공학박사 학위논문

Enhancement of Radio Access Technologies for 5G and Beyond Wireless Networks

5G 이후 무선 네트워크를 위한 무선 접속 기술 향상 연구

2020년 8월

서울대학교 대학원

전기·정보공학부

김 준 석

Enhancement of Radio Access Technologies for 5G and Beyond Wireless Networks

지도 교수 박 세 웅

이 논문을 공학박사 학위논문으로 제출함

2020년 7월

서울대학교 대학원

전기·정보공학부

김 준 석

김준석의 공학박사 학위 논문을 인준함

2020년 6월

위 원 장: 심 병 효 (인)

부위원장: 박 세 웅 (인)

위 원: 오 정 석 (인)

위 원: 이 경 한 (인)

위 원: 백 정 엽 (인)

Abstract

Recently, operators are creating services using 5G systems in various fields, e.g., manufacturing, automotive, health care, etc. 5G use cases include transmission of small packets using IoT devices to high data rate transmission such as high-definition video streaming. When a large-scale IoT device transmits a small packet, power saving is important, so it is necessary to disconnect from the base station and then establish a connection through random access to transmit data. However, existing random access procedures are difficult to satisfy various latency requirements. It is attractive to use a wide bandwidth of the millimeter wave spectrum for high data rate transmission. In order to overcome the channel characteristics, beamforming technology is applied. However, when determining a beam pair between a transmitter and a receiver, interference is not considered.

In this dissertation, we consider the following three enhancements to enable 5G and beyond use cases: (i) Two-step random access procedure for delay-sensitive devices, (ii) self-uplink synchronization framework for solving preamble collision problem, and (iii) interference-aware beam adjustment for interference coordination.

First, *RAPID*, two-step random access for delay-sensitive devices, is proposed to reduce latency requirement value for satisfying specific reliability. When devices, performing *RAPID* and contention-based random access, coexist, it is important to determine a value that is the number of preambles for *RAPID* to reduce random access load. Simulation results show that *RAPID* achieves 99.999% reliability with 80.8% shorter uplink latency, and also decreases random access load by 30.5% compared with state-of-the-art techniques.

Second, in order to solve preamble collision problem, we develop self-uplink synchronization framework called EsTA. Preamble collision occurs when multiple devices transmit the same preamble. Specifically, we propose a framework that helps the UE to estimate the timing advance command using a deep neural network model and to determine the TA value. Estimation accuracy can achieve 98–99% when subcarrier spacing is 30 and 60 kHz.

Finally, we propose IBA, which is interference-aware beam adjustment method to reduce interference in millimeter wave networks. Unlike existing methods of reducing interference by scheduling time and frequency resources differently, interference is controlled through beam adjustment. In IBA, it is important to reduce search space of finding new beam pair to reduce interference. In practical, it is impossible to search beam pair of all combinations. Therefore, through Monte Carlo method, we can reduce search space to achieve local optimum. IBA achieve enhancement of lower 50% throughput up to 50% compared with only applying beam adjustment.

In summary, we propose a two-step random access, a self-uplink synchronization framework, and interference-aware beam adjustment for 5G and beyond use cases. Through these researches, we achieve enhancements of network performance such as latency and throughput compared with state-of-the-art techniques.

keywords: Fifth-generation (5G), random access, load analysis, supervised learning, millimeter wave, network simulator-3, and beam management.

student number: 2014-21637

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 5G Vision, Applications, and Keywords

One of the main challenges for operators is how to address new growth opportunities in the highly competitive telecommunications market. An example of growth opportunity is to pursue industrial digitization through 5th generation (5G) systems. Recently, operators are making services using 5G systems in various field, e.g., manufacturing, automotive, healthcare, etc. As shown in Fig. 1.1, services such as smart construction site and smart factory are have been launched by integrating information technology (IT) systems with traditional industries such as manufacturing and construction. The connected car is also a use case where operators are investing heavily due to the development of self-driving cars. In this use case, the car connected to the network provides various services, e.g., infotainment and high density (HD) 3-dimensional (3D) map.

In smart industry such as smart construction site, a number of Internet of things (IoT) devices are expected to improve productivity and safety. With the increasing number of IoT devices, machine-type communication (MTC) has become an important use case of 5G systems. Since MTC devices are mostly disconnected from next-generation nodeB (gNB) for power saving, random access procedure is required for

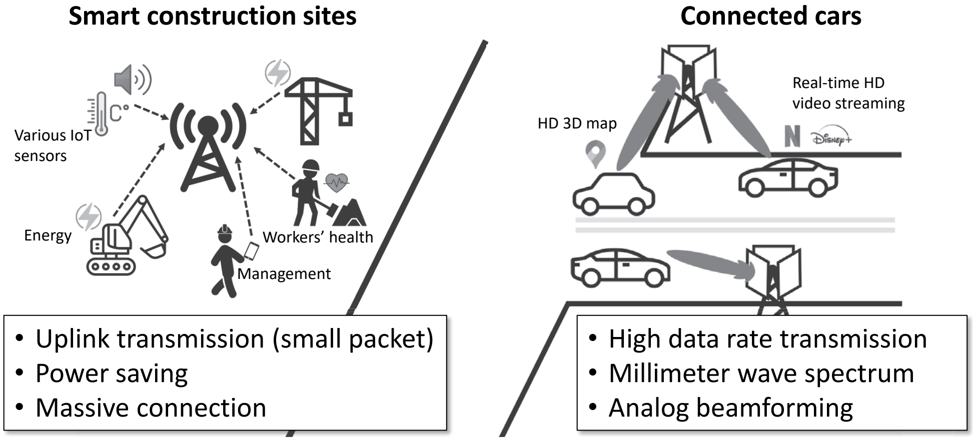


Figure 1.1: Representative examples of 5G services and keywords.

devices to transmit data. If many devices try random access simultaneously, preamble collision problem occurs, thus causing latency increase. In an environment where delay-sensitive and delay-tolerant devices coexist, the contention-based random access procedure cannot satisfy latency requirements of delay-sensitive devices. Firstly, therefore, we propose *RAPID*, a novel random access procedure, which is completed through two message exchanges for the delay-sensitive devices.

