Beta( $\alpha = 3, \beta = 4$ ) |
| $T_a$                | 10 seconds                                  |
| $I_a$                | 2000 slots                                  |
| $N_{max}$            | 10  |
| $T_{RAR}$            | 3 subframes                                 |
| $W_{RAR}$            | 5 subframes                                 |
| $W_{BO}$             | 20 subframes                                |
| $T_{RAREP}$          | 5 subframes                                 |
| $T_{UPDATE}$         | 320 subframes                               |
| $R_{min}$            | 10  |
| $R_1, R_{1,1}$       | 54  |
| $R_{2,1}$            | 0   |

in device. Thus, we take 1,500 simulations for each point in Figures from 3.5 to 3.12. The number of simulations is selected to reduce the standard error to be below 0.5% of the average value of each metric. For the proposed protocol,  $P_{S,i}$  becomes

$$P_{S,i} = \begin{cases} \left(1 - \frac{1}{R_{1,r}}\right)^{M_i-1} & ; n \leq N_{th}, \\ \left(1 - \frac{1}{R_{2,r}}\right)^{M_i-1} & ; n > N_{th}. \end{cases} \quad (3.32)$$



**Fig. 3.5.** The throughput of the cell with proposed preamble partition approach for  $M$  of 100,000



**Fig. 3.6.**  $M_i$ ,  $R_{m,r}$ , and  $R_r$  for  $M$  of 100,000

Figure 3.5 shows the throughput at time slot  $i$  ( $T_i$ ) in a simulation using 100,000 devices, respectively. The uniform distribution for arrival and proposed preamble partition protocol is applied in both simulations. The throughput during the first  $T_{UPDATE}$  subframes decreases rapidly because initial size of pool,  $R_{1,1}$ , is much smaller than the number of arrivals. By calibrating the size of pool in the second and third si-period, the throughput approaches the maximum throughput.

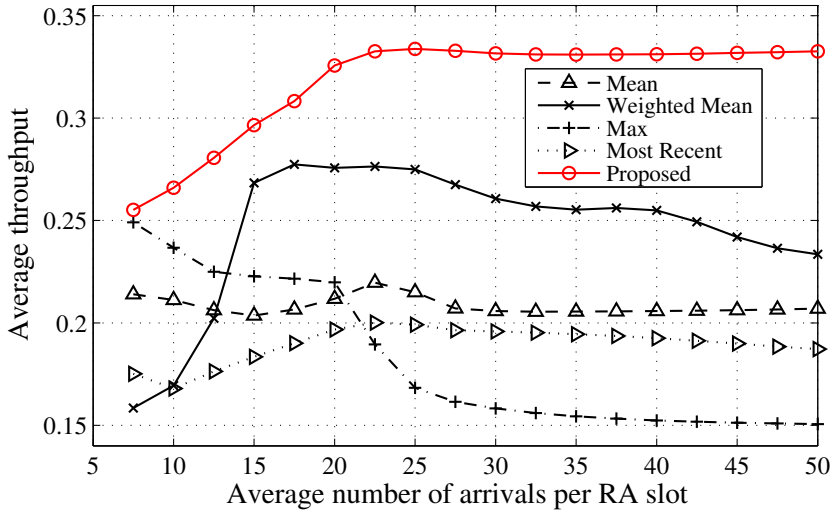
Figure 3.6 shows the values of  $M_i$ ,  $R_{m,r}$ , and  $R_r$ , where  $R_r = R_{1,r} + R_{2,r}$ , for  $M$  of 100,000, respectively. Because  $R_{1,1}$  and  $R_{2,1}$  are set to 54 and 0, respectively,  $M_i$  rapidly increases in the first 320 subframes due to the small number of preambles. By allocating  $R_{2,2}$  at 320 subframes, the deep backlogged devices are contended and estimated independently to the non-deep backlogged devices. Therefore, the BS can know the estimates of  $B_N(i)$ . After the third interval, the variation of  $M_i$  becomes small by independent selection of  $R_{1,r}$  using estimated  $B_N(i)$ . Note that the small variations after 960 subframes can be observed by random arrivals during a si-periodicity.

Figure 3.7 shows the comparison of the average throughputs for uniformly distributed arrivals with different arrival rates. The arrival rate is the number of newly arrived devices per RACH period. The average throughput in a simulation, denoted by  $\bar{U}$ , is equal to

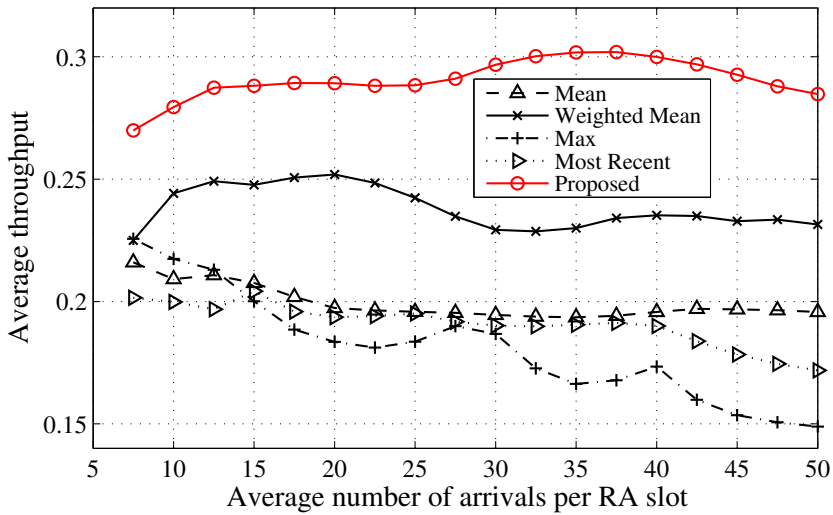
$$\bar{U} = \frac{1}{I_a} \sum_{i=1}^{I_a} U_i. \quad (3.33)$$

For the weighted mean policy, the weight is set as

$$w_i = 1 + \text{mod}(i - 1, K). \quad (3.34)$$



**Fig. 3.7.** The average throughput vs. the number of arrivals per time slot for uniformly distributed arrival



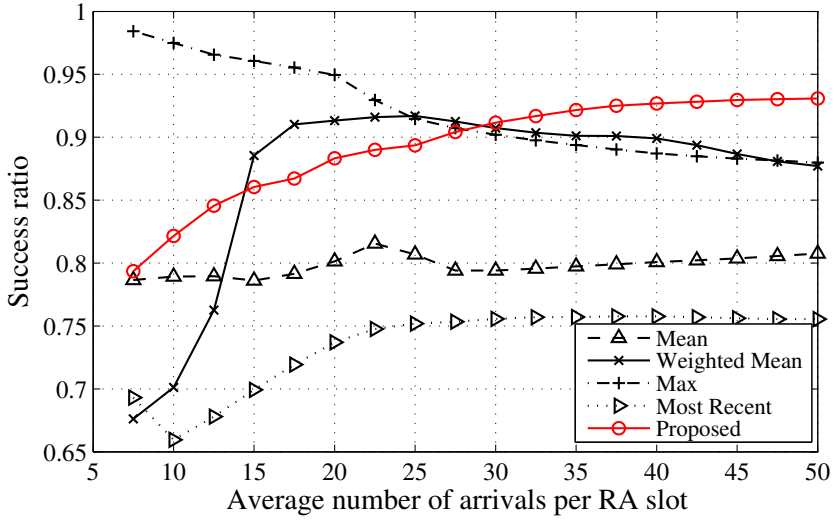
**Fig. 3.8.** The average throughput vs. the number of arrivals per time slot for Beta distributed arrival



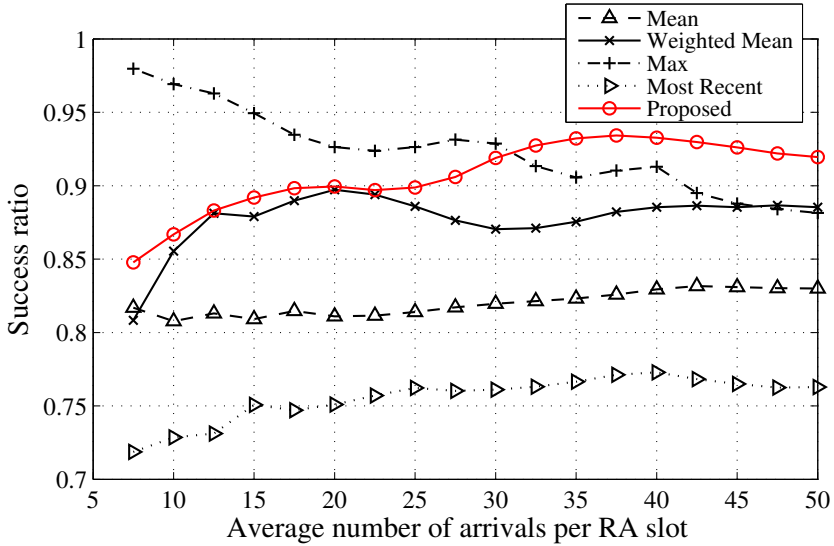
where  $\text{mod}(x, y)$  refers to the remainder of the division of  $x$  by  $y$ . The fluctuation problem decreases the throughput of the conventional policies. The preamble partition protocol reduces the fluctuation problem so the throughput of the proposed protocol is larger than that of the conventional policies. Note that the utilities for all approaches slightly increases when the average number of arrivals per RA slot is around 20, because the average number of devices contending in RACH period is similar to the initial size of the preamble pool.

Figure 3.8 shows the average throughput for the Beta-distributed arrivals. The preamble partition protocol also shows better throughput for the Beta distribution, where the arrival rates change with the  $\cap$ -shaped curves. The throughput is increased by approximately from 29.7% to 114.4% and from 23.0% to 91.3% compared with sample mean policy for the uniform and Beta distributed arrivals, respectively, when the arrival rate is equal to 50 devices per RACH period (corresponding to  $M$  of 100,000 devices).

Figures 3.9 and 3.10 show the success ratio for the uniform and Beta distributed arrival, respectively. The success ratio is equal to the ratio of the sum of  $S_i$  for all  $i$  to  $M$ . As in Figure 3.1, the fluctuation of  $M_i$  generally has two types of interval in turn, as denoted by  $W_1$  and  $W_2$ , where  $M_i > R^*$  during  $W_1$  and the  $M_i < R^*$  during  $W_2$ . If the duration of  $W_1$  increases or  $|M_i - R^*|$  increases in  $W_1$ , the sum of  $S_i$  generally decreases. If the duration of  $W_2$  increases or  $|M_i - R^*|$  increases in  $W_2$ , the sum of  $S_i$  generally increases. The max policy allocates excessively large pool sizes thus the mean duration of  $W_2$  increases. However, the success



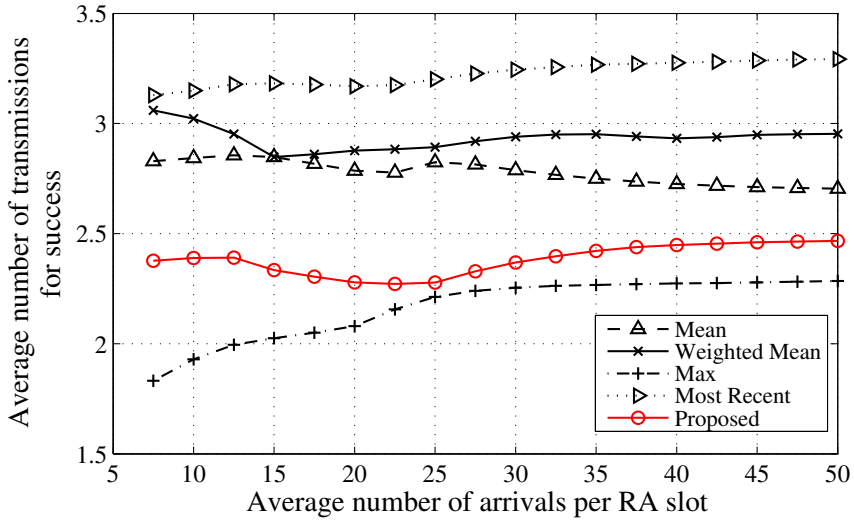
**Fig. 3.9.** Success ratio vs. arrival rate per RACH period for uniformly distributed arrival



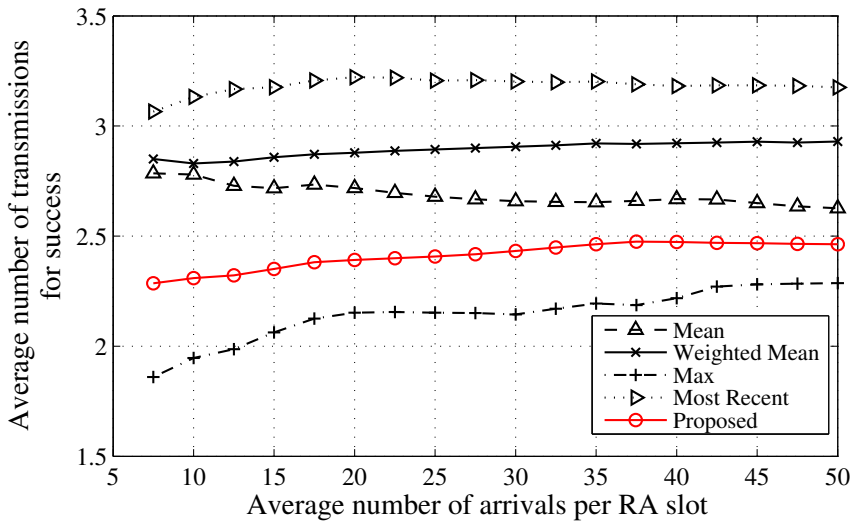
**Fig. 3.10.** Success ratio vs. arrival rate per RACH period for Beta distributed arrival

ratio decreases as the arrival rate increases since  $|M_i - R^*|$  increases in  $W_1$ . The average pool sizes of other three conventional policies are similar and smaller than that of max policy but their fluctuation patterns are different. The most recent policy selects very small pool size in  $W_2$ , which causes long duration of  $W_1$  and the increase of  $|M_i - R^*|$  in  $W_1$  compared with mean policy. The mean policy selects small pool size in the middle of  $W_1$  thus the time to enter  $W_2$  increases. The weighted mean policy selects sufficiently large pool size in  $W_1$  and not selects very low pool size in  $W_2$ . However, fluctuation is still observable. In addition, when arrival rate is small, it shows similar pattern to the most recent policy. The proposed protocol shows high success ratio by decrease  $M_i$  in  $W_1$  close to  $R^*$  even the average pool size shows smallest value compared with other policies.

Figures 3.11 and 3.12 show the average number of preamble transmissions for success with the uniform and Beta distributed arrivals, respectively. The high value implies a high number of collisions and long delay to success. The max policy shows the lowest values due to the low number of collisions from the large size of the preamble pool. The proposed protocol shows lower values than the other three conventional policies. The lower values imply the lower number of collisions than the other three conventional approaches.



**Fig. 3.11.** Average number of preamble transmissions for success vs. arrival rate per RACH period for uniformly distributed arrival



**Fig. 3.12.** Average number of preamble transmissions for success vs. arrival rate per RACH period for Beta distributed arrival

### 3.6. Summary

In this chapter, the challenge of the throughput degradation for DARR in LTE-A due to si-periodicity was discussed. To resolve the fluctuation problem, this dissertation proposed a preamble partition protocol for LTE-A. The proposed preamble partition protocol increases throughput compared to the system without the preamble partition protocol. The proposed preamble partition protocol can be used to improve the throughput of the RACH procedure in LTE-A, or RA schemes based on a frame slotted or multi-channel ALOHA.

Although the throughput degradation in DARR is analyzed and resolved, a similar problem is expected in the DFSA for other communication systems or for the ACB in LTE-A, because they also use an estimated number of contended devices. To increase the effectiveness of DFSA or ACB, studies on the throughput degradation from si-periodicity or similar periodicity need to be carried out.

# Chapter 4. Preamble Partition based Adaptive DARR and ACB Protocol

## 4.1. Introduction

The ACB is one of the congestion control methods to overcome traffic overload from massive number of devices [4]. The ACB intentionally defers some of devices to prevent traffic congestion. The ACB is useful when the resource is limited but the traffic is overloaded. Since the LTE-A operates in the licensed band and with a coverage, the number of preambles and the bandwidths can be limited. Therefore, the ACB is one of hot research issues in the expected solutions to overcome traffic congestion. In addition to the ACB, the BS should allocate the resources for IoT devices to not exceed their demands although the resource is sufficient since preambles and bandwidths can be shared with H2H communication users. As represented in previous chapter, DARR covers the resource allocation. Therefore, an adaptive DARR and ACB protocol is required to efficiently allocate resource and to overcome traffic overload with the limited resources.

Since the mobile networks has an interval to announce the resource

allocation and ACB factor as represented in 2.1, an adaptive DARR and ACB protocol requires to prevent the throughput degradation. However, the previous researches for ACB were rarely considered an interval to notify a message which includes resource allocation results and ACB factor.

In this chapter, the discussion about the changes in the throughput when an interval exists which is used to announce resource allocation results and ACB factor. In addition, a preamble partition and stochastic gradient descent approach based DARR and ACB protocol is presented to determine both the size of the pool and ACB factor adaptively. The preamble partition approach adaptively separates a pool into two pools or merges two pools into a pool to determine both the pool size and the ACB factor. The stochastic gradient descent approach is used to find optimal size of pool. The performance evaluation using simulations is represented where the results show that the proposed DARR and ACB protocol approach can achieve performance that is closer to that from ideal case algorithm than a conventional DARR and ACB protocol.

## 4.2. System Model

Suppose that a cell consists of a BS and  $M$  MTC devices. These devices are activated not simultaneously but with in a short period of time,  $T_a$ , which is referred as the "activation time" in this chapter [4]. Each devices are activated at time  $t$  with a probability density function of  $g(t)$ . Let  $I_a$  be the number of RACHs within the activation time. The

activation time is divided into  $I_a$  discrete slots, where each slot has longer duration than a RACH. The  $i$ -th slot begins with  $i$ -th RACH, where  $i = 1, 2, \dots, I_a, \dots$ . The length of each time slot is equal to the interval between two consecutive RACHs, where this interval is also referred as RACH period in this chapter. Let  $T_{RAREP}$  be the length of each time slot. The  $i$ -th time slot begins at time  $t_{i-1}$  and ends at time  $t_i$ . Let  $\lambda_i$  be the number of activations during time slot  $i$ , which is equal to

$$\lambda_i = \left\lceil M \int_{t_{i-1}}^{t_i} g(t) dt \right\rceil, \quad i = 1, 2, \dots \quad (4.1)$$

Let  $K$  be the number of time slots during a si-period. The  $r$ -th si-period will include time slots from  $((r-1)K+1)$ -th time slot to  $(rK)$ -th time slot. The  $i$ -th time slot ( $i=1,2,\dots$ ) is included in the  $r$ -th si-period ( $r=1,2,\dots$ ) with a relation given as

$$r = 1 + \lfloor (i-1)/K \rfloor. \quad (4.2)$$

Let the SIB2 for  $r$ -th period be changed and broadcasted before the start of  $(r-1)K+1$ -th time slot. Let  $T_{UPDATE}$  be the si-periodicity shown in Figure 2.1, which is the length of each si-period.  $K$  becomes

$$K = \lfloor T_{UPDATE}/T_{RAREP} \rfloor, \quad (4.3)$$

where  $\lfloor x \rfloor$  is the largest integer equal or smaller than  $x$ .

The SIB2 includes information for preamble pool and the access class barring (ACB). Suppose that there are  $R_r$  preambles in the pool for the  $r$ -th si-period, where the pool is given as  $\mathbf{C}_r = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{R_r}\}$ . The minimum size of pool is given as  $R_{\min}$ , i.e.,  $R_r \geq R_{\min}$ . In this



chapter, this dissertation assumes that the size of the preamble pool can be variable with upper limit of  $R_{\max}$ . For the ACB, the SIB2 includes the ACB factor  $p_r$  for  $r$ -th si-period.

In this chapter, an MTC device can transmit preambles until it succeeds the transmission of data [8]. This dissertation also assumes that the preamble is always detectable in the BS [4],[12]. A collision always occurs when two or more devices select the same preamble in the same RACH period [4]. Suppose that the transmissions of MSG3, MSG4, and data are always successful, since hybrid automatic repeat request (HARQ) provides very high transmission probability for MSG3 and MSG4 [8]. Thus, the performance of the RACH procedure thus depends on the collision probability.

Assume that  $M_i$  devices are activated and but have not finished their random access before the start of  $i$ -th time slot. An activated device chooses a random variable which is uniformly distributed between 0 and 1. If the selected random variable is larger than  $p_r$ , the device defers the random access trial to the next time slot, and repeats the selection of the random variable. Otherwise, the device transmits a preamble through a RACH, where the preamble is randomly chosen from the preamble pool.

Let  $T_{RAR}$  be the waiting time before the start of the random access response window, let  $W_{RAR}$  be the size of the RAR window, and let  $W_{BO}$  be the size of the backoff window. The devices failed the random access in from  $(i - k_2)$ -th RACH period to  $(i - k_1)$ -th RACH period will backoff

to  $i$ -th RACH period, where  $k_1$  and  $k_2$  are obtained by

$$k_1 = \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RAREP}} \right\rceil, \quad (4.4)$$

$$k_2 = \left\lceil \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RAREP}} \right\rceil. \quad (4.5)$$

Let  $N_i$  be the number of devices that passed the ACB check.  $N_i$  devices transmit their preambles randomly chosen from  $\mathbf{C}_r$  in  $i$ -th time slot. The statistical expectation of  $N_i$  for given  $M_i$  and  $p_r$  is equal to

$$\mathbb{E}[N_i | M_i, p_r] = p_r M_i. \quad (4.6)$$

In a practical system, a BS cannot know  $M_i$  and  $N_i$  since it cannot obtain  $\lambda_i$  without explicit signaling. However, the BS can estimate  $M_i$  and  $N_i$ . Let  $\hat{M}_i$  and  $\hat{N}_i$  be the estimate for  $M_i$  and  $N_i$ , respectively. Let  $\mathbf{O}_i = [O_{i,0}, O_{i,1}, O_{i,2}]$  be the observation vector for the preambles allocated for the  $i$ -th RACH period, where  $O_{i,0}$ ,  $O_{i,1}$ , and  $O_{i,2}$  are the observed number of unused, successful, and collided preambles, respectively. The BS can use an estimation function of  $\mathbf{O}_i$  and  $R_r$ ,  $f(\mathbf{O}_i, R_r)$ , to obtain  $\hat{N}_i$  after completing RA in the  $i$ -th RACH period, which can be represented as

$$\hat{N}_i = f(\mathbf{O}_i, R_r). \quad (4.7)$$

Once the BS obtained  $\hat{N}_i$ , the BS can estimate  $M_i$  from  $\hat{N}_i$  and  $p_r$ . The estimation can be represented as

$$\hat{M}_i = \frac{\hat{N}_i}{p_r}. \quad (4.8)$$

In this study, the BS attempts to increase the RACH throughput. Let  $S_i$  be the number of devices that successfully transmit their preamble in

the  $i$ -th RACH period. Let  $U_i$  be the RACH throughput for each  $i$ -th RACH period. The RACH throughput is the ratio of the number of successful preambles in a RACH period to the allocated number of preambles in the RACH period, which is equal to

$$U_i = S_i/R_r. \quad (4.9)$$

### 4.3. Problem Definition: Needs of adaptive MAC protocol considering si-periodicity

In the standard of LTE-A, the maximum number of preambles per RACH is fixed. When the number of M2M devices contending in a RACH exceeds the number of preambles per RACH, the congestion will occur. The access class barring (ACB), sometimes referred as access barring (AB), is one of hot issues in the random access of LTE-A. The ACB defers some access of devices without changing the number of preambles.

When the number of M2M devices does not exceed the number of preambles, the excessive allocation of preambles for M2M devices can take the opportunity of random access of the H2H devices or other devices those have high priority than M2M devices. Thus, the DARR is required in BS in addition to the ACB, since the traffic load is continuously changes over time. Therefore, the adaptive MAC protocol combining DARR and ACB is required. The adaptive MAC protocol also needs to consider the interval to update the resource allocation and ACB factor for mobile network similar to the discussion in the previous chapter.

### 4.3.1 Background for the adaptive MAC protocol with DARR and ACB

Without considering the si-periodicity, the optimal pool size and ACB factor are already studied. Remind that the DARR selects the pool size equal to the expected number of contending devices in  $(i + 1)$ -th time slot to maximize throughput for  $K = 1$ . With the system model for this chapter, the maximum pool size is limited by  $R_{\max}$ . In addition,  $S_{i+1}$  will increase as  $R_{r+1}$  increases. Therefore, the BS needs to select the pool size for  $K = 1$  as

$$R_{r+1} = \min[R_{\max}, M_i]. \quad (4.10)$$

The ACB controls the number of contending devices in a RACH by announcing ACB factor when the expected number of contending devices exceeds the number of preambles. The throughput of random access is maximized when  $N_{i+1} = R_{r+1}$ , and the expectation for  $N_{i+1}$  needs to be equal to  $R_{r+1}$ . Therefore, the BS needs to find  $p_{r+1}$  such as

$$\mathbb{E}[N_{i+1}|M_{i+1}, p_{r+1}] = p_{r+1}M_{i+1} = R_{r+1}. \quad (4.11)$$

However,  $M_{i+1}$  is a random variable and unknown for BS.  $\hat{M}_{i+1}$  is also unknown before the completion of  $(i + 1)$ -th RACH. As described in previous chapter,  $M_{i+1}$  includes the number of devices backoffed from previous time slots, and  $\lambda_i$  and  $\lambda_{i+1}$  are correlated according to the arrival distribution for the activation of devices [8, 14]. Therefore, the BS selects  $p_{r+1}$  for  $K = 1$  given  $M_i$  and  $R_r$  as

$$p_{r+1} = \frac{M_i}{R_r} = \frac{\mathbb{E}[N_i|M_i, p_r]}{p_r R_r}. \quad (4.12)$$

---

**Algorithm 4.1** Conventional algorithm to select pool size and ACB factor for  $K = 1$

---

- 1: On completing last RACH period in  $r$ -th si-period ( $i = r$ ):
  - 2: obtain  $\mathbf{O}_i = [O_i(0), O_i(1), O_i(2)]$
  - 3: obtain  $\hat{N}_i = f(\mathbf{O}_i, R_r)$
  - 4: obtain  $\hat{M}_i = \frac{\hat{N}_i}{p_r}$
  - 5:  $R_{r+1} = \max \left\{ R_{\min}, \min \left( \hat{M}_i, R_{\max} \right) \right\}$
  - 6:  $p_{r+1} = \min \left( 1, \frac{R_{r+1}}{\hat{M}_i} \right)$
- 

The conventional algorithms select ACB factor and pool size for  $(i+1)$ -th ( $= (r+1)$ -th for  $K = 1$ ) RACH period based on  $\hat{M}_i$  and  $\hat{N}_i$ , where this approach can be simplified as the algorithm in Algorithm 4.1.

### 4.3.2 Throughput degradation in DARR due to the resource allocation update interval

As the analysis in the previous chapter, the si-periodicity causes the fluctuation problem in the DARR. The fluctuation problem causes the degradation of throughput for DARR. When the si-periodicity is given, which means that  $K > 1$ , the BS needs to select  $R_{r+1}$  and  $p_{r+1}$  after passing multiple time slots. Before construction of SIB2, BS will have  $\hat{N}_{(r-1)K+1}, \dots, \hat{N}_{rK}$ , and  $\hat{M}_{(r-1)K+1}, \dots, \hat{M}_{rK}$ . If the BS uses the basic algorithm, the BS will select  $R_{r+1}$  and  $p_{r+1}$  based on  $\hat{M}_{rK}$  and  $\hat{N}_{rK}$ .

If the network is operating with  $R_r^* < R_{\max}$ , the BS will use DARR. If the basic algorithm is used to select  $R_r$ , the oscillation of  $M_i$  will occur. Remind that  $\mathbb{E}[S_i | M_i, p_r, R_r]$  is maximized when  $N_i = R_r$  for given  $R_r$ .

where  $N_i = M_i$  in this case. Let  $T_{RAR} = W_{RAR} = W_{BO} = 0$ , and let devices arrive with uniformly distributed arrival with  $\mathbb{E}[\lambda_i] = \bar{\lambda}$ . In this case,  $\mathbb{E}[S_i|M_i, p_r, R_r]$  needs to be equal to  $\bar{\lambda}$  to maximize throughput. If  $R_1$  is selected as  $R_1 < R_1^*$ ,  $\mathbb{E}[S_i|M_i, p_r, R_r] < \bar{\lambda}$ , and  $M_i$  will increase in first si-period. The conventional algorithm will select  $R_2$  as  $R_2 > R_2^*$ . In second si-period,  $\mathbb{E}[S_i|M_i, p_r, R_r] > \bar{\lambda}$ , thus  $M_i$  will continuously decrease. The conventional algorithm will select  $R_3$  as  $R_3 < R_3^*$ , which returns to the situation in first si-period, and continues until the activations of devices stop. The random arrivals of  $\lambda_i$  also increase the probability to select improper  $R_r$ . This fluctuation of  $M_i$  must be resolved to increase the throughput. Thus, the algorithm combining DARR and ACB needs to consider the fluctuation problem.

### 4.3.3 Throughput of ACB with ACB factor update interval

If  $R_r^* = R_{\max}$ , ACB will be used to reduce congestion with  $R_r = R_{\max}$ . If  $M_i > R_r$  continuously,  $M_i$  increases gradually as  $i$  increases. In this case,  $\frac{R_r}{M_i}$  will decrease continuously.  $|p_r - \frac{R_r}{M_i}|$  also decreases continuously which means the error between selected ACB factor and optimal ACB factor. Thus, the  $p_r$  selected by conventional approach will gradually converge to the optimal ACB factor, and the throughput also converges to optimal throughput.

## 4.4. Proposed dynamic resource allocation and congestion control protocol

In this section, this dissertation proposes a dynamic congestion control algorithm. The proposed algorithm selects the pool size using preamble partition approach. In addition, the stochastic gradient descent method is used to search the better pool size for statistical system.

### 4.4.1 Background for device grouping

For the description, let  $T_{RAR} = W_{RAR} = W_{BO} = 0$ . Let  $P_i$  be the preamble success probability that device can successfully transmit its preamble at time slot  $i$  given that  $p_r = 1$ , where

$$P_i = \left(1 - \frac{1}{R_{\lfloor (i-1)/K \rfloor + 1}}\right)^{(M_i-1)}. \quad (4.13)$$

The activated device backoffs until success. Thus the expectation of  $M_i$  becomes

$$\mathbb{E}[M_i] = \lambda_i + \sum_{n=1}^{i-1} \lambda_{i-n} \prod_{x=1}^n (1 - P_{i-x}). \quad (4.14)$$

Let  $\mathbb{E}^*[M_i]$  be the expectation of  $M_i$  when throughput is continuously optimized. If the network is continuously optimized in the throughput, i.e.  $P_i \simeq e^{-1}, \forall i$ ,  $\mathbb{E}^*[M_i]$  is equal to

$$\mathbb{E}^*[M_i] = \lambda_i + \sum_{n=1}^{i-1} (1 - e^{-1})^n \lambda_{i-n}. \quad (4.15)$$

If  $\mathbb{E}[M_i]$  is similar to  $\mathbb{E}^*[M_i]$ , the decision for  $R_{r+1}$  by  $\hat{M}_i$  will be good selection. With  $K = 1$ ,  $\hat{M}_{i+1}$  and  $\hat{M}_i$  can be similar because  $P_i$  can be

controlled to be close to  $e^{-1}$ . However, when  $K > 1$ ,  $\mathbb{E}[M_i]$  can have large difference with  $\mathbb{E}^*[M_i]$ .  $(1 - e^{-1})^n$  converges to 0 as  $n$  increases but  $\prod_{x=1}^n (1 - P_{i-x})$  is not due to the fluctuation problem. In this case, referring  $\hat{M}_i$  will be bad selection for  $R_{r+1}$ .

If the BS can have observed values which is close to  $\mathbb{E}^*[M_i]$  and can use for the selection of pool size for DARR, the oscillation of  $M_i$  can be decreased. Let  $B^*(i)$  be the expected number of devices at time slot  $i$  which has the number of backlogs equal or less than  $N_{th}$  and when throughput is continuously optimized.  $N_{th}$  is a threshold for the number of backlogs.  $B_N^*(i)$  is equal to

$$B^*(i) = \lambda_i + \sum_{n=1}^{N_{th}} (1 - e^{-1})^n \lambda_{i-n}. \quad (4.16)$$

Let  $B(i)$  be the expected number of devices which has the number of backlogs less than  $N_{th}$  at time slot  $i$ .  $B(i)$  is equal to

$$B(i) = \lambda_i + \sum_{n=1}^{N_{th}} \lambda_{i-n} \prod_{x=1}^n (1 - P_{i-x}). \quad (4.17)$$

If  $N_{th}$  is sufficiently large,  $B^*(i) \simeq \mathbb{E}^*[M_i]$ . If  $N_{th}$  is sufficiently small,  $(1 - e^{-1})^n$  has large value, thus the error between  $B_N(i)$  and  $B_N^*(i)$  will decrease compared to the error between  $\mathbb{E}[M_i]$  and  $\mathbb{E}^*[M_i]$ . Thus,  $B_N(i)$  with a selected  $N_{th}$  can provide an observation which is close to  $\mathbb{E}^*[M_i]$ .

#### 4.4.2 Background for the optimization with stochastic process

The stochastic gradient descent (SGD) method is a stochastic approximation of the gradient descent optimization method [66]. The gradient



descent method gradually finds the optimal parameters using the difference between current parameter and the gradient [67]. The SGD method is useful to find optimal input with low computational cost, which is enabled by the iterative computation [12]. Since the objective in the system is to maximize the throughput, the BS needs to find  $R_r$  and  $p_r$  which maximize  $\mathbb{E}[U_i|M_i, p_r, R_r]$ . For the SGD method, the derivative of  $\mathbb{E}[U_i|M_i, p_r, R_r]$  with respect to  $R_r$  is required. Let the  $\nabla_i(R_r)$  be the derivative of  $\mathbb{E}[U_i|M_i, p_r, R_r]$ .  $\nabla_i(R_r)$  is equal to

$$\begin{aligned}\nabla_i(R_r) &= \frac{d\mathbb{E}[U_i|M_i, p_r, R_r]}{dR_r} \\ &= \frac{N_i(N_i - R_r)}{R_r^3} \left(1 - \frac{1}{R_r}\right)^{N_i - 2}.\end{aligned}\quad (4.18)$$

Since  $\nabla_i(R_r)$  is small for large  $R_r$  and  $N_i$ , the convergence rate of the SGD method can be slow. To increase the convergence rate, the second order SGD method can be used [66]. For the second order SGD, the second order derivative is required. Let the  $\nabla_i^2(R_r)$  be the second order derivative of  $\mathbb{E}[U_i|M_i, p_r, R_r]$ .  $\nabla_i^2(R_r)$  is equal to

$$\begin{aligned}\nabla_i^2(R_r) &= \frac{d^2\mathbb{E}[U_i|M_i, p_r, R_r]}{dR_r^2} \\ &= \frac{N_i}{R_r^5} \left(1 - \frac{1}{R_r}\right)^{N_i - 3} \\ &\quad \times [R_r(1 - R_r) - (N_i - R_r)(R_r - 3N_i + 1)].\end{aligned}\quad (4.19)$$

When  $R_r \simeq N_i$ ,

$$\nabla_i^2(R_r) \simeq H_0 = -\frac{N_i}{R_r^3} \left(1 - \frac{1}{R_r}\right)^{N_i - 2}.\quad (4.20)$$

Let  $\bar{\nabla}_i(R_r)$  be the ratio of the first order derivative to the approximated second order derivative.  $\bar{\nabla}_i(R_r)$  is equal to

$$\bar{\nabla}_i(R_r) = \frac{\nabla_i(R_r)}{H_0} = R_r - N_i.\quad (4.21)$$

The solution by gradient method can be the input for local maximum when multiple candidates for inputs exist. Thus, we need to check whether the solution from gradient method can result global maximum or not. First, if  $N_i = 0$ , throughput is always zero regardless to  $R_r$ . Second, if  $N_i = 1$ , the maximum throughput becomes 1 with  $R_r = 1$  and decreases as  $R_r$  increases. Third, when  $N_i \geq 2$ ,  $R_r = N_i$  be the sole input for maximum or minimum output from  $\nabla_i(R_r) = 0$  in  $0 \leq N_i, R_r < \infty$ . In third case,  $R_r = N_i$  be the inputs for global maximum or global minimum. To check whether the input  $R_r = N_i$  derives global maximum or not, we need to check second order gradient at  $R_r = N_i$ . (4.19) with  $R_r = N_i$  is equal to

$$\nabla_i^2(R_r = N_i) = \frac{1}{N_i^4} \left(1 - \frac{1}{N_i}\right)^{N_i-3} \{N_i(1 - N_i)\} < 0, \quad (4.22)$$

for  $R_r \geq 1$  and  $N_i \geq 2$ . Since  $\nabla_i^2(R_r = N_i)$  is always lower than zero, the output with  $R_r = N_i$  becomes the global maximum with  $N_i \geq 2$ . From the analysis on three cases, the gradient method will find input for global maximum output in the valid ranges for  $R_r$  and  $N_i$ .

For the second order SGD method, the BS internally estimates the value for  $R_r$ . Let  $\hat{R}_r$  be the estimate for  $R_r$ . With the step size  $\alpha_{dg}$ , the BS can estimate the value for  $R_{r+1}$  using the following iterative method:

$$\begin{aligned} \hat{R}_{r+1} &= \max[R_{\min}, \hat{R}_r - \alpha_{dg} \bar{\nabla}_i(\hat{R}_r)] \\ &= \max[R_{\min}, \hat{R}_r - \alpha_{dg}(\hat{R}_r - N_i)]. \end{aligned} \quad (4.23)$$

The BS then can broadcast  $R_{r+1} = \lceil \hat{R}_{r+1} \rceil$  to all devices. Note that the flooring, ceiling, or rounding operation should not be used for (4.23) otherwise the error due to small error will be accumulated.