Meanwhile, after finalizing Release 15 specifications for the 5G new radio (NR) in June 2018, the 3GPP worked on not only technical improvements over the previous release but also the introduction of new features in Release 16. One of the new features is the use of two-step random access channel (2-step RACH) that enhances contention-based random access with respect to radio resource control connection setup and resume procedures. We need to look into details of 2-step random access defined in 3GPP Release 16. We also briefly introduce recent literature related to 2-step random access. Among the challenges derived from the above random access schemes, we focus on how a user equipment (UE) performs self-uplink synchronization with the gNB to resolve preamble collisions, which occur when multiple UEs transmit the same preamble. Specifically, we propose a framework that helps the UE to estimate the timing advance (TA) command using a deep neural network (DNN) model and to determine the TA value.

HD 3D map and real time HD video streaming for connected cars are require high data rate transmission. 5G NR utilize millimeter Wave (mmWave) spectrum providing wide bandwidth to support these services. At such high frequencies compared to sub-6 GHz, propagation properties are different, with less diffraction, higher penetration losses, and in general higher path losses. This can be compensated for by having more antenna elements both at the transmitter and receiver, to be used for narrower antenna beams. In order to overcome degradation, analog beamforming is introduced, which forms a beam at specific locations through antenna processing in the analog domain.

The ultimate task of beam management is, under these conditions, to establish and retain a suitable beam pair, that is, a transmitter-side beam direction and a corresponding receiver-side beam direction that jointly provide good connectivity. However, the current beam management determines the beam pair considering only the link between the transmitter and the receiver. Therefore, we propose a new beam adjustment method, which control beam pair after beam adjustment to coordinate interference.

1.2 Overview of Existing Approach

A number of random access procedures for MTC devices have been studied [1]. The authors in [2] proposed prioritized random access with dynamic access barring to support various quality-of-service (QoS). In [3], the authors proposed a new random access scheme considering characteristics of RACH structure for MTC devices defined in 3GPP. These approaches decrease latency for random access by reducing preamble collision probability during random access. Although two-step random access specified in 3GPP [4] reduces the procedures from four to two, there is still preamble collision problem when the number of UEs trying to random access increases.

In general, TA estimation for a UE is performed at gNB based on the time of arrival of random access preamble. The authors in [5] proposed spatial averaging-based TA estimation at gNB to support initial random access in high user density scenarios. In 5G NR, accurate knowledge is a key requirement for services such as emergency and autonomous driving. Therefore, 3GPP specified positioning support for 5G NR, and several solutions are provided [6]. However, the proposed solutions use addition reference signals for measuring time difference of arrival of signals.

As the mmWave network becomes denser, co-channel interference becomes a factor limiting performance [7]. To overcome this, many studies have been conducted to overcome interference through scheduling by dividing frequency resources or time resources [7], [8]. However, since this approach uses time-frequency resources separately, it may not be possible to obtain the maximum throughput performance. In addition, since most of the previous studies were assumed to be omni-directional receivers, the environment, which is affected by interference, was artificially created.

1.3 Main Contributions

1.3.1 RAPID: Two-Step Random Access

We propose a contention resolution-based random access (*RAPID*), which is complete random access procedures using two messages.

Contributions of *RAPID* are summarized as follows:

- We propose a new random access procedure, *RAPID*, for delay-sensitive UEs to reduce the uplink latency.
- We develop access pattern analyzer (APA) which predicts traffic characteristics of UEs to efficiently use radio resources while UEs perform *RAPID* procedure.
- Markov chain model is developed to analyze random access load of random access procedures. We also develop an optimization problem to find the optimal number of preambles for *RAPID* through the analysis.

- We evaluate latency and random access load of *RAPID* through system-level simulation, and validate that the proposed scheme outperforms state-of-the-art technologies.

With these contributions, the *RAPID* can be used for delay-sensitive MTC devices.

1.3.2 EsTA: Self-Uplink Synchronization

To help a UE to determine its own TA value, we propose EsTA, a framework helps a UE to determine its own TA value. Specifically, we design EsTA to achieve the following goals: (a) it estimate the location of UE coarsely (TA command) to determine TA value. (b) it estimate the TA command by only using reference signal received power without other contexts.

Two key contributions of this chapter can be summarized as follows:

- We tackle the preamble collision problem cause by recent 2-step random access schemes.
- We propose a framework for TA command estimation using a DNN model and TA value determination for each UE.

1.3.3 IBA: Interference-Aware Beam Fine Adjustment

To coordinate interference in 5G mmWave networks, we propose IBA, a interference-aware beam adjustment.

We claim the following major contributions:

- We propose a interference-aware beam adjustment to coordinate interference in 5G mmWave networks.
- We reduce search space of beam pairs for IBA so that the additional adjustment is done as soon as possible.

1.4 Organization of the Dissertation

The rest of the dissertation is organized as follows.

Chapter 2 presents *RAPID*, contention resolution-based random access procedures using context ID. *RAPID* completes the random access procedure by exchanging two messages using access stratum (AS) context ID of UE. We then develop an optimization problem to obtain the number of preambles for RAPID based on random access load analysis. Next, we provide simulations and mathematical analysis considering mMTC devices,and demonstrate that RAPID can support delay-sensitive UEs by satisfying more strict latency requirement.

In Chapter 3, we present EsTA, a self-uplink synchronization framework. First, we introduce not only 2-step random access defined 3GPP but also recent literature to tackle the preamble collision problem. Next, we describe the self-uplink synchronization framework that allow a UE to determine it TA value using DNN model.

Chapter 4 presents IBA, a beam adjustment method that coordinates interference in 5G mmWave networks. First, we present overall procedures of IBA and challenges in perspective of search space size of beam pairs. We then look at how to reduce the search space of beam pairs during IBA. Next, we provide simulation results and demonstrate that IBA can increase throughput by reducing interference.