### 4.4.3 Preamble partition based adaptive resource allocation and congestion control protocol

Based on two backgrounds and the analysis on throughputs, this dissertation proposes the preamble partition based adaptive resource allocation and congestion control protocol. This protocol includes the algorithm for BS and the algorithm for devices. To know the estimation of  $B(i)$  in BS, the BS requires the observation vector to estimate the number of devices which have the number of backlogs less than  $N_{th}$ . This dissertation proposes the device grouping method for the estimation of  $B(i)$ . The BS allocates two preamble pools before the start of  $r$ -th si-period, where the size of first pool is  $R_{1,r}$  and that of second pool is  $R_{2,r}$ . The BS also calculates two ACB factors, where the ACB factor for first pool is  $p_{1,r}$  and that for second pool is  $p_{2,r}$ . The BS broadcasts the information of preamble pools and the ACB factors for each group before the start of  $r$ -th si-period.

Let a device experienced backlogs  $n$  times before the start of  $i$ -th time slot. In  $i$ -th time slot, the device does ACB check using  $p_{1,r}$  and select a preamble from the first pool if  $n < N_{th}$  or  $R_{2,r} = 0$ . The device is regarded as in the first group in  $i$ -th time slot. Otherwise, i.e. if  $n \geq N_{th}$  and  $R_{2,r} > 0$ , the device does ACB check using  $p_{2,r}$  and selects a preamble from the second pool. The device is regarded as in the second group in  $i$ -th time slot.

Let  $\mathbf{O}_{1,i}$  and  $\mathbf{O}_{2,i}$  be the observation vectors for each group in  $i$ -th time slot, respectively. Let  $\hat{N}_{1,i}$  and  $\hat{N}_{2,i}$  be estimation of the number of

contending devices for each group in  $i$ -th time slot, respectively. For the group index  $m = 1$  and  $2$ ,

$$\hat{N}_{m,i} = f(\hat{\mathbf{O}}_{m,i}, R_{m,r}). \quad (4.24)$$

Note that  $\hat{N}_{1,i} = \hat{B}(i)$ , where  $\hat{B}(i)$  is the estimation of  $B(i)$ . Let  $\hat{M}_{1,i}$  and  $\hat{M}_{2,i}$  be estimation of the number of activated devices in each group in  $i$ -th time slot, respectively, which can be obtained by

$$\hat{M}_{m,i} = \hat{N}_{m,i}/p_{m,r}. \quad (4.25)$$

Algorithm 4.2 shows the preamble partition based DARR and ACB algorithm for BS. The BS estimates the pool size of each group based on the SGD method as the lines from 5 to 10, where  $\hat{R}_{m,r}$  denotes the estimates for the pool size of  $m$ -th group. If one of the estimated pool size exceeds the maximum pool size, which means that  $R_{\max}$  is not sufficient to serve activated devices, the BS uses single pool with ACB as the lines from 12 to 14. If the sum of the estimated pool size exceeds the maximum pool size but each pool size not exceeds, the BS uses ACB for first group and DARR for second group to reduce access delay as lines from 15 to 17. Otherwise, the BS uses DARR for both group as the lines from 20 to 26 to increase resource efficiency.

Algorithm 4.3 shows the algorithm for UE. The UE selects a preamble in second group if the pool for second group is given by BS and the number of backoff is larger than a threshold  $N_{th}$ . Otherwise, the UE selects a preamble in first group.

---

**Algorithm 4.2** The proposed DARR and ACB algorithm for BS

---

- 1: On completing last RACH period in  $r$ -th si-period ( $i = rK$ ):
  - 2: obtain  $\mathbf{O}_{m,i}, m = 1, 2$
  - 3: obtain  $\hat{N}_{m,i} = f(\mathbf{O}_{m,i}, R_{m,r})$
  - 4: obtain  $\hat{M}_{m,i} = \frac{\hat{N}_{m,i}}{p_{m,r}}$
  - 5:  $\hat{R}_{1,r+1} = \max[0, \hat{R}_{1,r} + \alpha_{dg}\{\min(R_{\max}, \hat{M}_{m,i}) - \hat{R}_{1,r}\}]$
  - 6: if ( $\hat{M}_{2,r} > 0$ )
  - 7:      $\hat{R}_{2,r+1} = \max[0, \hat{R}_{2,r} + \alpha_{dg}\{\min(R_{\max}, \hat{M}_{m,i}) - \hat{R}_{2,r}\}]$
  - 8: else
  - 9:      $\hat{R}_{2,r+1} = 0.2\hat{R}_{1,r+1}$
  - 10: end if
  - 11: if( $\lceil \hat{R}_{1,r+1} + \hat{R}_{2,r+1} \rceil \geq R_{\max}$ )
  - 12:     if( $\lceil \hat{R}_{1,r+1} \rceil \geq R_{\max}$  or  $\lceil \hat{R}_{2,r+1} \rceil \geq R_{\max}$ )
  - 13:          $R_{1,r+1} = R_{\max}, R_{2,r+1} = 0$
  - 14:          $\hat{p}_{1,r+1} = \min\left[1, \frac{R_{\max}}{\hat{M}_{1,i} + \hat{M}_{2,i}}\right]$
  - 15:     else
  - 16:          $R_{1,r+1} = R_{\max} - \lceil \hat{R}_{2,r+1} \rceil, R_{2,r+1} = \lceil \hat{R}_{2,r+1} \rceil$
  - 17:          $\hat{p}_{1,r+1} = \min\left[1, \frac{\hat{R}_{1,r+1}}{\hat{M}_{1,i}}\right]$
  - 18:     end if
  - 19:      $\hat{p}_{2,r+1} = 1$
  - 20: else
  - 21:      $R_{1,r+1} = \lceil \hat{R}_{1,r+1} \rceil, R_{2,r+1} = \lceil \hat{R}_{2,r+1} \rceil$
  - 22:      $p_{1,r+1} = 1, p_{2,r+1} = 1$
  - 23:     if ( $R_{1,r+1} + R_{2,r+1} < R_{\min}$ )
  - 24:          $R_{1,r+1} = R_{\min}$
  - 25:     end if
  - 26: end if
-

---

**Algorithm 4.3** The proposed DARR and ACB algorithm for devices

---

- 1: On before preamble transmission
  - 2: if ( $R_{2,r} > 0$  and  $n > N_{th}$ )
  - 3:     Do ACB check using  $p_{2,r}$
  - 4:     Select a preamble from the preamble pool for second group
  - 5: else
  - 6:     Do ACB check using  $p_{1,r}$
  - 7:     Select a preamble from the preamble pool for first group
  - 8: end if
-

## 4.5. Performance evaluation

In this section, this dissertation presents simulation results of the throughput in time slot  $i$  over time and the average throughput with respect to  $M$ . The values for parameters used in simulation are summarized in Table 4.1. In the evaluation, the arrival distribution for devices is set as Beta ( $\alpha = 3, \beta = 4$ ) distribution.  $T_a$  is set as 10 seconds. In this evaluation, the maximum number of preamble transmission is not limited, i.e.  $N_{max}$  is infinity [4]. In the simulation, 1 subframe is equal to 1 milliseconds.  $T_{RAR}, W_{RAR}, W_{BO}$  are set as 3, 5, and 20 subframes, respectively [8].  $T_{RAREP}$  is 5 subframes and  $T_{UPDATE}$  is changed according to the si-periodicity [26, 37]. For example, when si-periodicity is 32,  $T_{UPDATE}$  becomes 160 ms.  $R_{min}$  is set as 3 [12], and  $R_1$ , and  $R_{1,1}$  of 54 [8].  $R_{max}$  is selected as 64 [11]. The  $N_{th}$  is selected as 4 as in previous chapter. A controllable parameter for D-ACB with DRA,  $b$ , is selected as 1.0 [4]. In this section, the BS can know the exact number of contended devices in a RACH, i.e.  $\hat{N}_i = N_i$ .

The Riverbed Modeler (known as OPNET Modeler) is used for the simulation of a conventional DARR and ACB protocol, which is dynamic ACB with dynamic resource allocation (D-ACB with DRA) [4], and for the simulation of proposed protocol. The state machines and the operations in each state for the devices and the BS have implemented, which simulate RACH procedure with given system model and the parameters for evaluation. In the simulation, each device selects an arrival time according to the arrival distribution and  $I_a$ .  $M_i[1]$  and  $M_i$  changes

randomly during simulation due to the arrival distribution and the probability to select a preamble in device. Thus, 6,000 simulations for each point in Figures from 4.2 to 4.9 are taken.

The performances of the proposed protocols are compared with the results with ideal case simulation. For the ideal case simulation, a C based simulator is implemented. The C based simulator includes algorithm in Algorithm 4.4, where the algorithm is referred as an ideal case algorithm in this dissertation. In Algorithm 4.4, the BS knows the number of activating devices in future time slots. In addition, the BS evaluates the performance of selected pool size and ACB factor to maximize the expected throughput of future time slots based on the knowledge of the number of activating devices in future time slots. 6,000 simulations for each point in Figures from 4.2 to 4.9 are taken for the ideal case algorithm simulation.

Figure 4.1 shows the average throughput for three DARR and ACB algorithms. The ideal case algorithm is expected to show best throughput due to future information, and the actual results are presented as expected. The throughput increases as the number of devices increases. For small number of devices, the DARR is used to change pool size. For large number of devices, the ACB is used to change ACB factor. Note that the DARR is difficult to obtain better throughput than ACB because the DARR generally controls pool size but the pool size can be selected in the domain of integer. The D-ACB with DRA shows low throughput for small number of devices, which means that it does not select the pool size to increase throughput. The proposed algorithm shows the throughputs



**Table 4.1.** Parameters for performance evaluation of preamble partition approach

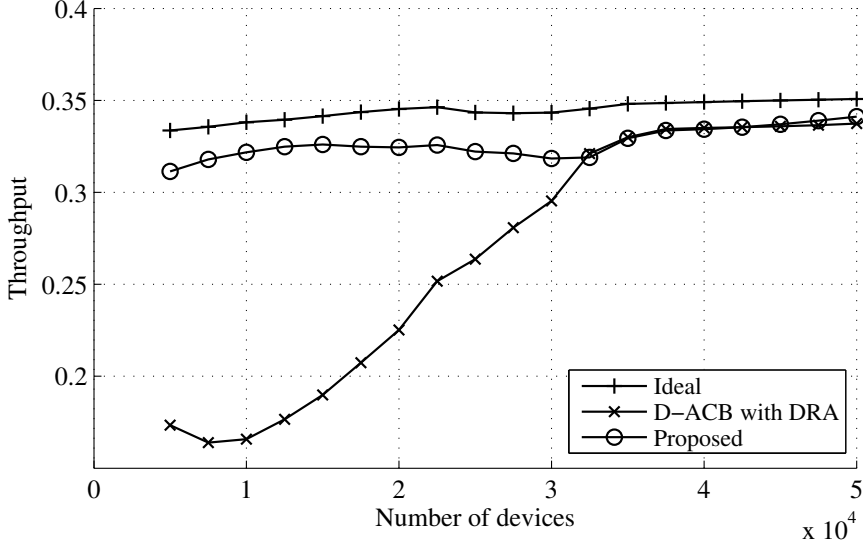
| <b>Parameters</b>    | <b>Value</b>                    |
|----------------------|---------------------------------|
| Arrival distribution | Beta( $\alpha = 3, \beta = 4$ ) |
| $T_a$                | 10 seconds                      |
| $I_a$                | 2000 slots                      |
| $N_{max}$            | infinity                        |
| $T_{RAR}$            | 3 subframes                     |
| $W_{RAR}$            | 5 subframes                     |
| $W_{BO}$             | 20 subframes                    |
| $T_{RAREP}$          | 5 subframes                     |
| $T_{UPDATE}$         | variable                        |
| $R_{min}$            | 3                               |
| $R_{max}$            | 64                              |
| $R_1, R_{1,1}$       | 54                              |
| $R_{2,1}$            | 0                               |
| $b$                  | 1.0                             |

---

**Algorithm 4.4** Ideal pool size selection for the comparison

---

- 1: On completing last RACH period in  $r$ -th si-period:
  - 2: assume that BS knows  $N_{rK}$  and  $M_{rK}$
  - 3: assume that BS knows  $\lambda_i$  for  $i = rK + 1, \dots, (r+1)K$
  - 4:  $R_{\text{sel}} = R_{\text{min}}, p_{\text{sel}} = 1, U_{\text{max}} = 0$
  - 5: for  $R = R_{\text{min}}$  to  $R_{\text{max}}$
  - 6:  $p = \frac{R}{M_{rK}}$
  - 7: obtain  $U_{\text{sum}} = \sum_{i=rK+1}^{(r+1)K} E[U_i | M_i, p_r = p, R_r = R]$
  - 8: if( $U_{\text{max}} > U_{\text{sum}}$ )
  - 9:  $U_{\text{max}} = U_{\text{sum}}$
  - 10:  $p_{\text{sel}} = p, R_{\text{sel}} = R$
  - 11: end if
  - 12: end for
  - 13:  $R_{r+1} = R_{\text{sel}}$
  - 14:  $p_{r+1} = p_{\text{sel}}$
-



**Fig. 4.1.** Average throughput vs. number of devices for the si-period of 32

close to that with ideal case algorithm. The proposed algorithm shows from 92.32% to 97.25% of average throughput of ideal case algorithm.

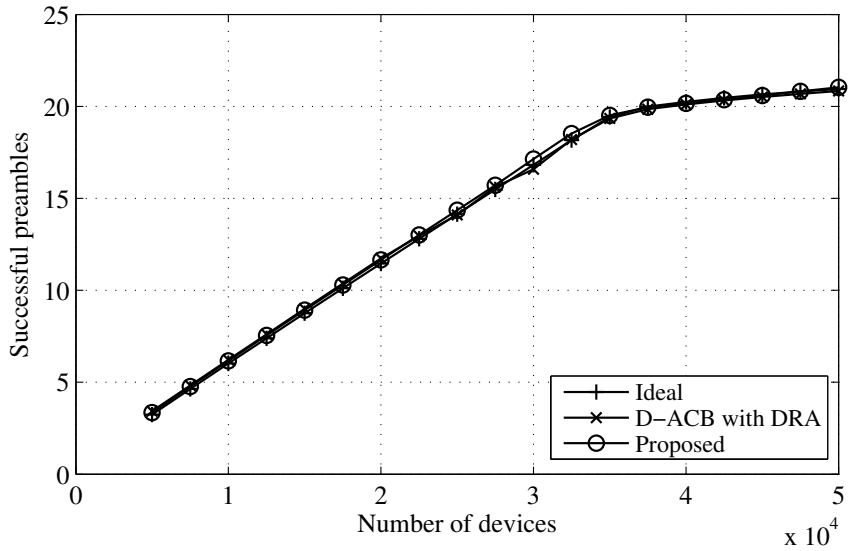
Figure 4.2 shows the average number of successful preambles per RACH and the average number of allocated preambles per RACH. The si-periodicity was set as 32. The average numbers of successful preambles per RACH are similar for three algorithms. Thus, the throughput is decided by the average number of allocated preambles per RACH. The ideal case algorithm shows lowest number of allocated preambles per RACH from the information for the future arrivals. The D-ACB with DRA allocates large number of preambles for DARR. The objective of D-ACB with DRA is to increase the success of random access rather than increase the throughput. In addition, the D-ACB with DRA selects very

low number for pool size at the start of random access, which decrease the speed to converge. The proposed algorithm shows the average number of allocated preambles per RACH close to the ideal case algorithm, although the proposed algorithm allocates little high pool size due to the lack of future information.

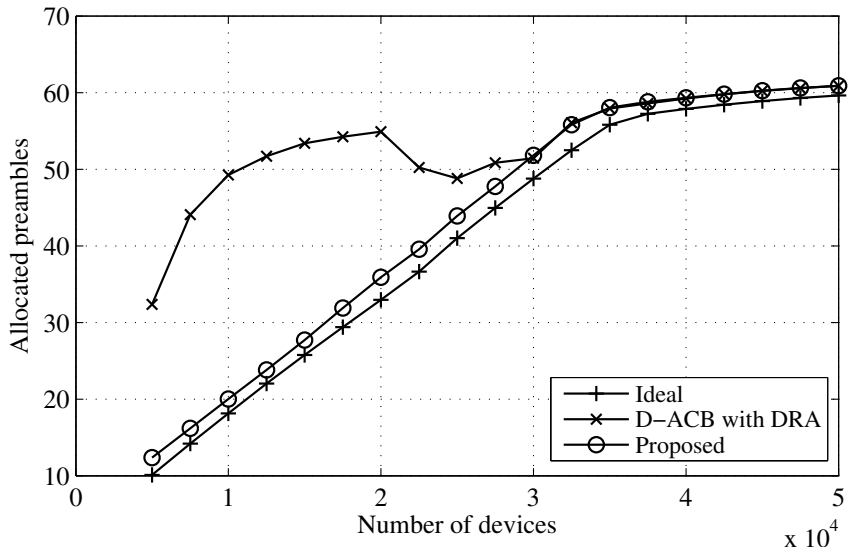
Figure 4.3 shows the cumulative distribution function (CDF) of throughputs with si-periodicity of 32 and different number of devices. The step size of throughput is selected as 0.0005. For the number of devices of 10,000 and 30,000, the CDF ensures that the actual throughput of proposed protocol will be not far from the average throughput. For the number of devices of 50,000, the proposed algorithm has higher probability for better throughput compared with probability distribution for the throughput of the D-ACB with DRA.

Figure 4.4 shows the average delay for three algorithms, and Figure 4.5 shows the same results but magnified for delay less than 0.5 s. Generally, the delay increases as the number of allocated preambles decreases. Thus, the ideal case shows high delay and D-ACB with DRA shows low delay. The proposed algorithm shows the delay close to the ideal case algorithm for small number of devices and the delay close to the D-ACB with DRA for large number of devices.

Figures 4.6, 4.7, and 4.8 shows the average throughput with respect to different si-periodicity when the number of devices is 10,000, 20,000, and 30,000 devices, respectively. As the si-periodicity increases, the BS hards to make  $M_i$  near to  $R_r$ , thus the average throughput decreases. D-ACB with DRA shows low throughput with small number of devices

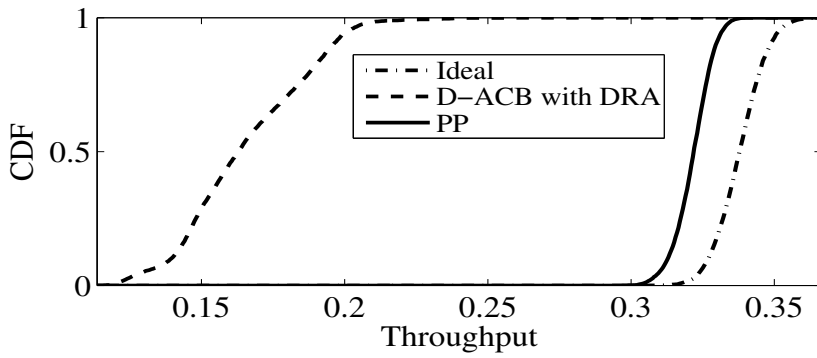


(a)

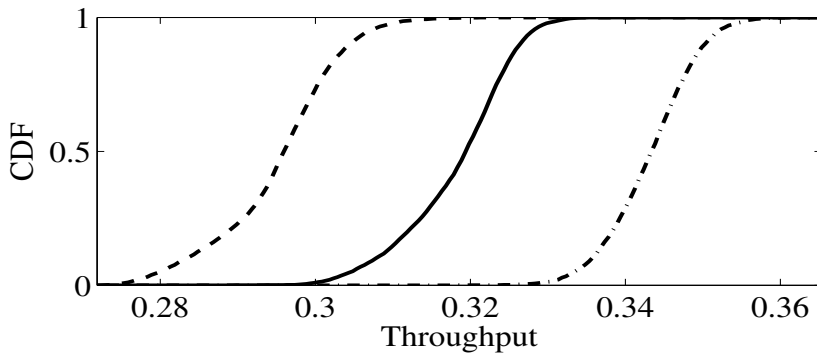


(b)

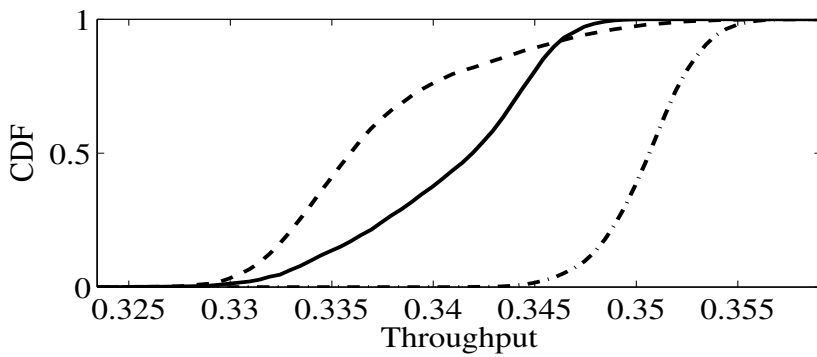
**Fig. 4.2.** (a) Average number of successful preambles per RACH (b) Average number of allocated preambles per RACH. Si-periodicity is equal to 32.



(a)

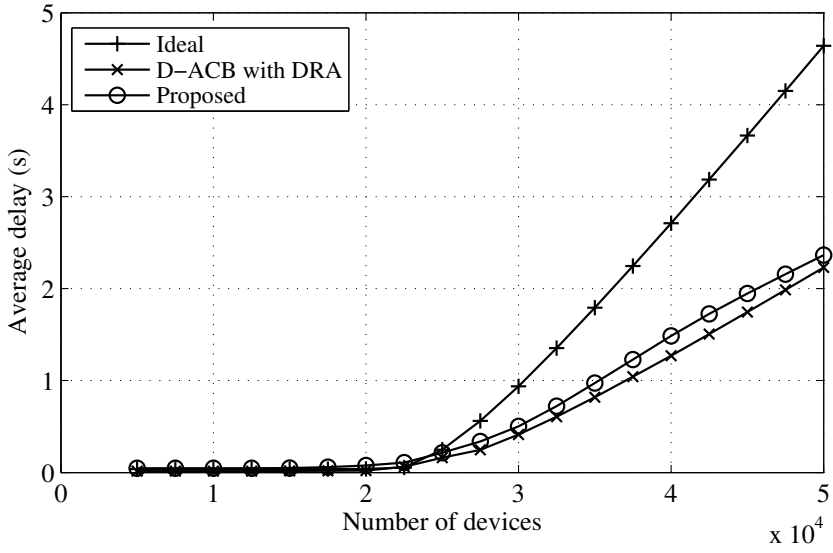


(b)

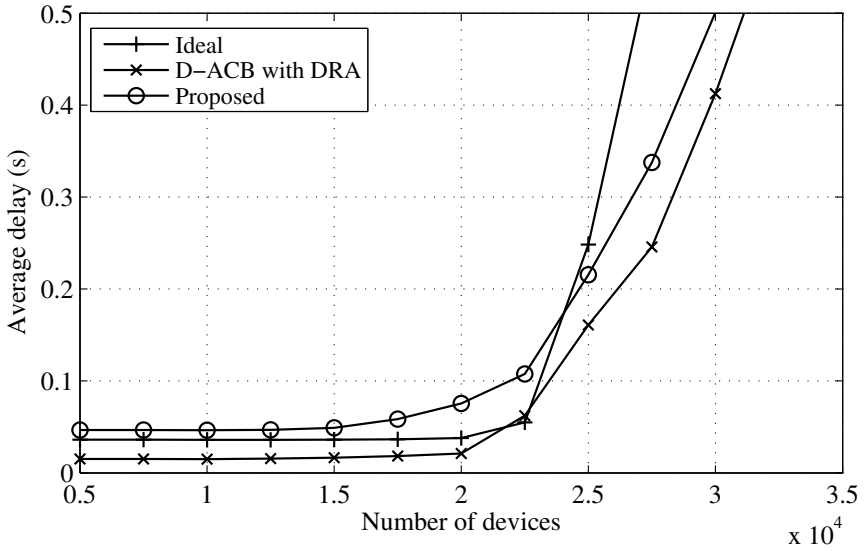


(c)

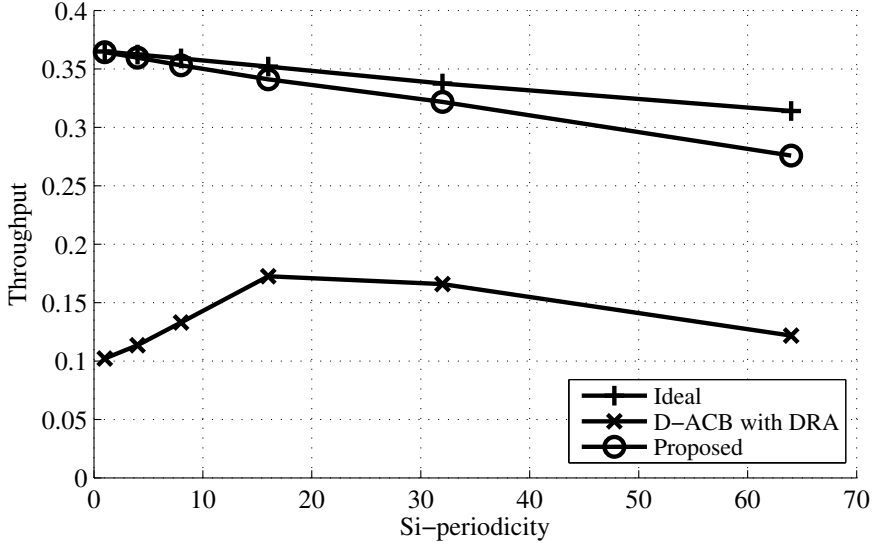
**Fig. 4.3.** CDF of throughputs for si-period of 32 with (a) 10,000 devices, (b) 30,000 devices, (c) 50,000 devices



**Fig. 4.4.** Average delay vs. number of devices for si-period of 32



**Fig. 4.5.** Average delay vs. number of devices for si-period of 32, magnified for delay less than 0.5 s



**Fig. 4.6.** Average throughput vs. si-periodicity for 10,000 devices

since it changes the number of preambles to maximize the number of successful devices, not the throughput. The D-ACB with DRA shows better throughput with 30,000 devices, since the ACB is frequently required with high number of devices. The proposed algorithm improves the throughput for DARR, thus the proposed algorithm shows better throughput which also close to the ideal case simulation.

Figure 4.9 shows the CDF of throughputs with 30,000 devices and different si-periodicity. The step size of throughput is selected as 0.0005. The CDF ensures that the actual throughput of proposed protocol will be not far from the average throughput.



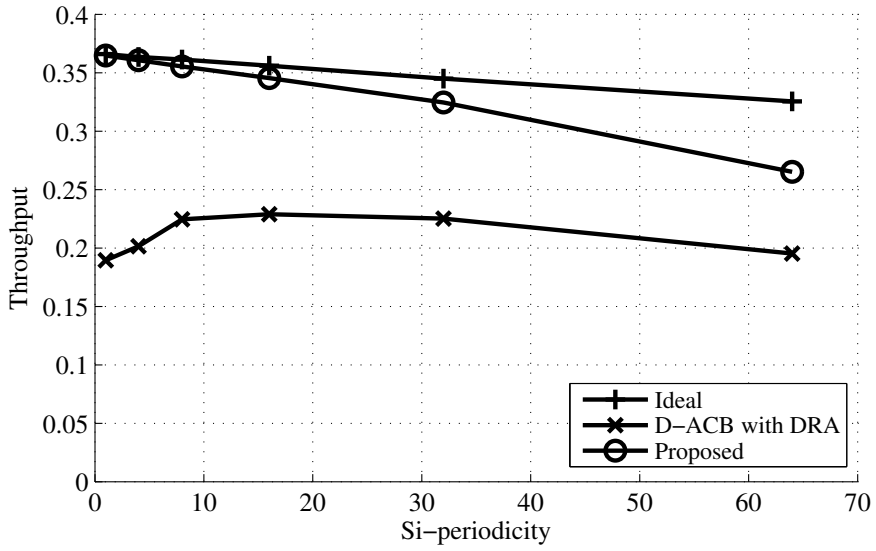


Fig. 4.7. Average throughput vs. si-periodicity for 20,000 devices

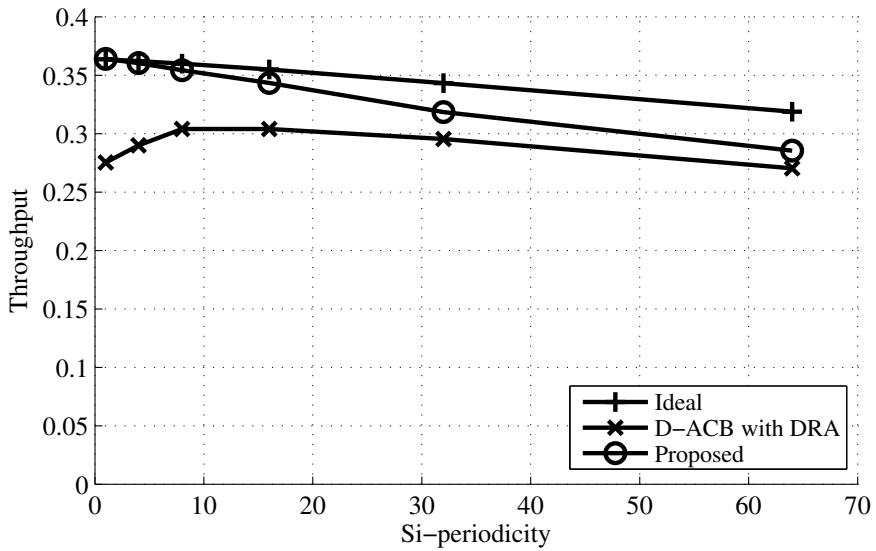
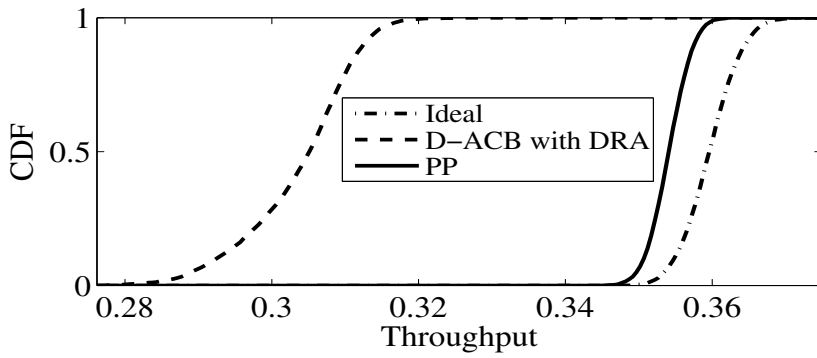
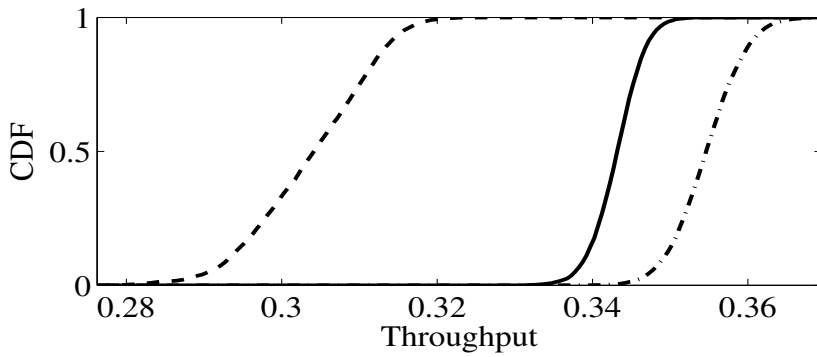


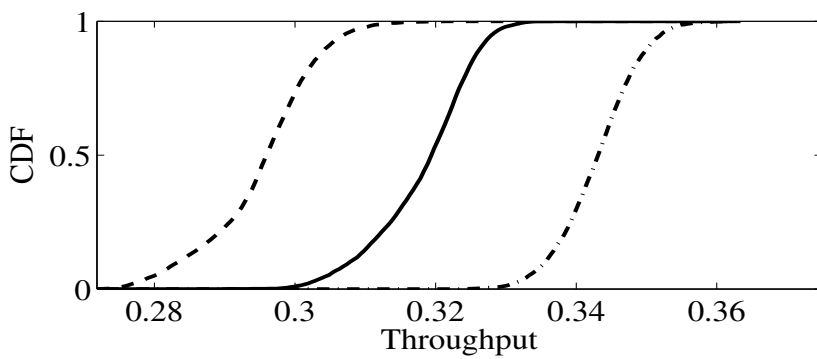
Fig. 4.8. Average throughput vs. si-periodicity for 30,000 devices



(a)



(b)



(c)

**Fig. 4.9.** CDF of utilities for different si-periodicity with  $M$  of 30,000 and si-period of (a) 8 time slots, (b) 16 time slots, (c) 32 time slots

## 4.6. Summary

In this chapter, the challenge of the throughput degradation for DARR and ACB in LTE-A due to si-periodicity was discussed. To overcome the traffic overload and to provide efficient preamble resource usage, a novel DARR and ACB protocol is proposed in this dissertation. The proposed protocol provides high throughput in both DARR and ACB by mixing the preamble partition approach in DARR and the basis of ACB. The proposed DARR and ACB protocol can be used to adaptively improve the throughput of the RACH procedure in LTE-A, or RA schemes based on a frame slotted or multi-channel ALOHA.

# Chapter 5. Optimizing Random Access Procedure for the Data Transmission of MTC Devices

## 5.1. Introduction

The data transmission for a H2H device in idle state requires the initial attach, and RRC connection to setup SRB1, where SRB means signaling radio bearer. In addition, the device should obtain the resource in uplink after the RRC connection. However, the M2M or IoT devices will frequently enter to the idle state since the traffic generation interval will be very long. Thus, the conventional data transmission procedure requires random access procedure which causes the large overhead for the transmission of data of M2M or IoT devices [68]. Therefore, 3GPP recently changes the specification for not to remove the information of devices in the BS after the initial attach. In addition, the 3GPP developed two small data transmission (SDT) procedures, where one is the data over non-access stratum (NAS) and the other is the data over user plane [19, 26].

However, the data over NAS or the data over user plane requires some overheads since they still exchange the messages for RRC connection. Two alternative data transmission procedures in previous studies can be considered as another SDT procedures to reduce overhead. One is the data in MSG1, where the device transmits the data instead of a preamble (MSG1) [69, 70]. The other is the data in MSG3, where the device transmits the data with RRC connection request (MSG3) or instead of MSG3 [2]. These approaches can reduce the number of message exchanges between devices and the BS. However, the collision can be detected after the transmission of data when the data in MSG1 or the data in MSG3 is used. Therefore, when the size of data becomes larger than the size of MSG3, these schemes can decrease overall resource efficiency although the number of signaling decreases.

The performance of four SDT procedures needs to be evaluated in terms of the resource usage. The previous studies evaluated their SDT procedures for the number of access success or delay, but the resource usage is not well discussed. In addition, LTE-A allocates the time-frequency resources to the devices with the unit of resource block (RB), where the previous studies assumed with the unit of bits.

In this Chapter, this dissertation evaluates the expected number of resource blocks occurred from the contention in a RACH for these conventional random access procedures. In addition, the uplink resource efficiency is also evaluated where the uplink resource efficiency is the ratio of the number of successful data transmission to the the number of RBs allocated from a RACH. Based on the evaluation, the alternative

SDT procedure for the M2M services in LTE-A is proposed to reduce overhead. The comparison of performance with other SDT procedures is also performed.

## 5.2. System Model

Suppose that a cell consists of a BS and  $M$  MTC devices. Let these devices already did their initial attach procedure and entered idle state. Let these devices need to transmit their data. Since they are entered in idle state, they require the random access for the data transmission.

The BS allocates a RACH periodically. Let a time slot be an interval between two RACHs, where each time slot starts with a RACH. Suppose that there are  $R$  preambles in the pool, where the pool is given as  $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_R\}$ . Assume that  $M$  devices transmit preambles where each preamble is randomly chosen from  $\mathbf{C}$  in a time slot. The sizes of MSG2, MSG3, MSG4, and MSG5 in bits are given as  $L_2$ ,  $L_3$ ,  $L_4$ , and  $L_5$  respectively. Let the size of data in bits be  $L_D$ , where the size of data includes upper layer headers and payload from application layer. The size of data is a random variable where the probability that the size of data is equal to  $l$  and its probability density function is given as  $g(l)$ . The minimum and maximum size of data are given as  $L_{\min}$  and  $L_{\max}$ , respectively. The RBs allocated for the data transmission is sometimes called as "data block" in this section. The device can transmit  $L_{UL}$  bits using a RB in uplink band, and the BS can transmit  $L_{DL}$  bits using a RB in downlink band. The BS can allocate the resources in the unit of

RB, but it cannot allocate the resource in the unit of bit.

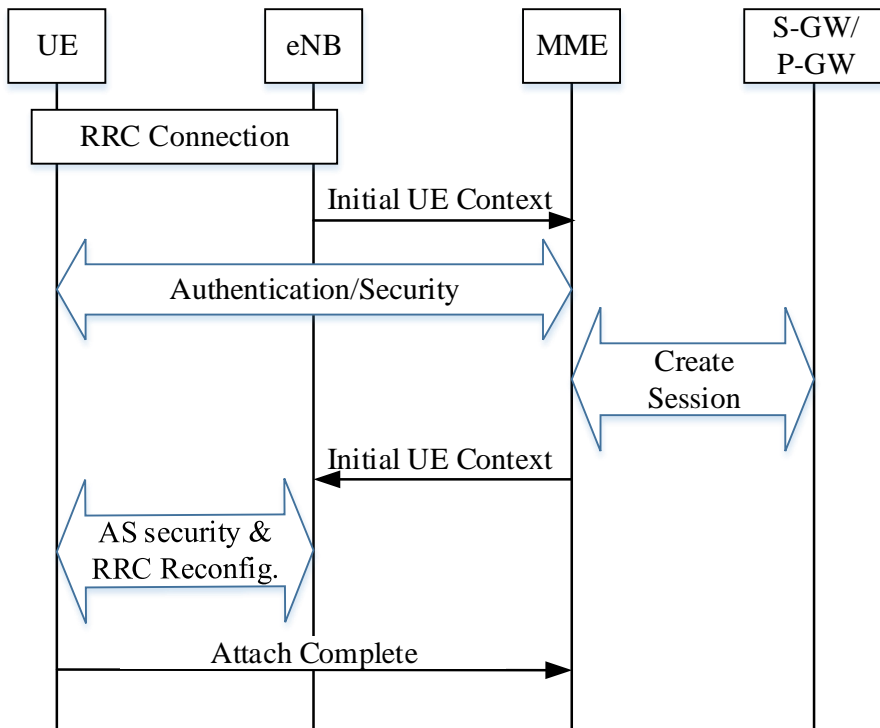
A MSG2 requires a transmission in common space PDCCH and a transmission in PDSCH. The multiple MSG2s can be concatenated for the transmission in PDSCH, where they share one common space PDCCH. Let  $N_{RAR}$  be the number of MSG2s which share one common space PDCCH.  $N_{RAR}$  can be given as the number of uplink (UL) grants per RAR [8]. Each transmission for MSG3, MSG4, MSG5 and data requires a transmission in control channel and a transmission in shared channels.