Finally, Chapter 5 concludes the dissertation with the summary of contributions and discussion on the future work.

Chapter 2

RAPID: Contention Resolution-based Random Access Procedure using Context ID for IoT

2.1 Introduction

In recent years, 3GPP mobile communication systems have evolved in a different direction than ever before in order to provide services for MTC devices [9]. With the development of various services such as eHealth, smart city, and smart factory, the number of MTC connections is expected to grow to 3.3 billion by 2021 [10]. In light of this prediction, 3GPP specifies mMTC as a new use case of the 5G communication systems [11].

MTC devices are battery powered and may be located out of people's reach. Therefore, reducing power consumption of MTC devices is essential to extend their lifetime. In long term evolution-advanced (LTE-A) system, if a UE does not perform any operation for certain amount of time, gNB releases connection with the UE, i.e., releases radio resource control (RRC) connection. The connection between gNB and core network for the UE is also released. This state is called RRC_IDLE state. The UE in RRC_IDLE state transits to RRC_CONNECTED state to transmit or receive data, by involving a number of message exchanges. However, such transition incurs large ran-

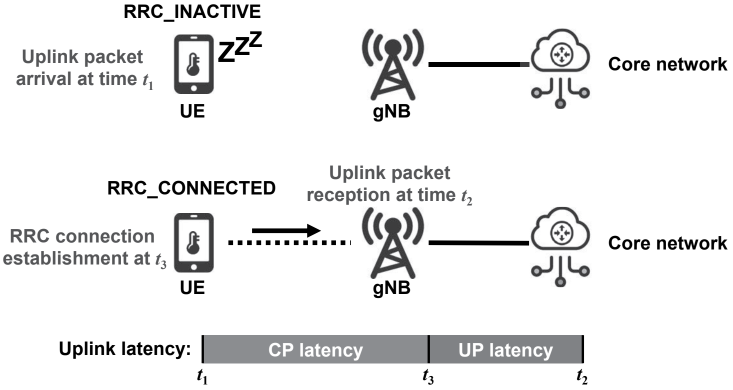

RRC_INACTIVE
Uplink packet arrival at time t_1
UE
gNB
Core network
RRC_CONNECTED
Uplink packet reception at time t_2
RRC connection establishment at t_3
UE
gNB
Core network
Uplink latency:
CP latency
UP latency
t_1
t_3
t_2

Figure 2.1: Uplink packet transmission in RRC_INACTIVE state and its latency components.

dom access load for MTC UEs, which frequently transmit small size packets. Therefore, a new RRC state, i.e., RRC_INACTIVE state, is proposed as a primary sleeping state prior to RRC_IDLE state [12, 13], and is included in the 3GPP standards [14].

Fig. 2.1 shows uplink packet transmission when a UE is in RRC_INACTIVE state. At time t_1, an uplink packet is generated in the UE. Because the UE is in RRC_INACTIVE state, random access procedure is needed to establish a RRC connection. If gNB receives the packet successfully at time t_2, uplink latency[1] is simply computed as $t_2 - t_1$. We can divide uplink latency into two parts. Control plane (CP) latency is the amount of time to transit from RRC_INACTIVE to RRC_CONNECTED state. Therefore, t_3 is the time when random access procedure is completed. user plane (UP) latency is the amount of time required for transmitting packets when the UE is active, i.e., in RRC_CONNECTED state. The contention-based random access defined in LTE-A increases CP latency as the number of UEs increases. This is because many UEs simultaneously attempt contention-based random access, thus resulting in a preamble collision problem.

[1] In this paper, we observe the uplink latency only in the radio access network (RAN), i.e., between UE and gNB.

In this paper, we propose *RAPID*, a novel random access procedure to support delay-sensitive UEs by reducing CP latency. In addition, we develop APA that works in the RRC layer. APA determines traffic characteristics of MTC UEs for gNB to efficiently use radio resources during *RAPID* operation.

The key idea beneath *RAPID* is to reduce the number of message exchanges from four to two. One of the components to achieve this purpose is allocating preambles for *RAPID* by decreasing preambles for contention-based random access.[2] For each procedure, random access load needed for random access increases according to the decrease of the number of preambles for each random access. For this reason, when UEs performing contention-based random access or *RAPID* coexist, it is important to determine the number of preambles for *RAPID*. Therefore, we analyze random access load of random access procedures, i.e., contention-based random access and *RAPID*, using a Markov chain model to determine the optimal number of preambles for *RAPID* (or preambles for contention-based random access).

In this chapter, we claim the following four major contributions.

- We propose a new random access procedure, *RAPID*, for delay-sensitive UEs in RRC_INACTIVE state to reduce the uplink latency.
- We develop APA which predicts traffic characteristics of UEs to efficiently use radio resources while UEs perform *RAPID* procedure.
- Markov chain model is developed to analyze random access load of random access procedures. We also develop an optimization problem to find the optimal number of preambles for *RAPID* through the analysis.
- We evaluate latency and random access load of *RAPID* through system-level simulation, and validate that the proposed scheme outperforms state-of-the-art technologies.

[2] The sum of the number of preambles for *RAPID* and contention-based random access is fixed.

The rest of the chapter is structured as follows. In Section 2.2, we introduce the RRC states, random access procedure in LTE-A, and uplink latency addressed in this chapter. We also discuss the related work. *RAPID* and APA are detailed in Section 2.3 and Section 2.4, respectively. In Section 2.5, we analyze random access load of random access procedures, and develop an optimization problem to find the number of preambles for *RAPID*. We then evaluate the performance of *RAPID* via simulation under the environment where mMTC UEs exist in Section 2.6. Finally, the paper concludes in Section 2.7.

2.2 Background

2.2.1 RRC State

In the LTE-A system, only two RRC states are defined, i.e., RRC_CONNECTED and RRC_IDLE states. If UE is in RRC_IDLE state, RRC connection must be established to allow the UE to transfer data to gNB. Therefore, uplink latency required in RRC_IDLE state is larger than in RRC_CONNECTED state. gNB uses an inactivity timer to manage the RRC state of each UE.

Fig. 2.2 shows RRC states and characteristics of each state in the 5G system. In the 5G system, RRC_INACTIVE state is introduced [14]. The main characteristics of the new state are as follows. First, while the RRC connection is released, both UE and gNB keep the context information of UE's RRC connection, such as UE capabilities and security context. When releasing the RRC connection, the gNB allocates an Access Stratum (AS) context ID to the UE in order to activate context information when resuming the RRC connection [15]. Second, connections between the gNB and core network for UE remain alive. These properties provide a way to reduce the latency for establishing the RRC connection when UE is in RRC_INACTIVE state. Since RRC_INACTIVE state is used as the primary sleeping state, the gNB should have a new inactivity timer to convert the RRC state from RRC_CONNECTED to

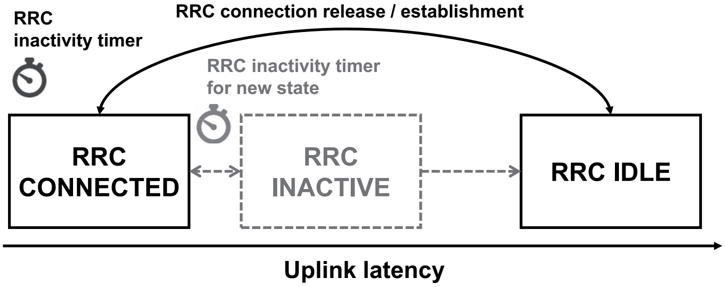

RRC inactivity timer
RRC connection release / establishment
RRC inactivity timer for new state
RRC CONNECTED
RRC INACTIVE
RRC IDLE
Uplink latency

Figure 2.2: RRC state machine and state transition in the 5G system: Solid rectangles and arrow represent RRC states and state transition of the LTE-A system, respectively. Dashed rectangle and arrows are newly added in the 5G system.

RRC_INACTIVE.

2.2.2 Random Access Procedure

In the LTE-A system, there are two types of random access procedures, i.e., contention-based and contention-free [16].

Contention-based random access: This procedure is initiated by a UE when a gNB does not allocate a preamble to the UE. It consists of four steps, and details are as follows.

1. **Preamble transmission:** The UE transmits a preamble randomly selected from a set of preambles for contention-based random access. The time when the UE transmits the preamble is determined by a list of allowed time slots allocated by the gNB.

2. **RAR:** The gNB, successfully receiving the preamble, transmits random access response (RAR) including timing advancement value for adjusting the uplink synchronization and uplink resource allocation information for a RRC connection resume request message. The gNB and the UE use random access-radio network temporary identifier (RA-RNTI) to transmit and receive RAR, respectively. Specifically, when transmitting (or receiving) RAR, the gNB scrambles (or the UE de-

scrambles) bits for error check of control channel with RA-RNTI which is determined by time-frequency resources of preamble transmitted by the UE. The gNB also includes the received preamble ID in the RAR so that the UE can identify whether the RAR is for itself or not. If the UE does not receive the RAR for certain amount of time, i.e., the size of *RAR window*, the UE tries random access again after performing a backoff procedure.

3. **RRC connection resume request:** The UE transmits the RRC connection resume request message using uplink resources allocated through the RAR. If two or more UEs simultaneously send the same preamble, UEs transmit RRC connection resume request messages using the same uplink resources. This is because the UEs have the same RA-RNTI, and thus receive the same RAR. In this case, it is difficult for the gNB to successfully decode RRC connection resume request message of each UE.

4. **RRC connection setup:** If the gNB successfully receives the third message, it sends a RRC connection setup message including cell-RNTI (C-RNTI) to identify the UE in the cell. As UE successfully receives the fourth message, the random access procedure is completed. However, if the UE does not receive the fourth message for certain amount of time, i.e., the value of *contention resolution timer* becomes zero, the UE tries random access again.

Contention-free random access: This procedure is performed when the gNB assigns a preamble to the UE transitioning to a RRC_INACTIVE state. The preamble is selected from a set of preambles for contention-free random access. The gNB that successfully received the preamble transmits the RAR as in the case of contention-based random access. Contention-free random access is completed through exchanging two messages because the preamble does not collide.

Table 2.1: Latency components of contention-based random access.

No.	Description	Time (ms)
1	Average delay due to RACH scheduling period	0.25
2	RACH preamble transmission	0.5
3	Preamble detection and RAR transmission	1.5
4	UE processing delay	1.25
5	RRC connection resume request transmission	0.5
6	gNB processing delay	1
7	RRC connection setup transmission	0.5
8	UE processing delay	3
	Total latency	8.5

2.2.3 Uplink Latency in RRC_INACTIVE State

In RRC_INACTIVE state, when an uplink packet is generated, UE must perform random access for resuming the RRC connection. Table 2.1 shows contention-based random access procedure and its latency components [17]. We set the transmit time interval (TTI) value to 0.5 ms.[3] In this case, the CP latency becomes 8.5 ms assuming that the processing time is reduced by one-fourth compared with the LTE-A [17]. With this assumption, the UP latency[4] becomes 3 ms. Therefore, the uplink latency is 11.5 ms. If the number of UEs performing contention-based random access increases, the uplink latency even increases further due to preamble collisions, thus increasing the latency requirement that can be satisfied. In the case of contention-free random access, because the procedure is completed in two steps, uplink latency becomes 6.5 ms, i.e., the CP latency is 3.5 ms (No. 1–4 in Table 2.1) and the UP latency is 3 ms. However, it is impossible for a large number of UEs to perform contention-free random access because the number of preambles is required to be equal to the number of UEs.

[3] We assume subcarrier spacing is 30 kHz which is doubled compared with the LTE-A [18]. Therefore, TTI value is 0.5 ms that is halved compared with the LTE-A.