Assume that the BS can always decode preambles and the devices can always decode MSG2s. Assume that a transmission for MSG3, MSG4, MSG5, or data cannot be decoded with probability of  $p$ . The hybrid automatic repeat request (HARQ) is used for the transmission of MSG3, MSG4, MSG5 and data. For HARQ, the maximum number of transmissions for a message or a data is given as  $N$ , i.e.  $1 \leq n \leq N$ .

### **5.3. Conventional procedures for the data transmission using RA**

#### **5.3.1 Control plane (CP) solution**

Any devices for IoT services do the initial attach procedure before the data transmission [71]. Figure 5.1 shows the initial attach procedure. An activated device does RRC connection using the RACH procedure. After the RRC connection, the initial UE context message is transmitted from the BS to the mobile management entity (MME). The UE and the MME



**Fig. 5.1.** Initial attach procedure

do authentication and security setup. The session between the MME and the serving gateway(S-GW) or between the MME and the packet data network gateway (P-GW) is then created. The MME transmits initial UE context message to BS. The UE and BS do the security setup and RRC reconfiguration if CP solution is used. The initial attach procedure is completed by transmitting attach complete message to MME.

Since IoT devices has long interval between data generation, the UE is in idle state in most case. For the transmission of data from idle state,



the CP solution is proposed for NB-IoT [72]. Figure 5.2 shows the data transmission procedure of CP solution. As shown in the figure, the main idea of the CP solution is the transmission of data by piggybacking to MSG5. If the BS receives MSG5 and data after the random access, the BS transmits the data to MME. MME decrypts the data and does integrity check. Then, the MME forward the data to S-GW/P-GW using the session created during initial attach procedure. This scheme is also called as control plane solution (CP solution) [25], since the data is transmitted to serving gateway through the control network.

The RA procedure for the CP solution can be summarized as follows:

- **Preallocation:** The BS allocates 6 RBs for preamble transmission in every RACH periodicity.
- **Step 1 (Preamble):** The device transmits one of preamble from the preamble pool in the RACH.
- **Step 2 (RAR):** When the BS can detect the preamble, the BS responses with RAR.
- **Step 3 (MSG3):** If the device receives RAR corresponding to the transmitted preambles, it transmits its RRC connection request in addition to the medium access control element (MAC CE). The MAC CE includes the data volume which assists an BS to efficiently allocate radio resources for the data transmission. The hybrid automatic repeat request (HARQ) can be used for the reliability.
- **Step 4 (MSG4):** When the BS can decode the MSG3, the BS

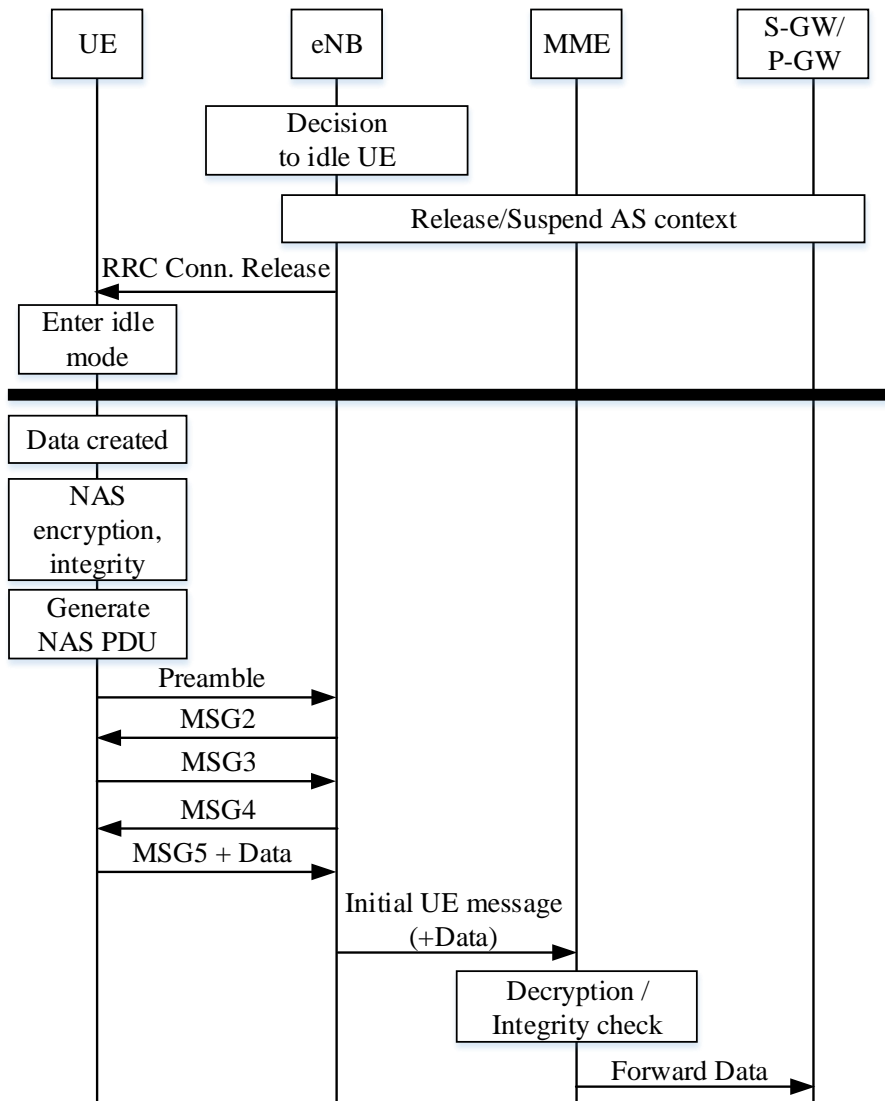


Fig. 5.2. Data transmission procedure of CP solution

responses with contention resolution and RRC connection setup message. The HARQ can be used for the reliability.

- **Step 5 (MSG5 and Data):** If the device receives MSG4 with the ID of itself, the device transmits MSG5 and data to the BS. The HARQ can be used for the reliability.

### 5.3.2 User plane (UP) solution

The device can use UP solution for the IoT service. The UP solution is very similar to the RRC connection procedure in conventional RACH procedure. Figure 5.3 shows the data transmission procedure of UP solution. As shown in the figure, the UE and BS exchanges preambles and four RRC messages (MSG2, MSG3, MSG4, MSG5). After the exchange of RRC messages, the BS activates AS context between BS and S-GW/P-GW. The BS then allocates the RBs for the transmission of data. The UE then transmits its data to BS, and the BS forwards the data to S-GW/P-GW. The RA procedure for the UP solution can be summarized as follows:

- **Preallocation:** The BS allocates 6 RBs for preamble transmission in every RACH periodicity.
- **Step 1 (Preamble):** The device transmits one of preamble from the preamble pool in the RACH.
- **Step 2 (RAR):** When the BS can detect the preamble, the BS responds with RAR.

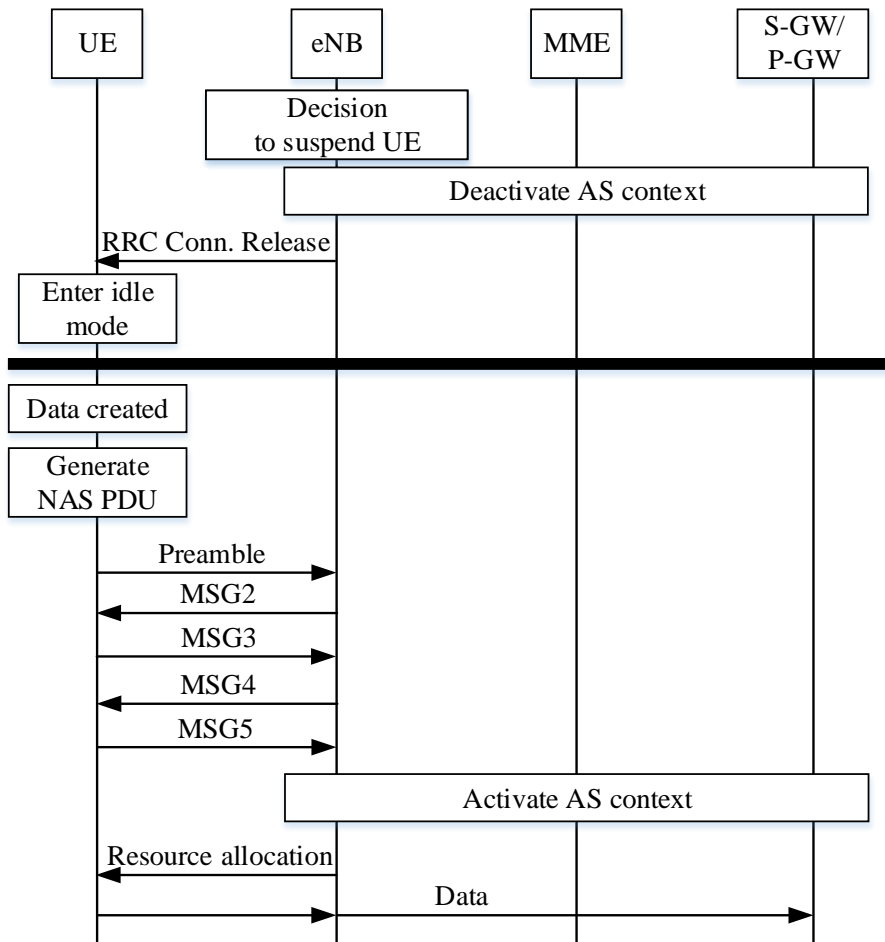


Fig. 5.3. Data transmission procedure of CP solution

- **Step 3 (MSG3):** If the device receives RAR corresponding to the transmitted preambles, it transmits its RRC connection request in addition to the medium access control element (MAC CE). The MAC CE includes the data volume which assists an BS to efficiently allocate radio resources for the data transmission. The hybrid automatic repeat request (HARQ) can be used for the reliability.
- **Step 4 (MSG4):** When the BS can decode the MSG3, the BS responds with contention resolution and RRC connection setup message. The HARQ can be used for the reliability.
- **Step 5 (MSG5):** If the device receives MSG4 with the ID of itself, the device transmits MSG5 to the BS. The HARQ can be used for the reliability.
- **Step 6 (Resource allocation):** If the BS receives MSG5, the BS allocates the resources in uplink band for the transmission of data. The BS sends UL grant for the data.
- **Step 7 (Data transmission):** If the device receives UL grant, the device transmits its data to BS.

### 5.3.3 Data in MSG1

The main idea of the data in MSG1 is the transmission of data in the first step instead of the transmission of preambles. The data in MSG1 removes the transmission of from MSG3 to MSG5 in the conventional

data transmission procedure in LTE-A. Thus, the RA procedure for the data in MSG1 can be summarized as follows:

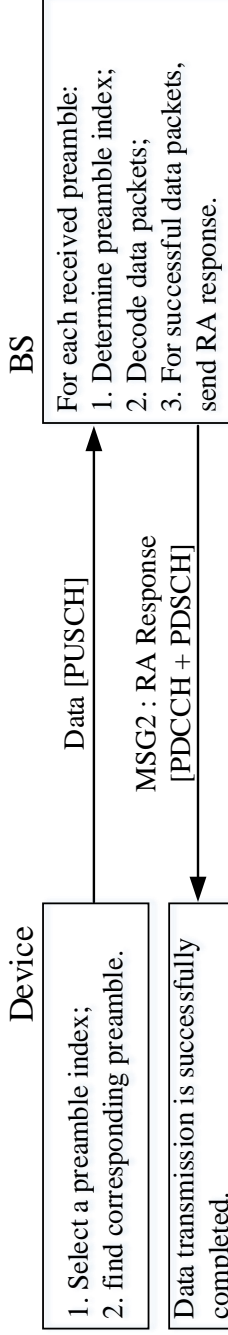
- **Preallocation:** The BS allocates  $R$  of data blocks in every RACH periodicity instead of RACH.
- **Step 1 (Data):** The device transmits their data in the one of data blocks.
- **Step 2 (RAR):** When the BS can decode the data, the BS responds with RAR. If the device receives its RAR, the data transmission is successfully completed.

Figure 5.4 shows the procedure for data in MSG1.

### 5.3.4 Data in MSG3

The data can be transmitted with MSG3 to reduce resource usage for MSG5 and MSG6 in conventional data transmission procedure. In addition, MSG4 can be replaced to an acknowledgment to reduce the size MSG4. This approach is the data in MSG3 which is presented in [2] and [25]. The main idea of the data in MSG3 is the transmission of data in the third step in addition to the RRC connection request message, where RRC connection request message is transmitted to the recognition of MTC device. The RA procedure for the data in MSG3 can be summarized as follows:

- **Preallocation:** The BS allocates 6 RBs for preamble transmission in every RACH periodicity.



**Fig. 5.4.** RA procedure with the data in MSG1

- **Step 1 (Preamble):** The device transmits one of preamble from the preamble pool in the RACH.
- **Step 2 (RAR):** When the BS can detect the preamble, the BS responses with RAR.
- **Step 3 (Data):** If the device receives RAR corresponding to the transmitted preambles, it transmits its data and the ID of the device using the RBs allocated by the BS. The hybrid automatic repeat request (HARQ) is used for the reliability.
- **Step 4 (Contention resolution):** When the BS can decode the data, the BS responses with MSG4. If the device receives MSG4 with the ID of itself, the data transmission is successfully completed. The HARQ is used for the reliability.

#### 5.4. Numerical evaluation for the conventional SDT procedures

Let  $M_I$ ,  $M_S$ , and  $M_C$  be the number of idle, successful, and collided data for a RACH, respectively. These numbers are random variables, thus they can be changed in every RACH. Let the statistical mean of the number of idle preambles, successful preambles, and collided preambles be  $\mathbb{E}[M_I]$ ,  $\mathbb{E}[M_S]$ , and  $\mathbb{E}[M_C]$ , respectively. The contention in a RACH can be modeled as a balls and bins problem [64]. Let the probability that there are  $k$  balls fall in a bin  $b$  be  $Pr[N_b = k]$ . Since there are  $R$  bins in



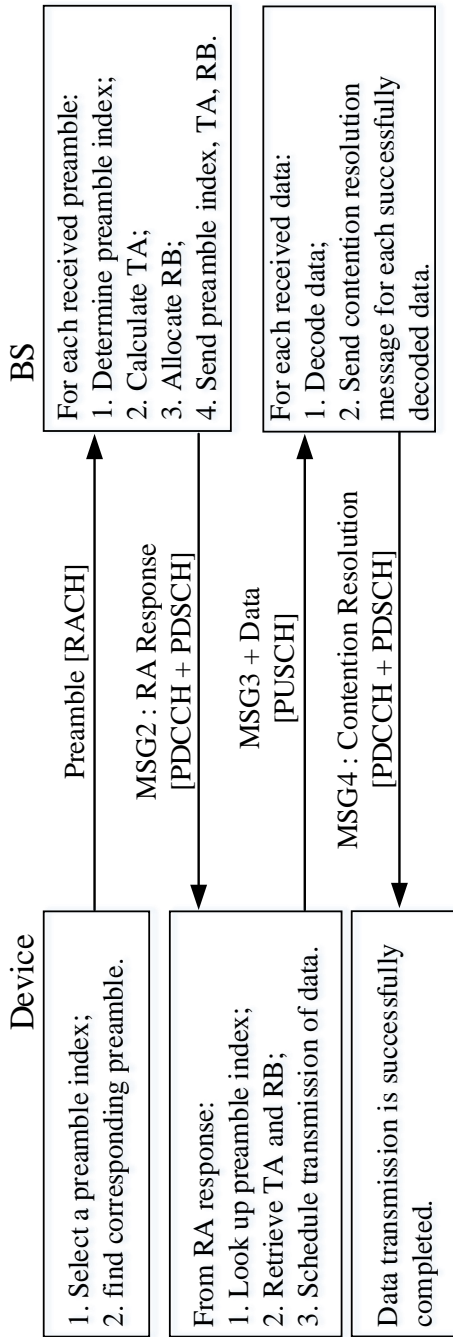


Fig. 5.5. RA procedure with the data in MSG3

RACH,  $Pr[N_b = k]$  can be obtained using following equation:

$$Pr[N_b = k] = \binom{M}{k} \frac{1}{R} \left(1 - \frac{1}{R}\right)^{M-k}. \quad (5.1)$$

Let  $P_I$ ,  $P_S$ , and  $P_C$  be the probability that the bin is empty, has one ball, and has two or more balls, respectively. These probabilities represent the probabilities that a preamble is idle, successful, and collided, respectively. Three probabilities are equal to

$$P_I = Pr[N_b = 0] = \left(1 - \frac{1}{R}\right)^M, \quad (5.2)$$

$$P_S = Pr[N_b = 1] = \frac{M}{R} \left(1 - \frac{1}{R}\right)^{M-1}, \quad (5.3)$$

$$P_C = Pr[N_b > 2] = 1 - P_I - P_S. \quad (5.4)$$

$\mathbb{E}[M_I]$ ,  $\mathbb{E}[M_S]$ , and  $\mathbb{E}[M_C]$  can be obtained by the multiplication of  $M$  and their probabilities. Thus, they are equal to

$$\mathbb{E}[M_I] = RP_I = R \left(1 - \frac{1}{R}\right)^M, \quad (5.5)$$

$$\mathbb{E}[M_S] = RP_S = M \left(1 - \frac{1}{R}\right)^{M-1}, \quad (5.6)$$

$$\mathbb{E}[M_C] = RP_C = R \{1 - P_I - P_S\}. \quad (5.7)$$

### 5.4.1 CP solution

When the CP solution is applied to the conventional LTE-A, the BS preallocates 6 RBs in uplink band for step 1 in every time slot. Let  $S_{cp,1}$  be the number of preambles arrived to the BS in a RACH. In a