[4] The UP latency value is the time from when UE transmits the buffer status report (BSR) message to when gNB successfully receives the data [19].

2.2.4 Related Work

In recent years, many studies have proposed random access procedures for MTC devices [1]. We review two representative random access schemes [2], [3]. We also introduce 2-step random access discussed in 3GPP temporary documents [20], [21]. Lastly, we discuss sparse code multiple access (SCMA) [22], one of non-orthogonal multiple access schemes.

Prioritized random access: The main idea of this technique is allocating different random access resources for each access class and preventing a large number of UEs from performing random access procedure at the same time. Specifically, it is possible to reduce the competition by allocating different subframe numbers according to each UE's class. Based on this idea, prioritized random access with dynamic access barring (PRADA) is proposed [2]. PRADA is proven to be superior to access class barring [23] in terms of random access success probability and average latency. However, since PRADA does not reduce contention among UEs in the same access class, it is difficult to satisfy the latency requirement of delay-sensitive UEs.

Random access for low cost-MTC: 3GPP RAN working group introduced a new random access channel (RACH) structure for low cost-MTC (LC-MTC) [3]. RACH for LC-MTC consists of multiple narrow band channels. Each channel has a pair of physical RACH (PRACH) and downlink control channel. The authors of [3] propose a new random access scheme using characteristics of the new RACH structure. In this scheme, using different PRACHs, multiple UEs can transmit the same preamble without collision. Also, gNB transmits separate RARs through multiple downlink control channels to reduce the collision of uplink resources. Although this scheme achieves low CP latency by reducing collision probability, it requires four message exchanges which are identical to contention-based random access. Therefore, we reduce the number of massage exchanges from four to two in *RAPID* to achieve lower CP latency.

2-step random access: In the 3GPP RAN working group, simplified contention-based random access with 2-step is defined [4]. 2-step random access has the advantage of reducing latency by simplifying the existing contention-based random access procedures. In the first step of 2-step random access, a UE transmits a preamble with payload, i.e., control message or data, using uplink resources randomly selected by UEs. In this case, however, it is possible for the first-step messages from different UEs to collide. Especially, because the probability of collision increases as the number of UEs increases, the 2-step random access proposed in [20], [21] cannot support delay-sensitive UEs. The difference between *RAPID* and 2-step random access is whether the collision problem can be resolved in two message exchanges. That is, we solve the collision problem in *RAPID* to achieve lower CP latency.

SCMA: Uplink grant-free transmission based on SCMA allows UEs to transmit data in an arrive-and-go manner. Different UEs may use the same radio resource, but use different codebooks and pilot sequences. In this case, a gNB is able to detect the data as long as different codebooks (or pilot sequences) are used [22]. For SCMA operation, uplink synchronization should be maintained with RRC connection established. However, for UEs in RRC_INACTIVE state, RRC connection is released and uplink synchronization is also lost. Therefore, SCMA is not appropriate for UEs in RRC_INACTIVE state.

2.3 *RAPID*: Proposed Random Access Procedure

2.3.1 Overview

We propose *RAPID* to overcome the limitations of the conventional random access procedures which are mentioned in Section 2.2.3. The key feature of *RAPID* is to complete the random access procedure by exchanging only two messages. *RAPID* enables this by using AS context ID for the following two procedures:

- Selection of the preamble and the set of allowed slot numbers to transmit preamble
- Scrambling of error check bits for control channel when sending RAR

Table 2.2: List of frequently-used parameters.

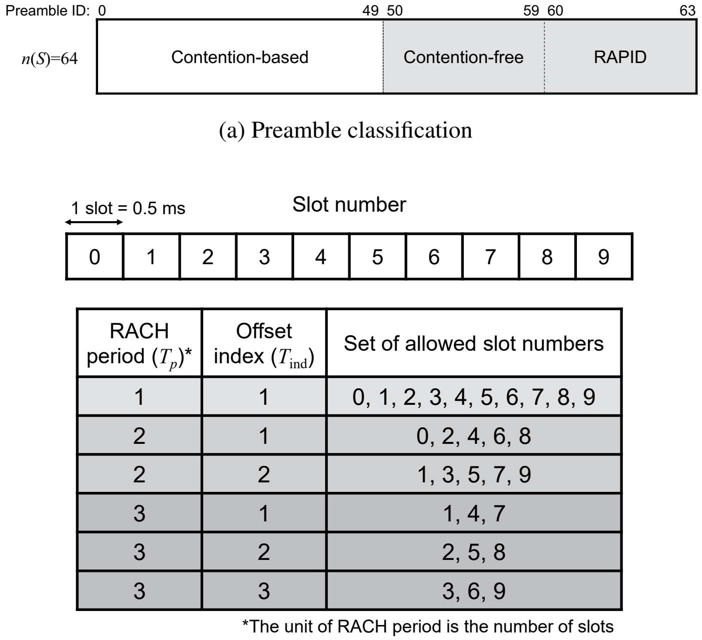
Symbol	Description
$n(S)$	Total number of preambles
$n(S_{\mathrm{cb}})$	Number of preambles for contention-based random access
$n(S_{\mathrm{cr}})$	Number of preambles for *RAPID*
id	AS context ID
pid	Preamble ID
i	UE index
r	Received uplink packet index
T_p	RACH period
T_{ind}	Offset index
t_{TTI}	TTI value
t_{up}	UP latency
t_I	Inactivity timer value for transition to RRC_INACTIVE state
t_{r-1}	Reception time of the r-th uplink packet
N_{cb}	The number of UEs performing contention-based random access
N_{cr}	The number of UEs performing *RAPID*
N_{ed}	The number of UEs whose traffic type is ED among N_{cr} UEs

In the proposed scheme, different UEs would try random access by selecting the same preamble in the same slot. However, contention can be resolved by sending different RARs scrambled by AS context ID of each UE. Therefore, *RAPID* is a contention resolution-based *R*andom *A*ccess *P*rocedure using AS context *ID*. The detailed description of *RAPID* is provided in the following subsections. Table 2.2 provides the list of parameters used in this paper along with their definition.

2.3.2 Criterion of Applying *RAPID*

The existing random access cannot satisfy the latency requirement of a UE according to the number of UEs served by a gNB. Therefore, the gNB should determine the random access method for the UE in consideration of latency requirement and the number of UEs served by the gNB. The gNB notifies this information to the UEs via the AS context ID. That is, the AS context ID includes information whether to apply *RAPID* or not. For example, if the most significant bit of AS context ID[5] is zero, *RAPID* should be used. Otherwise, the contention-based random access procedure should be

[5] We assume that AS context ID consists of enough bits to cover the number of UEs we handle.



(a) Preamble classification

RACH period (T_p)*	Offset index (T_{ind})	Set of allowed slot numbers
1	1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
2	1	0, 2, 4, 6, 8
2	2	1, 3, 5, 7, 9
3	1	1, 4, 7
3	2	2, 5, 8
3	3	3, 6, 9

*The unit of RACH period is the number of slots

(b) RACH period and offset index

Figure 2.3: Preamble set and RACH time resources.

used. We assume that contention-free random access procedure is not used for UEs in RRC_INACTIVE states.

2.3.3 Preamble Set and RACH Period Allocation

A UE selects a preamble and a set of allowed slot numbers to transmit preamble using AS context ID. For this purpose, gNB must inform UE of $n(S_{\text{cr}})$ representing the number of preambles for *RAPID* and T_p representing RACH period using a broadcast message. The RACH period is the interval between slot numbers that UE can transmit the preamble. Fig. 2.3(a) shows an example of preamble allocation to support *RAPID*. We consider a total $n(S)$ of preambles where $n(S) = 64$ and allocate 10 preambles for contention-free random access [2]. In addition, we allocate four preambles for *RAPID*. Fig. 2.3(b) represents sets of allowed slot numbers according to each T_p. We consider three types of T_p values. Offset index, denoted by T_{ind}, is defined to distinguish the sets of allowed slot numbers when T_p is fixed. For example, when $T_p = 3$, T_{ind} can range from one to three.

2.3.4 Preamble Transmission

Using AS context ID, UE selects a preamble ID, denoted by pid, from the given preamble set and determines a offset index to transmit the selected preamble.

Preamble selection: Using $n(S)$ and $n(S_{\text{cr}})$ broadcast by gNB, the preamble ID of the i-th UE is calculated by

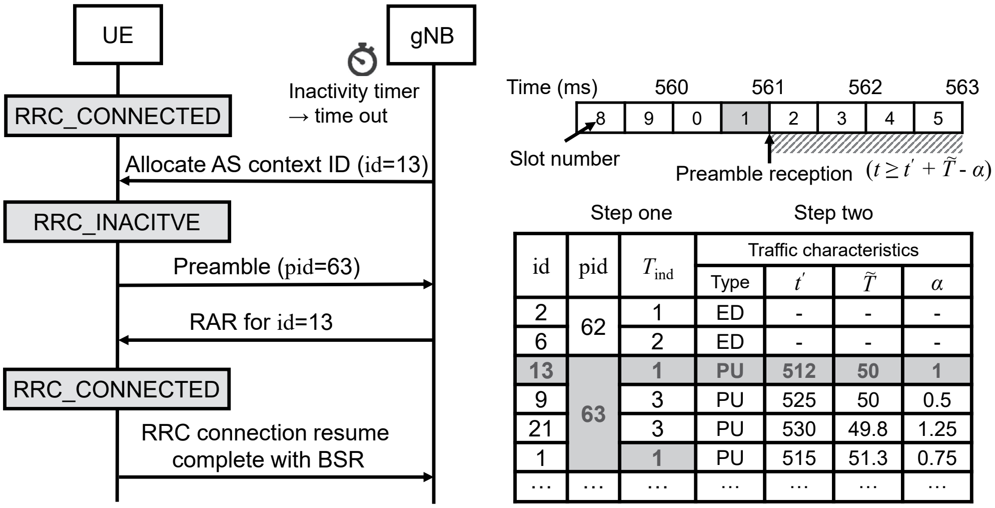
$$\text{pid}(i) = n(S) - 1 - \big(\text{id}(i) - 1\big) \bmod n(S_{\text{cr}}), \tag{2.1}$$

where $\text{id}(i)$ represents the i-th UE's AS context ID in decimal and mod is modulo operation. The zero value is not assigned to $\text{id}(i)$, and each $\text{pid}(i)$ is one of the values from $n(S) - n(S_{\text{cr}})$ to 63. Since UE selects the preamble using AS context ID, gNB does not need to use additional resources to inform the preamble ID that UE will use in RRC_INACTIVE state.

Offset index selection: Using $n(S_{\text{cr}})$ and T_p broadcast by gNB, the offset index of the i-th UE is given by

$$T_{\text{ind}}(i) = \left\lfloor \frac{\big(\text{id}(i) - 1\big) \bmod \big(n(S_{\text{cr}})T_p\big)}{n(S_{\text{cr}})} \right\rfloor + 1. \tag{2.2}$$

The numerator value of the floor function input can have an integer from 0 to $n(S_{\text{cr}})T_p - 1$. If we divide this value by $n(S_{\text{cr}})$ and apply the floor function, the output of the function has an integer value from 0 to $T_p - 1$. Therefore, we add one at the end to determine $T_{\text{ind}}(i)$ in (2.2). The UE transmits the selected preamble in the slot number closest to a slot in which traffic is generated among the set of allowed slot numbers corresponding to T_{ind}.



id	pid	T_{ind}	Traffic characteristics			
			Type	t'	$\tilde{T}$	α
2	62	1	ED	-	-	-
6		2	ED	-	-	-
13	63	1	PU	512	50	1
9		3	PU	525	50	0.5
21		3	PU	530	49.8	1.25
1		1	PU	515	51.3	0.75
⋯	⋯	⋯	⋯	⋯	⋯	⋯

Figure 2.4: Overall *RAPID* procedure of UE whose AS context ID is 13 and process of selecting UEs to receive RAR.