RACH, the BS can decode multiple preambles, which include successful and collided preambles. Thus, the expectation of  $S_{cp,1}$  is equal to

$$\mathbb{E}[S_{cp,1}] = \mathbb{E}[M_S] + \mathbb{E}[M_C]. \quad (5.8)$$

The BS will response for each decoded preamble using multiple MSG2, where an MSG2 includes multiple UL grants for multiple preambles, where an UL grant is corresponding to a preamble. Let  $S_{cp,2}$  be the number of MSG2s arrived to the device. The BS responses all decoded preambles. Thus, the expectation of  $S_{cp,2}$  is equal to

$$\mathbb{E}[S_{cp,2}] = \mathbb{E}[S_{cp,1}] = \mathbb{E}[M_S] + \mathbb{E}[M_C]. \quad (5.9)$$

The devices received RA response will send their MSG3. For each transmission for the MSG3, HARQ will cause additional allocations of RBs. Let  $S_{cp,3}$  be the number of MSG3s which are successfully decoded in the BS. A MSG3 can be decoded in the BS if the MSG3 is transmitted for the successful preamble and if HARQ is successful. Thus, the expectation of  $S_{cp,3}$  is equal to

$$\mathbb{E}[S_{cp,3}] = \mathbb{E}[M_S] (1 - p^N). \quad (5.10)$$

Let  $F_{cp,3}$  be the number of transmissions for MSG3 which fail to decode in BS. For the collided preambles, each device transmits  $N$  times of MSG3. For the successful preambles, each device transmits additional MSG3 until success or reaching to  $N$ . Therefore, the expectation of  $F_{cp,3}$  is equal to

$$\mathbb{E}[F_{cp,3}] = N\mathbb{E}[M_C] + \mathbb{E}[M_S] \left\{ Np^N + \sum_{n=1}^{N-1} (n-1)p^n(1-p) \right\}. \quad (5.11)$$

The transmission of MSG4 is similar to the MSG3. Let  $S_{cp,4}$  be the number of MSG4 arrived to the device. The BS transmits MSG4 for each successful arrival of MSG3. Thus, the expectation for the  $S_{cp,4}$  is equal to

$$\mathbb{E}[S_{cp,4}] = \mathbb{E}[S_{cp,3}] (1 - p^N). \quad (5.12)$$

Let  $F_{cp,4}$  be the number of transmissions for MSG4 which fail to decode in BS, which is equal to

$$\mathbb{E}[F_{cp,4}] = \mathbb{E}[S_{cp,3}] \left\{ Np^N + \sum_{n=1}^{N-1} (n-1)p^n(1-p) \right\}. \quad (5.13)$$

The transmission of MSG5 with data is similar to the MSG4. Let  $S_{cp,5}$  be the number of MSG5 with data arrived to the BS. A device transmits MSG4 if the arrival of MSG4 is successful. The transmission of MSG5 with data is also retransmitted by HARQ. Thus, the expectation for the  $S_{cp,5}$  is equal to

$$\mathbb{E}[S_{cp,5}] = \mathbb{E}[S_{cp,4}] (1 - p^N). \quad (5.14)$$

Let  $F_{cp,5}$  be the number of transmissions for MSG3 and data which fail to decode in BS, which is equal to

$$\mathbb{E}[F_{cp,5}] = \mathbb{E}[S_{cp,4}] \left\{ Np^N + \sum_{n=1}^{N-1} (n-1)p^n(1-p) \right\}. \quad (5.15)$$

Let  $B_{cp,1}$ ,  $B_{cp,2}$ ,  $B_{cp,3}$ ,  $B_{cp,4}$ , and  $B_{cp,5}$  be the average number of allocated RBs for the transmission of preambles, MSG2, MSG3, MSG4, and MSG5 with data, respectively. For the transmission of preambles, the BS allocate 6 RBs in LTE-A, thus  $B_{cp,1} = 6$ . The  $N_{RAR}$  of MSG2 can be concatenated for a PDCCH. For the given  $S_{cp,2}$  transmissions of

MSG2,  $\left\lfloor \frac{S_{cp,2}}{N_{RAR}} \right\rfloor$  of MSG2s share each PDCCH, where  $\lfloor x \rfloor$  is the largest integer smaller than  $x$ . Remaining  $\left( S_{cp,2} - N_{RAR} \left\lfloor \frac{S_{cp,2}}{N_{RAR}} \right\rfloor \right)$  MSG2s share 1 PDCCH. Thus, the expectation for  $B_{cp,2}$  is equal to

$$\mathbb{E}[B_{cp,2}] = \left\lfloor \frac{\mathbb{E}[S_{cp,2}]}{N_{RAR}} \right\rfloor \left\lfloor \frac{N_{RAR}L_2}{L_{DL}} \right\rfloor + \left\lfloor \frac{\left( \mathbb{E}[S_{cp,2}] - N_{RAR} \left\lfloor \frac{\mathbb{E}[S_{cp,2}]}{N_{RAR}} \right\rfloor \right) L_2}{L_{DL}} \right\rfloor. \quad (5.16)$$

where  $\lceil x \rceil$  is the smallest integer larger than  $x$ . For each transmission of MSG3, which includes success and failure, the BS allocates  $\left\lceil \frac{L_3}{L_{UL}} \right\rceil$  RBs. Thus, the expectation for  $B_{cp,3}$  is equal to

$$\mathbb{E}[B_{cp,3}] = (\mathbb{E}[S_{cp,3}] + \mathbb{E}[F_{cp,3}]) \left\lceil \frac{L_3}{L_{UL}} \right\rceil. \quad (5.17)$$

For each transmission of MSG4, the BS allocates  $\left\lceil \frac{L_4}{L_{UL}} \right\rceil$  RBs. Thus, the expectation for  $B_{cp,4}$  is equal to

$$\mathbb{E}[B_{cp,4}] = (\mathbb{E}[S_{cp,4}] + \mathbb{E}[F_{cp,4}]) \left\lceil \frac{L_4}{L_{UL}} \right\rceil. \quad (5.18)$$

For each transmission of MSG5 with data, where the size of data is given as  $l$ , the BS allocates  $\left\lceil \frac{l+L_5}{L_{UL}} \right\rceil$  RBs. Thus, the expectation for  $B_{cp,5}$  is equal to

$$\mathbb{E}[B_{cp,5}] = \sum_{l=L_{\min}}^{L_{\max}} (\mathbb{E}[S_{cp,5}] + \mathbb{E}[F_{cp,5}]) \left\lceil \frac{l+L_5}{L_{UL}} \right\rceil g(l). \quad (5.19)$$

Let  $B_{cp,UL}$  and  $B_{cp,DL}$  be the total number of allocated RBs for a time slot in uplink and downlink band, respectively. The expectation of  $B_{cp,UL}$  is the sum of the expectations of  $B_{cp,1}$ ,  $B_{cp,3}$ , and  $B_{cp,5}$ . The expectation

of  $B_{cp,DL}$  is the sum of the expectations of  $B_{cp,2}$  and  $B_{cp,4}$ .

$$\mathbb{E}[B_{cp,UL}] = \mathbb{E}[B_{cp,1}] + \mathbb{E}[B_{cp,3}] + \mathbb{E}[B_{cp,5}], \quad (5.20)$$

$$\mathbb{E}[B_{cp,DL}] = \mathbb{E}[B_{cp,2}] + \mathbb{E}[B_{cp,4}]. \quad (5.21)$$

Let  $U_{cp,UL}$  be the uplink resource efficiency for CP solution, where the uplink resource efficiency be the ratio of the number of MTC devices that succeeded the transmission of data to  $B_{cp,UL}$ . Since the expectation for the number of MTC devices that succeeded the transmission of data is  $\mathbb{E}[S_{cp,5}]$ ,  $U_{cp,UL}$  can be obtained as

$$\mathbb{E}[U_{cp,UL}] = \frac{\mathbb{E}[S_{cp,5}]}{\mathbb{E}[B_{cp,UL}]}. \quad (5.22)$$

## 5.4.2 UP solution

Let  $B_{up,1}$ ,  $B_{up,2}$ ,  $B_{up,3}$ , and  $B_{up,4}$  be the average number of allocated RBs for the transmission of preambles, MSG2, MSG3, and MSG4 respectively, invoked from a RACH for the UP solution. From the transmission of preambles to the transmission of MSG4, the CP solution and the UP solution has same procedure. Thus, the expectation of  $B_{up,1}$ ,  $B_{up,2}$ ,  $B_{up,3}$ , and  $B_{up,4}$  are equal to that of  $B_{cp,1}$ ,  $B_{cp,2}$ ,  $B_{cp,3}$ , and  $B_{cp,4}$ , respectively.

Let  $B_{up,5}$  be the average number of allocated RBs for the transmission of MSG5. Since the data is not transmitted with MSG5 in UP solution, the BS allocates  $\left\lceil \frac{L_5}{L_{UL}} \right\rceil$  RBs for each transmission of MSG5. Thus, the expectation for  $B_{cp,5}$  is equal to

$$\mathbb{E}[B_{up,5}] = (\mathbb{E}[S_{cp,5}] + \mathbb{E}[F_{cp,5}]) \left\lceil \frac{L_5}{L_{UL}} \right\rceil. \quad (5.23)$$

Let  $B_{up,6}$  be the average number of allocated RBs for the transmission of resource allocation. For each successful transmission of MSG5, the resource allocation is transmitted, and HARQ is not applied. Let  $L_6$  be the length of resource allocation message. The expectation of  $B_{up,6}$  is equal to

$$\mathbb{E}[B_{up,6}] = \mathbb{E}[S_{cp,5}] \left[ \frac{L_6}{L_{DL}} \right]. \quad (5.24)$$

Let  $S_{up,7}$  be the number of data arrived to the BS. A device transmits its data for its resource allocation if the arrival of MSG4 is successful. The transmission of data is also retransmitted by HARQ. Thus, the expectation for the  $S_{up,7}$  is equal to

$$\mathbb{E}[S_{up,7}] = \mathbb{E}[S_{cp,5}] (1 - p^N). \quad (5.25)$$

Let  $F_{up,7}$  be the number of transmissions for data which fail to decode in BS. The expectation of  $F_{up,7}$  is equal to

$$\mathbb{E}[F_{up,7}] = \mathbb{E}[S_{cp,5}] \left\{ Np^N + \sum_{n=1}^{N-1} (n-1)p^n(1-p) \right\}. \quad (5.26)$$

Let  $B_{up,7}$  be the average number of allocated RBs for the transmission of data.  $B_{up,7}$  is equal to

$$\mathbb{E}[B_{up,7}] = \sum_{l=L_{\min}}^{L_{\max}} (\mathbb{E}[S_{up,7}] + \mathbb{E}[F_{up,7}]) \left[ \frac{l}{L_{UL}} \right] g(l). \quad (5.27)$$

Let  $B_{up,UL}$  and  $B_{up,DL}$  be the total number of allocated RBs for a time slot in uplink and downlink band, respectively. The expectation of  $B_{up,UL}$  is the sum of the expectations of  $B_{up,1}$ ,  $B_{up,3}$ ,  $B_{up,5}$ , and  $B_{up,7}$ . The expectation of  $B_{up,DL}$  is the sum of the expectations of  $B_{up,2}$ ,  $B_{up,4}$ ,

and  $B_{up,6}$ .

$$\mathbb{E}[B_{up,UL}] = \mathbb{E}[B_{up,1}] + \mathbb{E}[B_{up,3}] + \mathbb{E}[B_{up,5}] + \mathbb{E}[B_{up,7}], \quad (5.28)$$

$$\mathbb{E}[B_{up,DL}] = \mathbb{E}[B_{up,2}] + \mathbb{E}[B_{up,4}] + \mathbb{E}[B_{up,6}]. \quad (5.29)$$

Let  $U_{up,UL}$  be the uplink resource efficiency for UP solution. Since the expectation for the number of MTC devices that succeeded the transmission of data is  $\mathbb{E}[S_{up,7}]$ ,  $U_{up,UL}$  can be obtained as

$$\mathbb{E}[U_{up,UL}] = \frac{\mathbb{E}[S_{up,7}]}{\mathbb{E}[B_{up,UL}]} \quad (5.30)$$

### 5.4.3 Data in MSG1

For the data in MSG1, the BS needs to allocate  $R$  of data blocks in uplink band in every RACH periodicity instead of the RACH for the transmission of data from devices. The size of each data block should be the maximum of the size of data, otherwise the data cannot be transmitted. Thus, the preallocation requires  $R \left\lceil \frac{L_{\max}}{L_{UL}} \right\rceil$  RBs in uplink band. The data from an MTC device can be decoded if only one device selects a data block. In this case, the BS responses with RA response message. If a data block is unused, or the collision happens in a data block, the BS will not response for the allocated data block.

Let  $B_{d1,UL}$  and  $B_{d1,DL}$  be the total number of allocated RBs for a time slot in uplink and downlink band, respectively. They are equal to

$$\mathbb{E}[B_{d1,UL}] = R \left\lceil \frac{L_{\max}}{L_{UL}} \right\rceil, \quad (5.31)$$



$$\mathbb{E}[B_{d1,DL}] = \left\lfloor \frac{\mathbb{E}[M_S]}{N_{RAR}} \right\rfloor \left\lfloor \frac{N_{RAR}L_2}{L_{DL}} \right\rfloor + \left\lfloor \frac{\left( \mathbb{E}[M_S] - N_{RAR} \left\lfloor \frac{\mathbb{E}[M_S]}{N_{RAR}} \right\rfloor \right) L_2}{L_{DL}} \right\rfloor. \quad (5.32)$$

Let  $U_{d1,UL}$  be the uplink resource efficiency for the data in MSG1.  $M_S$  of devices will send their preamble without collision, and  $p$  of them will be decodable in BS. Thus,  $U_{d1,UL}$  can be obtained as

$$\mathbb{E}[U_{d1,UL}] = \frac{(1-p)\mathbb{E}[M_S]}{\mathbb{E}[B_{d1,UL}]}. \quad (5.33)$$

#### 5.4.4 Data in MSG3

For the data in MSG3, the BS preallocates 6 RBs in uplink band for step 1 in every time slot. Let  $S_{d3,1}$  be the number of preambles arrived to the BS in a RACH. In a RACH, the BS can decode multiple preambles, which include successful and collided preambles. Thus, the expectation of  $S_{d3,1}$  is equal to

$$\mathbb{E}[S_{d3,1}] = \mathbb{E}[M_S] + \mathbb{E}[M_C]. \quad (5.34)$$

The BS will response for each decoded preamble using multiple MSG2, where an MSG2 includes multiple UL grants for multiple preambles, where an UL grant is corresponding to a preamble. Let  $S_{d3,2}$  be the number of MSG2s arrived to the device. The BS responses all decoded preambles. Thus, the expectation of  $S_{d3,2}$  is equal to

$$\mathbb{E}[S_{d3,2}] = \mathbb{E}[S_{d3,1}] = \mathbb{E}[M_S] + \mathbb{E}[M_C]. \quad (5.35)$$

The devices received RA response will send their data with MSG3. For each transmission for the data and MSG3, HARQ will cause additional

allocations of RBs. Let  $S_{d3,3}$  be the number of MSG3s arrived to the BS. An MSG3 can be decoded in the BS if the MSG3 is transmitted for the successful preamble. For each successful preambles, the MSG3 and data arrive to the BS if HARQ is successful. Thus, the expectation of  $S_{d3,3}$  is equal to

$$\mathbb{E}[S_{d3,3}] = \mathbb{E}[M_S] (1 - p^N). \quad (5.36)$$

Let  $F_{d3,3}$  be the number of transmissions for MSG3 and data which fail to decode in BS. For the collided preambles, each device transmits  $N$  times of MSG3 and data. For the successful preambles, each device transmits additional MSG3 and data until success or reaching to  $N$ . Therefore, the expectation of  $F_{d3,3}$  is equal to

$$\mathbb{E}[F_{d3,3}] = \mathbb{E}[M_C]N + \mathbb{E}[M_S] \left\{ Np^N + \sum_{n=1}^{N-1} (n-1)p^n(1-p) \right\}. \quad (5.37)$$

The transmission of MSG4 is similar to the MSG3 and data. Let  $S_{d3,4}$  be the number of MSG4 arrived to the device. The BS transmits MSG4 for each successful arrival of MSG3 and data. Thus, the expectation for the  $S_{d3,4}$  is equal to

$$\mathbb{E}[S_{d3,4}] = \mathbb{E}[S_{d3,3}] (1 - p^N). \quad (5.38)$$

Let  $F_{d3,4}$  be the number of transmissions for MSG4 which fail to decode in BS, which is equal to

$$\mathbb{E}[F_{d3,4}] = \mathbb{E}[S_{d3,3}] \left\{ Np^N + \sum_{n=1}^{N-1} (n-1)p^n(1-p) \right\}. \quad (5.39)$$

For the data in MSG3, let  $B_{d3,1}$ ,  $B_{d3,2}$ ,  $B_{d3,3}$ , and  $B_{d3,4}$  be the average number of allocated RBs for the transmission of preambles, MSG2, MSG3

and data, and MSG4 for a time slot, respectively. For the transmission of preambles, the BS allocate 6 RBs in LTE-A, thus  $B_{d3,1} = 6$ . The  $N_{RAR}$  of MSG2 can be concatenated for a PDCCH. For a given  $S_{d3,2}$ ,  $\left\lfloor \frac{S_{d3,2}}{N_{RAR}} \right\rfloor$  of MSG2s are share each PDCCH. Remaining  $\left( S_{d3,2} - N_{RAR} \left\lfloor \frac{S_{d3,2}}{N_{RAR}} \right\rfloor \right)$  MSG2s are share 1 PDCCH. Thus, the expectation for  $B_{d3,2}$  is equal to

$$\mathbb{E}[B_{d3,2}] = \left\lfloor \frac{\mathbb{E}[S_{d3,2}]}{N_{RAR}} \right\rfloor \left\lfloor \frac{N_{RAR} L_2}{L_{DL}} \right\rfloor + \left\lfloor \frac{\left( \mathbb{E}[S_{d3,2}] - N_{RAR} \left\lfloor \frac{\mathbb{E}[S_{d3,2}]}{N_{RAR}} \right\rfloor \right) L_2}{L_{DL}} \right\rfloor. \quad (5.40)$$

For each transmission of MSG3 and data, which includes success and failure, the BS allocates  $\left\lfloor \frac{L_3+l}{L_{UL}} \right\rfloor$  RBs. Thus, the expectation for  $B_{d3,3}$  is equal to

$$\mathbb{E}[B_{d3,3}] = \sum_{l=L_{\min}}^{L_{\max}} (\mathbb{E}[S_{d3,3}] + \mathbb{E}[F_{d3,3}]) \left\lfloor \frac{L_3 + l}{L_{UL}} \right\rfloor g(l). \quad (5.41)$$

For each transmission of MSG4, which includes success and failure, the BS allocates  $\left\lfloor \frac{L_4}{L_{UL}} \right\rfloor$  RBs. Thus, the expectation for  $B_{d3,4}$  is equal to

$$\mathbb{E}[B_{d3,4}] = (\mathbb{E}[S_{d3,4}] + \mathbb{E}[F_{d3,4}]) \left\lfloor \frac{L_4}{L_{DL}} \right\rfloor. \quad (5.42)$$

Let  $B_{d3,UL}$  and  $B_{d3,DL}$  be the total number of allocated RBs for a time slot in uplink and downlink band, respectively. The expectation of  $B_{d3,UL}$  is the sum of the expectations of  $B_{d3,1}$  and  $B_{d3,3}$ . The expectation of  $B_{d3p,DL}$  is the sum of the expectations of  $B_{d3,2}$  and  $B_{d3,4}$ .

$$\mathbb{E}[B_{d3,UL}] = \mathbb{E}[B_{d3,1}] + \mathbb{E}[B_{d3,3}], \quad (5.43)$$

$$\mathbb{E}[B_{d3,DL}] = \mathbb{E}[B_{d3,2}] + \mathbb{E}[B_{d3,4}]. \quad (5.44)$$

Let  $U_{d3,UL}$  be the uplink resource efficiency for the data in MSG3. Similar to the CP or UP solution,  $U_{d3,UL}$  can be obtained as

$$\mathbb{E}[U_{d3,UL}] = \frac{\mathbb{E}[S_{d3,4}]}{\mathbb{E}[B_{d3,UL}]}.$$
 (5.45)

## 5.5. Proposed short data transmission procedure for IoT devices

The four SDT procedures have their advantages, but also can waste the resources. If the data in MSG1 or the data in MSG3 is used, a number of RBs will be allocated for the transmission of messages or datas which will be collided. In addition, the data in MSG1 also requires the allocation of RBs which replace the idle preambles. The CP solution or UP solution does not allocate the RBs for the transmission of colliding data, but it does the RRC connection which is the overhead for the transmission of data. Due to the characteristic of the traffic from IoT devices, the RRC connection for a IoT device is hard to maintain.