2.3.5 RAR Transmission

After receiving preambles for *RAPID*, gNB transmits one or more RARs[6] to candidate UEs that could send the preambles received by gNB. When transmitting RARs for each UE, gNB scrambles error check bits of control channel for each RAR with id of each candidate UE, respectively. Therefore, UE can successfully receive the RAR by descrambling error check bits of control channel with allocated id. It is important to send RARs to UEs having high probability of sending the preamble. Therefore, the gNB should select UEs to receive the RAR through two steps as shown in Fig. 2.4. In this example, we consider $n(S_{\text{cr}}) = 4$ and $T_p = 3$.

1. **Step one:** The gNB has a table that contains allocated AS context ID, preamble ID, offset index, and traffic characteristics. The gNB filters UEs that could send the preamble based on the received preamble and the slot number at which the preamble

[6] Since MTC UEs covered in this paper are fixed in position, timing advancement value obtained when UE first accesses to gNB is applied. To this end, the timing advancement value of UE should be stored in AS context.

is received. In Fig. 2.4, for example, the preamble with $\mathrm{pid} = 63$ is received at slot number one, and hence, UEs whose $\mathrm{id} = 1, 13$ could send the preamble (Shaded region in Step one columns in Fig. 2.4).

2. **Step two:** In this step, the gNB additionally exploits the traffic characteristics to predict UEs more precisely. For this purpose, we develop APA to estimate the traffic characteristics, i.e., traffic type, estimated traffic period, and margin value. We consider two traffic types, i.e., periodic update (PU) and event driven (ED) [24]. The PU traffic continuously generates uplink packets with a constant period, T. It should be noted that T is a separate parameter not related to RACH period, i.e., T_p. The ED traffic follows a Poisson process traffic model with an arrival rate, λ_{ed}. In case of the PU traffic, the gNB exploits $\tilde{T}$ representing estimated traffic period and α representing margin value obtained from APA. If preamble reception time, denoted by t, satisfies the equation below, the gNB transmits RAR for that UE (Shaded region in Step two columns in Fig. 2.4).

$$t \geq \ t' + \tilde{T} - \alpha, \tag{2.3}$$

where t' is the most recent time at which the gNB receives the preamble for a successful random access procedure. In case of ED traffic, the gNB always transmits RAR because of the difficulty of predicting traffic characteristics. In Fig. 2.4, the gNB transmits the RAR to UE whose id is 13 because its traffic characteristics satisfy (2.3). The detailed procedures that APA obtains the traffic characteristics are further described in Section 2.4.

UEs who perform *RAPID* do not carry out backoff procedure if random access is failed due to channel error or APA operation error. This is because the latency requirement of the delay-sensitive UE should be satisfied using *RAPID* procedure.

2.3.6 AS Context ID Allocation

gNB allocates AS context ID (id) to UE when its inactivity timer value becomes zero as shown in Fig. 2.4. Allocating id to a specific UE means that gNB determines preamble ID (pid) and offset index (T_{ind}) the UE will use. Therefore, the way to allocate id should reflect the following two elements for reducing random access load and satisfying latency requirement.

- Traffic type of UE
- Slot numbers at which gNB receives uplink packets from UE in RRC_CONNECTED state

If PU traffic UE and ED traffic UE[7] are allocated the same preamble, random access load increases because gNB always sends RAR when receiving preamble allocated to the ED traffic UE. Therefore, gNB should allocate different preambles depending on the traffic type of UE. For example, all PU traffic UEs are allocated a common preamble, and each ED traffic UE is randomly allocated one preamble among the remaining $n(S_{\mathrm{cr}}) - 1$ preambles.

In case of PU traffic UE, slot numbers at which gNB receives uplink packets are also considered. For instance, when $T_p = 3$, the maximum waiting time for PU traffic UE to transmit preamble is four slots, i.e., 2 ms when t_{TTI} is 0.5 ms. This value could affect the satisfaction of latency requirement for delay-sensitive UEs. Therefore, gNB selects T_{ind} based on slot numbers receiving uplink packets when UE is in RRC_CONNECTED state. The procedure of determining T_{ind} is detailed in Section 2.4.4. Otherwise, in case of ED traffic UE, because it is difficult to predict traffic characteristics, gNB randomly allocates candidate values of T_{ind} to UEs.

In short, when gNB receives uplink packets from a UE, gNB first determines pid and T_{ind} based on the UE's traffic type and the slot numbers at which the uplink packets

[7] PU (or ED) traffic UE is the UE whose traffic type is the PU (or ED).

are received, respectively. After that, gNB randomly selects id among the candidate IDs mapped to the given pid and T_{ind}.

2.3.7 Number of Preambles for *RAPID*

As the number of preambles for *RAPID*, i.e., $n(S_{\mathrm{cr}})$, increases, fewer ED traffic UEs are allocated to the same pid and T_{ind}. Therefore, random access load caused by both unnecessary RAR transmissions and unnecessary uplink resource allocation for RRC connection resume request messages decreases as $n(S_{\mathrm{cr}})$ increases. On the other hand, random access load for contention-based random access increases by increasing $n(S_{\mathrm{cr}})$. This is because the total number of two types of preambles, i.e., $n(S_{\mathrm{cb}}) + n(S_{\mathrm{cr}})$, is fixed at 54. For a given scenario, it is therefore of great importance to determine the optimal $n(S_{\mathrm{cr}})$ considering such trade-off relationship. For this purpose, in Section 2.5, we analyze random access load of two random access procedures, i.e., contention-based random access and *RAPID*, and develop an optimization problem to determine $n(S_{\mathrm{cr}})$.