This dissertation thus proposes the exchange of signaling messages between the device and the BS for the SDT. The proposed random access procedure focuses on removing unnecessary resource allocation for unused and collided preambles. In addition, the proposed random access procedure bypass the RRC connection to reduce overhead. The proposed random access procedure is given in Figure 5.6. The description for each step is as follows:

- **Preallocation:** The BS allocates 6 RBs for preamble transmission

in every RACH periodicity.

- **Step 1 (Preamble):** The device transmits one of preamble from the preamble pool in the RACH.
- **Step 2 (RAR):** When the BS can detect the preamble, the BS responses with RAR which includes the resource allocation information for step 3.
- **Step 3 (Alternative MSG3):** If the device receives RAR corresponding to the transmitted preambles, it transmits "small data transmission request (SDT request)" rather than transmitting RRC connection request for this step. This message should include the identification (ID) of device to identify the device in BS. The small data transmission request message can be additionally defined, or the RRC request message with a indicator bit can be used. The hybrid automatic repeat request (HARQ) can be used for the reliability.
- **Step 4 (Alternative MSG4):** When the BS detects a SDT request, the BS allocates the resource for the transmission of data. The ID of device and the information of allocated RBs are then transmitted to the device. The ID is used for the contention resolution.
- **Step 5 (Data):** If the device receives resource allocation with the ID of itself, the device transmits data to the BS. The HARQ can be used for the reliability.

Note that the security setup is also done in initial attach procedure [71]. Thus, the encryption and integration for the data is possible in the NAS layer in UE.

## 5.6. Numerical evaluation for the proposed SDT procedure

Let  $L'_3$  and  $L'_4$  be the size of SDT request and the size of message in step 4. Let  $B_{prop,1}$ ,  $B_{prop,2}$ ,  $B_{prop,3}$ ,  $B_{prop,4}$ , and  $B_{prop,5}$  be the average number of allocated RBs for the transmission of preambles, MSG2, SDT request, the size of message in step 4, and data, respectively.

For the step 1 and step 2, the proposed SDT procedure consumes same RBs as in CP solution. Thus,  $B_{prop,1} = 6$  and

$$\mathbb{E}[B_{prop,2}] = \left\lfloor \frac{\mathbb{E}[S_{cp,2}]}{N_{RAR}} \right\rfloor \left\lceil \frac{N_{RAR} L_2}{L_{DL}} \right\rceil + \left\lceil \frac{\left( \mathbb{E}[S_{cp,2}] - N_{RAR} \left\lfloor \frac{\mathbb{E}[S_{cp,2}]}{N_{RAR}} \right\rfloor \right) L_2}{L_{DL}} \right\rceil. \quad (5.46)$$

For the step 3, 4, and 5, the size of message is changed but their number of transmission is same with CP solution. Therefore,  $B_{prop,3}$ ,  $B_{prop,4}$ , and  $B_{prop,5}$  can be obtained as follows:

$$\mathbb{E}[B_{prop,3}] = (\mathbb{E}[S_{cp,3}] + \mathbb{E}[F_{cp,3}]) \left\lceil \frac{L'_3}{L_{UL}} \right\rceil, \quad (5.47)$$

$$\mathbb{E}[B_{prop,4}] = (\mathbb{E}[S_{cp,4}] + \mathbb{E}[F_{cp,4}]) \left\lceil \frac{L'_4}{L_{DL}} \right\rceil, \quad (5.48)$$

$$\mathbb{E}[B_{prop,5}] = \sum_{l=L_{\min}}^{L_{\max}} (\mathbb{E}[S_{cp,5}] + \mathbb{E}[F_{cp,5}]) \left\lceil \frac{l}{L_{UL}} \right\rceil g(l). \quad (5.49)$$

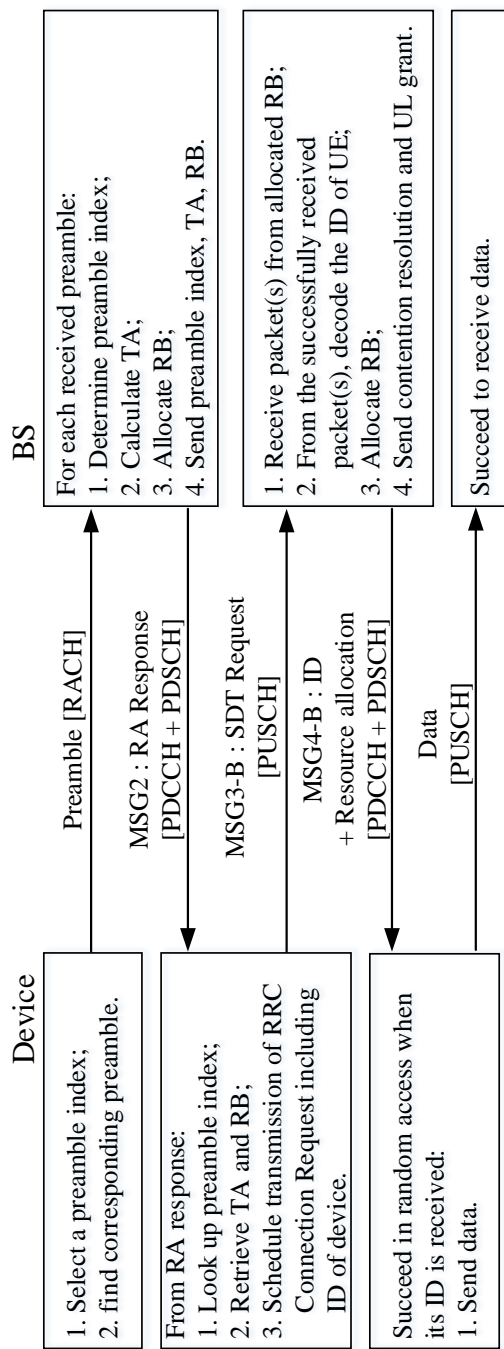


Fig. 5.6. The proposed SDT procedure

Let  $B_{prop,UL}$  and  $B_{prop,DL}$  be the total number of allocated RBs for a time slot in uplink and downlink band, respectively. The expectation of  $B_{cp,UL}$  is the sum of the expectations of  $B_{prop,1}$ ,  $B_{prop,3}$ , and  $B_{prop,5}$ . The expectation of  $B_{cp,DL}$  is the sum of the expectations of  $B_{prop,2}$ ,  $B_{prop,4}$ .

$$\mathbb{E}[B_{prop,UL}] = \mathbb{E}[B_{prop,1}] + \mathbb{E}[B_{prop,3}] + \mathbb{E}[B_{prop,5}], \quad (5.50)$$

$$\mathbb{E}[B_{prop,DL}] = \mathbb{E}[B_{prop,2}] + \mathbb{E}[B_{prop,4}]. \quad (5.51)$$

Let  $U_{prop,UL}$  be the uplink resource efficiency for the proposed SDT procedure. Since the expectation for the number of MTC devices that succeeded the transmission of data is  $\mathbb{E}[S_{prop,5}] = \mathbb{E}[S_{cp,5}]$ ,  $U_{cp,UL}$  can be obtained as

$$\mathbb{E}[U_{prop,UL}] = \frac{\mathbb{E}[S_{cp,5}]}{\mathbb{E}[B_{prop,UL}]}. \quad (5.52)$$

## 5.7. Performance Evaluation

In this section, the resource usages of conventional SDT procedures and that of the proposed SDT procedure are compared. For the comparison of the proposed SDT procedure and conventional SDT procedures, the parameters are selected as shown in Table 5.1. The maximum number of transmission for HARQ is selected as 5 as given in [8]. The decoding failure probability is equal to 10% as given in [8]. The number of preambles is selected as 54 [8]. This dissertation assumes that a RB consists of 12 sub-carriers and 7 symbols, thus there are total 84 symbols in a RB. Considering the modulation schemes, the number of bits carried by a RB



is equal to selected as 84, 168 bits when the modulation schemes are binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), respectively. The compared metrics are the average number of allocated resources in uplink channel and the resource efficiency in uplink channel. Note that the average number of allocated resources in downlink channel and resource efficiency in downlink channel are excluded since  $L_4$  can be vary dependent on the system [20, 26, 73]. The sizes of MSG3 and MSG5 are selected as 12 and 14 bytes, respectively [25].

Table 5.2 shows the assumptions for the size of data used in the performance evaluation of cellular IoT (CIoT) devices in 3GPP [14]. The data includes application layer payload, constrained application protocol (CoAP) header, datagram transport layer security (DTLS) header, user datagram protocol (UDP) header, and Internet protocol (IP) header. This dissertation assumes that the IP header compression is used to reduce the size of header. The size of application layer payload is from 20 bytes to 200 bytes. The size of application layer payload follows the Pareto distribution with  $L_{\min} = 20$  bytes,  $\alpha = 2.5$ , and cut off of 200 bytes. Let  $g(l)$  be the probability mass function for the size of application layer payload.  $g(l)$  is equal to

$$g(l) = \begin{cases} 0 & ; l < L_{\min}, \\ \frac{\alpha L_{\min}^{\alpha}}{l^{\alpha+1}} & ; L_{\min} \leq l \leq L_{\max}, \\ 0 & ; l > L_{\max}. \end{cases} \quad (5.53)$$

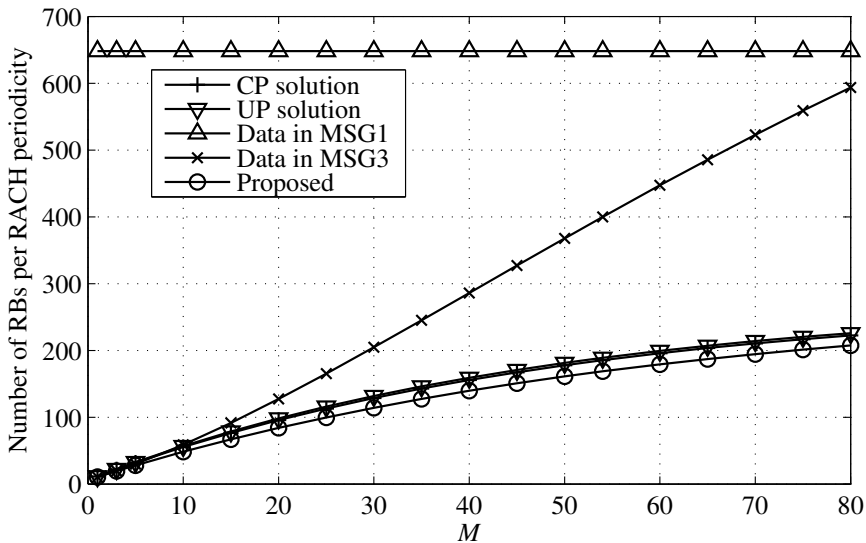
Figure 5.7 shows the number of allocated RBs per RACH periodicity when BPSK is used for modulation. The data in MSG1 requires the allocation of fixed number of RBs for the data transmission without pream-

**Table 5.1.** Parameters for evaluation of data in MSG5

| Parameter  | Definition                              | Values                  |
|------------|---|-------------------------|
| $N$        | Maximum number of transmission for HARQ | 5                       |
| $p$        | decoding failure probability            | 10%                     |
| $R$        | the number of preambles                 | 54                      |
| $L_3, L_5$ | Size of MSG3 and MSG5                   | 12, 14 bytes [25]       |
| $L_{RB}$   | Number of bits per RB                   | 84 (BPSK)<br>168 (QPSK) |

**Table 5.2.** Assumptions for the size of data,  $L_D$ 

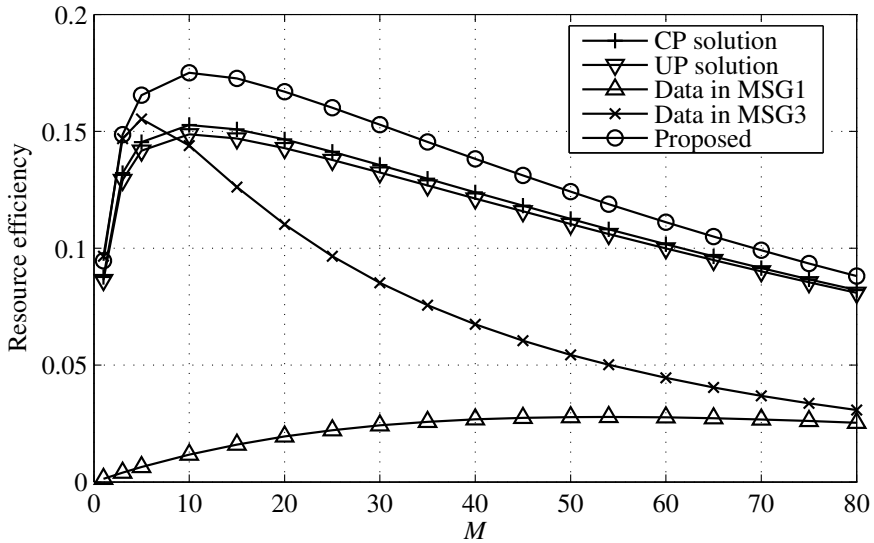
| Contents    | Size in bytes   |
|-------------|---|
| Payload     | Pareto Distribution<br>( $\alpha = 2.5, L_{\min} = 20$ bytes, $L_{\max} = 200$ bytes) |
| CoAP header | 4 bytes   |
| DTLS header | 13 bytes  |
| UDP header  | 8 bytes   |
| IP header   | 4 bytes (IP header compression)   |
| MAC header  | 3 bytes   |
| Total size  | 52-232 bytes  |



**Fig. 5.7.** Average number of allocated resources in uplink vs.  $M$  for BPSK

bles, thus it shows largest number of allocated RBs. The data in MSG3 shows lowest number of allocated RBs with small  $M$ , but it requires more resources compared to CP solution, UP solution, and proposed procedure as  $M$  increases. The CP and UP solution show similar resource usages. The proposed solution shows lowest resource usage compared to other SDT procedures.

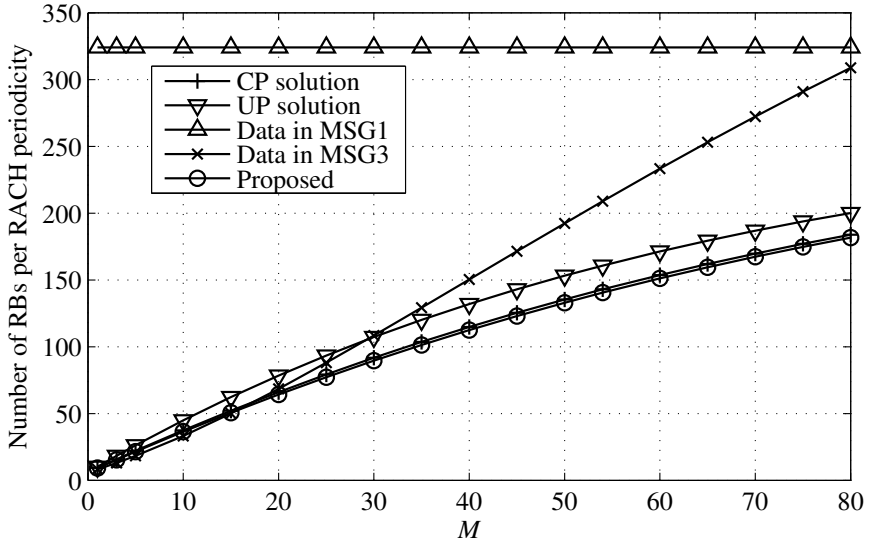
Figure 5.8 shows the uplink resource efficiency when BPSK is used for the modulation. The expected success probability for data transmission is lowest for the data in MSG1, and highest for the data in MSG3. In addition, the expected success probabilities for data transmission for all procedures except the data in MSG1 are similar. The data in MSG1 shows worst uplink resource efficiency due to largest resource usage and



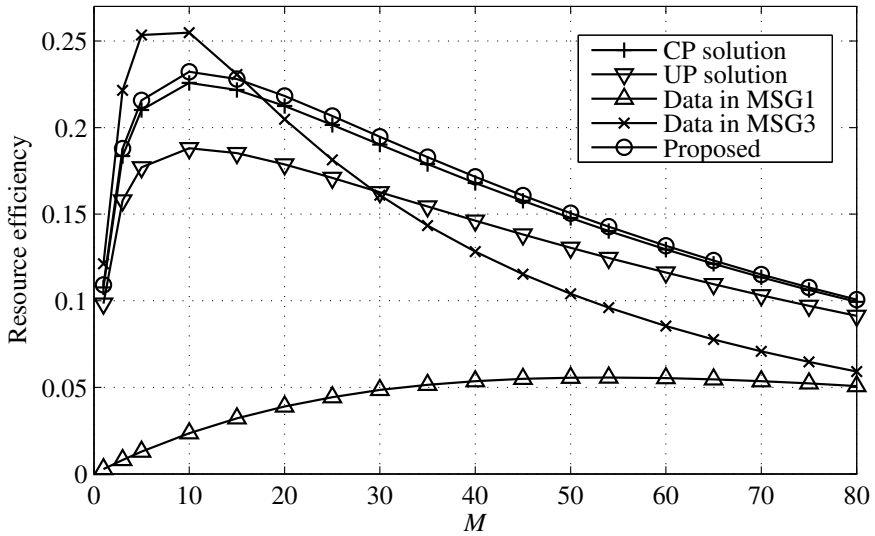
**Fig. 5.8.** Uplink resource efficiency vs.  $M$  for BPSK

lowest success probability. The proposed SDT procedure shows best resource efficiency since it uses lowest number of resources.

Figure 5.9 shows the number of allocated RBs per RACH periodicity when QPSK is used for modulation. Figure 5.10 shows the uplink resource efficiency with QPSK. The single RB is sufficient to transmit any message except large data when QPSK is used for modulation, thus the number of RBs per RACH periodicity is decreased. The resource efficiency of the data in MSG3 is best with small  $M$  since the effect from the number of message exchanges between device and BS is increased in QPSK. Even though the increment is reduced, the proposed SDT procedure shows best resource efficiency when  $M$  is above 18.



**Fig. 5.9.** Average number of allocated resources in uplink vs.  $M$  for QPSK



**Fig. 5.10.** Uplink resource efficiency vs.  $M$  for QPSK

## 5.8. Summary

In this chapter, the numerical analysis for conventional SDT procedures is performed where the analysis focus on the number of allocated resources per RACH. In addition, this dissertation proposes a SDT procedure which bypasses the RRC connection to reduce resource usage. The performance evaluations show that the CP and UP solution shows higher resource efficiency compared with the data in MSG1 and the data in MSG3 in most case. The performance evaluations also show that the proposed SDT procedure increases the resource efficiency of SDT procedure. Therefore, the proposed SDT procedure can improve the resource efficiency of RACH procedure for massive number of devices.

## Chapter 6. Conclusion

In this dissertation, the adaptive resource allocation protocols and the efficient SDT procedure are presented to alleviate traffic congestion and to provide resource efficiency for the massive number of devices in mobile network. First, a DARR protocol with preamble partition approach is proposed. The challenge of the decrease of throughput in DARR was discussed which is caused by periodicity to change information related to the RACH. The numerical evaluation in this dissertation showed that the fluctuation of the number of contending devices decreases the throughput. To resolve the fluctuation problem, this dissertation proposed a DARR protocol with preamble partition approach. The proposed protocol separately allocates the preambles based on the number of backlogs of devices. The devices select the allocated preambles according to their experienced number of backlogs. The performance evaluation shows that the proposed protocol increases throughput compared to that of the system without the preamble partition approach. For 100,000 devices in a cell, throughput is increased by 29.7%-114.4% and 23.0%-91.3% with uniform and Beta distributed arrivals of devices, respectively. Based on the evaluation results, this dissertation can conclude that the proposed

preamble partition based DARR protocol can be used to improve the throughput of the RACH procedure or the data transmission schemes based on a frame slotted or a multi-channel ALOHA.

Second, a DARR and ACB protocol with preamble partition and stochastic gradient descent approach is proposed. The numerical evaluation expects the fluctuation problem from the DARR, and the convergence of throughput in the ACB. Based on the numerical evaluation, an adaptive DARR and ACB protocol is proposed to improve the throughput in both low and high traffic loads. The performance evaluation results show that the proposed adaptive DARR and ACB protocol changes the ACB factor and the pool size according to the number of contending devices in the network regardless to the number of devices in the cell or the si-periodicity. The proposed algorithm shows the average throughput which is from 92.32% to 97.25% of that with ideal DARR and ACB algorithm. Based on the evaluation results, this dissertation can conclude that the proposed preamble partition based DARR and ACB protocol can be used to improve the throughput of the RACH procedure or the data transmission schemes based on a frame slotted or a multi-channel ALOHA.

Third, the efficient SDT procedure for IoT device is proposed. This dissertation has evaluated the number of required resource blocks and the uplink resource efficiency for conventional SDT procedures. Based on the evaluation of conventional SDT procedures, this dissertation also proposed a random access procedure which does five message exchanges between a device and a BS. The proposed procedure excludes the RRC



connection for SDT thus the overhead can be reduced. The numerical evaluation shows that the proposed SDT procedure can decrease the resource usage and can increase the number of successful devices per resource block in most cases.

Based on the evaluation in this dissertation, an MAC protocol which includes the proposed SDT procedure and the proposed adaptive DARR protocol can improve the throughput and resource efficiency when the mobile network system requires SDT procedures and DARR. In addition, an MAC protocol which includes the proposed SDT procedure and the proposed adaptive DARR and ACB protocol can improve the throughput and resource efficiency when the mobile network system requires SDT procedure and both DARR and ACB. Therefore, the throughput and resource efficiency for massive devices in mobile network can be increased if one of MAC protocol is applied in the RAN. Since the number of IoT devices will increase dramatically in the future, it is expected that the proposed adaptive MAC protocols can be applied to support massive number of devices in mobile networks more efficiently.

## References

- [1] 3GPP TR 22.888, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on enhancements for Machine-Type Communications (MTC) (Release 12),” March 2013. [Cited on page 1.]
- [2] A. Laya, L. Alonso, and J. Alonso-Zarate, “Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives,” *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 4–16, First 2014. [Cited on pages 10, 11, 18, 91, and 100.]
- [3] F. Ghavimi and H. H. Chen, “M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 525–549, Secondquarter 2015.
- [4] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, “D-ACB: Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9847–9861, Dec 2016. [Cited on pages 2, 4, 11, 18, 21, 23, 28, 30, 31, 33, 37, 60, 61, 63, and 77.]

- [5] 5GPPP, “5G empowering vertical industries,” February 2016. [Cited on page 1.]
- [6] Machina Research Sector Report, “Machine-to-Machine (M2M) communication in consumer electronics 2012-22,” February 2013. [Cited on page 1.]
- [7] Recommendation ITU-R M.2083-0, “IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond,” September 2015. [Cited on page 1.]
- [8] 3GPP TR 37.868 v11.0.0, “RAN Improvements for Machine-type Communications,” October 2011. [Cited on pages 2, 5, 11, 15, 17, 18, 19, 20, 21, 22, 23, 30, 32, 33, 35, 36, 50, 63, 66, 77, 93, and 118.]
- [9] ICT-317669 METIS Project Deliverable 1.1, “Scenarios, Requirements and KPIs for 5G Mobile and Wireless System,” April 2013. [Cited on page 2.]
- [10] P. Jain, P. Hedman, and H. Zisimopoulos, “Machine type communications in 3GPP systems,” *IEEE Communications Magazine*, vol. 50, no. 11, pp. 28–35, November 2012. [Cited on page 2.]
- [11] 3GPP TR 36.321 v12.6.0, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 14),” March 2017. [Cited on pages 2, 13, and 77.]
- [12] J. Choi, “On the Adaptive Determination of the Number of Preambles in RACH for MTC,” *IEEE Communications Letters*, vol. 20, no. 7, pp. 1385–1388, July 2016. [Cited on pages 3, 23, 28, 33, 34, 35, 36, 37, 50, 63, 71, and 77.]

- [13] A. G. Gotsis, A. S. Lioumpas, and A. Alexiou, "M2M Scheduling over LTE: Challenges and New Perspectives," *IEEE Vehicular Technology Magazine*, vol. 7, no. 3, pp. 34–39, Sept 2012. [Cited on page 4.]
- [14] 3GPP TR 45.820 v13.2.0, "Cellular System Support for Ultra-low Complexity and Low Throughput Internet of Things (CIoT)," November 2015. [Cited on pages 4, 36, 66, and 119.]
- [15] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: overload control," *IEEE Network*, vol. 26, no. 6, pp. 54–60, November 2012. [Cited on pages 5 and 30.]
- [16] Vina Krishnan, "LTE market share and growth in Q2 2016 revealed by OVUM," *Telecomlead*, September 2016. [Cited on page 10.]
- [17] KT TS 5G.321, "KT PyeongChang 5G Special Interest Group (KT 5G-SIG); KT 5th Generation Radio Access; Medium Access Control (MAC); Protocol specification (Release 1)," September 2016. [Cited on page 10.]
- [18] 3GPP TR 36.912, "Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced)," September 2014. [Cited on page 11.]
- [19] 3GPP TR 36.300, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 14)," March 2017. [Cited on page 90.]
- [20] "RACH." [Online]. Available: [http://www.sharetechnote.com/html/RACH\\_LTE.html](http://www.sharetechnote.com/html/RACH_LTE.html) [Cited on pages 11 and 119.]

- [21] Netmanias, “LTE EMM Procedure: 1. Initial Attach for Unknown UE (Part 2) - Call Flow of Initial Attach.” [Online]. Available: <http://www.netmanias.com/ko/post/techdocs/5320/attach-emm-lte/lte-emm-procedure-1-initial-attach-for-unknown-ue-part-2-call-flow-of-initial-attach> [Cited on page 11.]
- [22] 3GPP TR 36.211, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 14),” March 2017. [Cited on pages 13 and 15.]
- [23] H. Zepernick and A. Finger, *Pseudo Random Signal Processing: Theory and Application*, 2013. [Cited on page 13.]
- [24] 3GPP TR 36.212, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding (Release 14),” March 2017. [Cited on page 13.]
- [25] S. M. Oh and J. Shin, “An Efficient Small Data Transmission Scheme in the 3GPP NB-IoT System,” *IEEE Communications Letters*, vol. 21, no. 3, pp. 660–663, March 2017. [Cited on pages 14, 95, 100, 119, and 120.]
- [26] 3GPP TR 36.331 v13.2.0, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 14),” March 2017. [Cited on pages 15, 16, 18, 50, 77, 90, and 119.]

- [27] 3GPP TR 38.802, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio Access Technology Physical Layer Aspects (Release 14),” March 2017. [Cited on page 16.]
- [28] 3GPP TR 38.804, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio Access Technology; Radio Interface Protocol Aspects (Release 14) ,” March 2017.
- [29] “5G - Frame Structure.” [Online]. Available: [http://www.sharetechnote.com/html/5G/5G\\_FrameStructure\\_Candidate.html](http://www.sharetechnote.com/html/5G/5G_FrameStructure_Candidate.html) [Cited on page 16.]
- [30] C. Anton-Haro and M. Dohler, *Machine-to-machine (M2M) Communications: Architecture, Performance and Applications*. Elsevier, 2014. [Cited on page 18.]
- [31] T. P. C. d. Andrade, C. A. Astudillo, L. R. Sekijima, and N. L. S. d. Fonseca, “The random access procedure in long term evolution networks for the internet of things,” *IEEE Communications Magazine*, vol. 55, no. 3, pp. 124–131, March 2017. [Cited on page 18.]
- [32] S. Duan, V. Shah-Mansouri, and V. W. S. Wong, “Dynamic access class barring for M2M communications in LTE networks,” in *2013 IEEE Global Communications Conference (GLOBECOM)*, Dec 2013, pp. 4747–4752. [Cited on pages 18 and 21.]
- [33] H. He, Q. Du, H. Song, W. Li, Y. Wang, and P. Ren, “Traffic-aware ACB scheme for massive access in machine-to-machine networks,” in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 617–622. [Cited on pages 18 and 21.]

- [34] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, “Congestion control for bursty M2M traffic in LTE networks,” in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 5815–5820. [Cited on pages 19 and 21.]
- [35] J. Moon and Y. Lim, “Adaptive Access Class Barring for Machine-Type Communications in LTE-A,” in *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2016, pp. 398–402. [Cited on pages 19 and 21.]
- [36] M. Koseoglu, “Pricing-Based Load Control of M2M Traffic for the LTE-A Random Access Channel,” *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1353–1365, March 2017. [Cited on pages 19 and 21.]
- [37] R. G. Cheng, J. Chen, D. W. Chen, and C. H. Wei, “Modeling and Analysis of an Extended Access Barring Algorithm for Machine-Type Communications in LTE-A Networks,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 2956–2968, June 2015. [Cited on pages 19, 21, 50, and 77.]
- [38] T. M. Lin, C. H. Lee, J. P. Cheng, and W. T. Chen, “PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTE-A Networks,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2467–2472, Jun 2014. [Cited on pages 19 and 21.]
- [39] G. Y. Lin, S. R. Chang, and H. Y. Wei, “Estimation and Adaptation for Bursty LTE Random Access,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2560–2577, April 2016. [Cited on pages 19 and 22.]

- [40] X. Yang, A. Fapojuwo, and E. Egbogah, “Performance Analysis and Parameter Optimization of Random Access Backoff Algorithm in LTE,” in *2012 IEEE Vehicular Technology Conference (VTC Fall)*, Sept 2012, pp. 1–5. [Cited on pages 20 and 22.]
- [41] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, “Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks,” in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6. [Cited on pages 20 and 22.]
- [42] A. Lo, Y. Law, M. Jacobsson, and M. Kucharzak, “Enhanced LTE-Advanced Random-Access Mechanism for Massive Machine-to-Machine (M2M) Communications,” in *Proc. 27th Meeting of Wireless World Research Forum (WWRF)*, October 2011, pp. 1–5. [Cited on pages 20 and 28.]
- [43] H. Y. Hwang, S. M. Oh, C. Lee, J. H. Kim, and J. Shin, “Dynamic RACH preamble allocation scheme,” in *2015 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct 2015, pp. 770–772. [Cited on pages 20 and 28.]
- [44] W. Li, Q. Du, L. Liu, P. Ren, Y. Wang, and L. Sun, “Dynamic Allocation of RACH Resource for Clustered M2M Communications in LTE Networks,” in *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, Oct 2015, pp. 140–145. [Cited on pages 23 and 28.]
- [45] Q. Du, W. Li, L. Liu, P. Ren, Y. Wang, and L. Sun, “Dynamic RACH



Partition for Massive Access of Differentiated M2M Services,” *Sensors*, vol. 16, no. 4, pp. 1–19, April 2016. [Cited on page 23.]

- [46] Y. J. Choi, S. Park, and S. Bahk, “Multichannel random access in OFDMA wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 603–613, March 2006. [Cited on page 23.]
- [47] C. H. Wei, R. G. Cheng, and S. L. Tsao, “Modeling and Estimation of One-Shot Random Access for Finite-User Multichannel Slotted ALOHA Systems,” *IEEE Communications Letters*, vol. 16, no. 8, pp. 1196–1199, August 2012. [Cited on pages 23 and 28.]
- [48] B. Zhen, M. Kobayashi, and M. Shimizui, “Framed ALOHA for Multiple RFID Objects Identification,” *IEICE Transactions on Communications*, vol. E88-B, no. 3, pp. 991–999, 2005. [Cited on pages 24, 25, 27, and 28.]
- [49] G. Khandelwal, A. Yener, K. Lee, and S. Serbetli, “ASAP : A MAC Protocol for Dense and Time Constrained RFID Systems,” in *2006 IEEE International Conference on Communications*, vol. 9, June 2006, pp. 4028–4033. [Cited on pages 25, 27, 29, and 50.]
- [50] H. Vogt, “Efficient Object Identification with Passive RFID Tags,” in *PerCom 2002*, 2002. [Cited on pages 24, 25, 26, and 28.]
- [51] D. K. Klair, K. W. Chin, and R. Raad, “A Survey and Tutorial of RFID Anti-Collision Protocols,” *IEEE Communications Surveys Tutorials*, vol. 12, no. 3, pp. 400–421, Third 2010. [Cited on page 24.]
- [52] J.-R. Cha and J.-H. Kim, “Novel Anti-collision Algorithms for Fast Object Identification in RFID System,” in *11th International Conference on Par-*

*allel and Distributed Systems (ICPADS'05)*, vol. 2, July 2005, pp. 63–67.  
[Cited on pages 25, 27, 29, and 35.]

- [53] J.-R. Cha and J.-H. Kim, “Dynamic framed slotted ALOHA algorithms using fast tag estimation method for RFID system,” in *2006 3rd IEEE Consumer Communications and Networking Conference (CCNC 2006)*, vol. 2, Jan 2006, pp. 768–772. [Cited on pages 25, 27, 29, and 35.]
- [54] C.-H. L. Wen-Tzu Chen, “An Efficient Anti-Collision Method for Tag Identification in a RFID System,” *IEICE Transactions on Communications*, vol. E89-B, no. 12, pp. 3386–3392, 2006. [Cited on pages 25 and 27.]
- [55] Z. G. Prodanoff, “Optimal frame size analysis for framed slotted ALOHA based RFID networks,” *Computer Communications*, vol. 33, no. 5, pp. 648–653, 2010. [Cited on pages 26 and 29.]
- [56] D. Lee, J. Choi, and W. Lee, “OFSA: Optimum Frame-Slotted Aloha for RFID Tag Collision Arbitration,” *KSII Transactions on Internet and Information Systems*, vol. 5, no. 11, pp. 1929–1945, November 2011. [Cited on page 26.]
- [57] S. Dhakal and S. Shin, “Precise-Optimal Frame Length Based Collision Reduction Schemes for Frame Slotted Aloha RFID Systems,” *KSII Transactions on Internet and Information Systems*, vol. 8, no. 1, pp. 165–182, January 2014. [Cited on page 26.]
- [58] Y. B. Kim, “Determination of optimal frame sizes in framed slotted ALOHA,” *Electronics Letters*, vol. 50, no. 23, pp. 1764–1766, 2014. [Cited on pages 26 and 29.]

- [59] Auto-ID Center, “13.56 MHz ISM band class 1 radio frequency identification tag interface specification, version 1.0, HF RFID standard.” [Cited on pages 26 and 29.]
- [60] N. K. Pratas, H. Thomsen, C. Stefanovic, and P. Popovski, “Code-expanded random access for machine-type communications,” in *2012 IEEE Globecom Workshops*, Dec 2012, pp. 1681–1686. [Cited on pages 27 and 29.]
- [61] H. S. Jang, S. M. Kim, K. S. Ko, J. Cha, and D. K. Sung, “Spatial Group Based Random Access for M2M Communications,” *IEEE Communications Letters*, vol. 18, no. 6, pp. 961–964, June 2014. [Cited on pages 27 and 29.]
- [62] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, “Massive machine-type communications in 5G: physical and MAC-layer solutions,” *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, September 2016. [Cited on page 27.]
- [63] S.-H. Lee, S.-Y. Jung, and J.-H. Kim, “Dynamic Resource Allocation of the Random Access for MTC Devices,” *ETRI Journal*, 2017. [Cited on page 31.]
- [64] O. Arouk and A. Ksentini, “General Model for RACH Procedure Performance Analysis,” *IEEE Communications Letters*, vol. 20, no. 2, pp. 372–375, Feb 2016. [Cited on pages 35 and 102.]
- [65] F. Gravetter and L. Wallnau, *Essentials of Statistics for the Behavioral Science*. Cengage Learning, 2007. [Cited on page 37.]

- [66] M. G. Summa, L. Bottou, B. Goldfarb, C. P. Fionn Murtagh, and M. Touati, *Statistical Learning and Data Science*. CRC Press, 2012. [Cited on pages 70 and 71.]
- [67] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*. A Wiley-interscience publication, 2001. [Cited on page 71.]
- [68] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, “Toward massive machine type cellular communications,” *IEEE Wireless Communications*, vol. 24, no. 1, pp. 120–128, February 2017. [Cited on page 90.]
- [69] K. D. Lee, S. Kim, and B. Yi, “Throughput comparison of random access methods for M2M service over LTE networks,” in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, Dec 2011, pp. 373–377. [Cited on page 91.]
- [70] Y. Chen and W. Wang, “Machine-to-Machine Communication in LTE-A,” in *Vehicular Technology Conference Fall (VTC 2010-Fall)*, 2010 *IEEE 72nd*, Sept 2010, pp. 1–4. [Cited on page 91.]
- [71] 3GPP TR 24.301, “3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (Release 14),” March 2017. [Cited on pages 93 and 116.]
- [72] 3GPP TR 23.887, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements (Release 12),” December 2013. [Cited on page 95.]

[73] Qualcomm Europe, “Updating Encryption Keys For MBMS, 3GPP TSG SA WG3 Security — MBMS ad-hoc, S3z030022,” September 2003. [Cited on page 119.]