2.4 Access Pattern Analyzer

2.4.1 Overview

APA predicts traffic characteristics of each UE to help gNB allocate AS context ID to UE and send RAR messages during *RAPID* operation. As mentioned in Section 2.3.2, the gNB determines whether *RAPID* is applied to each UE in consideration of latency requirement. APA initially estimates the traffic type of each delay-sensitive UE, so that we define an initial phase in APA. The dashed arrows in Fig. 2.5 represent the operation in the initial phase. The traffic type is estimated based on the uplink packet reception time (Section 2.4.2). T_{ind} should be determined to allocate AS context ID. For this purpose, we define offset index decision procedure (Section 2.4.4). After the initial phase, if the traffic type is PU, APA estimates the traffic period and margin value

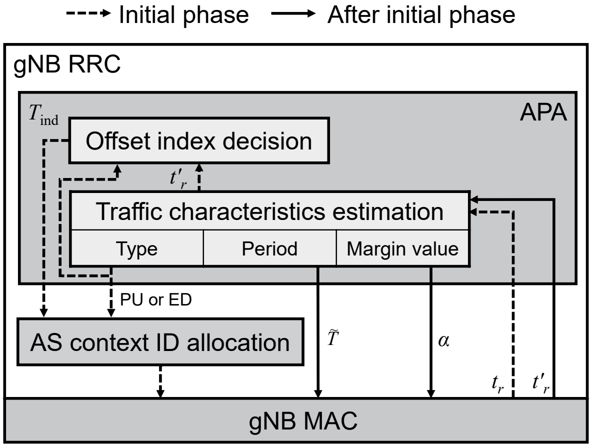


Figure 2.5: Overall APA operation.

of the corresponding UE (Section 2.4.2 and 2.4.3).

2.4.2 APA Operation

During the initial phase, UEs are in RRC_CONNECTED state, and APA obtains reception time values of uplink packets. Algorithm 1 shows the detailed procedure estimating the traffic type. At first, RRC inactivity timer for transition to RRC_INACTIVE state, denoted by t_I, is set to T_{init}. APA stores the time value of $t_{r-1}(i)$ when gNB receives the r-th uplink packet from the i-th UE (line 5). To estimate the time to receive the preamble after the initial phase, we keep track of the r-th preamble reception time using $t'_{r-1}(i)$ which is calculated using $t_{r-1}(i)$ (line 6). Specifically, UP latency, denoted by t_{up}, is subtracted from the $t_{r-1}(i)$, and then TTI value, denoted by t_{TTI}, is added to indicate the preamble reception time. When more than two uplink packets from the i-th UE are received, i.e., $r \geq 2$, the estimated traffic period of the i-th UE, denoted by $\tilde{T}_{r-1}(i)$, is calculated by linear regression (LR) using a normal equation (line 9) [25]. For r uplink packets of the i-th UE, the normal equation is defined

Algorithm 1 APA operation for the i-th UE.

```
Initialize:
$t_I \leftarrow T_{\text{init}}, r \leftarrow 0$
During initial phase:
while $t_I \neq 0$ do
    if New uplink packet is received then
        $r \leftarrow r + 1$
        $t_{r-1}(i) \leftarrow$ current time
        $t'_{r-1}(i) = t_{r-1}(i) - t_{\text{up}} + t_{\text{TTI}}$
        if $r \geq 2$ then
            $\mathbf{t}(i) = \big(t'_0(i), \cdot\cdot\cdot, t'_{r-1}(i)\big)^T$
            $\tilde{t}'_0(i), \tilde{T}_{r-1}(i) \leftarrow LR\big(r, \mathbf{t}(i)\big)$
        end if
        if $r = R_{\text{th}}$ then
            $\sigma^2 = \text{Var}\big(\tilde{T}_1(i), \cdot\cdot\cdot, \tilde{T}_{R_{\text{th}}-1}(i)\big)$
            if $\sigma^2 \leq \delta_{\text{th}}$ then
                UE has PU traffic type
                $k^* \leftarrow \text{MAS}\big(\mathbf{t}(i),\ R_{\text{th}}\big)$                ▷ Algorithm 2
            else
                UE has ED traffic type
            end if
            $t_I \leftarrow T_I$
            break
        end if
    else
        $t_I \leftarrow t_I - t_{\text{TTI}}$
    end if
end while
if $r < R_{\text{th}}$ then
    UE has ED traffic type, $t_I \leftarrow T_I$
end if
After initial phase for PU traffic UE:
if Preamble for the $i$-th UE is received then
    $t_{\text{temp}}(i) \leftarrow$ current time
    if Random access succeed then
        $r \leftarrow r + 1$
        $t'_{r-1}(i) \leftarrow t_{\text{temp}}(i)$
        $\mathbf{t}(i) = \big(t'_0(i), \cdot\cdot\cdot, t'_{r-1}(i)\big)^T$
        $\tilde{t}'_0(i), \tilde{T}_{r-1}(i) \leftarrow LR\big(r, \mathbf{t}(i)\big)$
    end if
end if
```

as

$$\Theta = LR\big(r, \mathbf{t}(i)\big) = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{t}(i),$$
$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & r-1 \end{pmatrix}, \ \mathbf{t}(i) = \begin{pmatrix} t'_0(i) \\ t'_1(i) \\ \vdots \\ t'_{r-1}(i) \end{pmatrix}, \tag{2.4}$$

where $(\cdot)^T$ and $(\cdot)^{-1}$ represent transpose and inverse of a matrix, respectively. The result of LR, denoted by Θ, is a 2×1 vector, which is $\Big(\tilde{t}'_0(i), \tilde{T}_{r-1}(i)\Big)^T$, where $\tilde{t}'_0(i)$ is the estimated initial uplink packet reception time of the i-th UE.

When the number of received packets becomes R_{th}, the traffic type is estimated based on the variance of the stored estimated traffic period values (lines 11–18). If the UE's estimated traffic type is PU, Algorithm 2 is called to determine T_{ind} of the UE (line 15). After the traffic type of UE is estimated, t_I is set to T_I, which is much smaller than T_{init}. This is for making the UE go to RRC_INACTIVE state. If the traffic type of UE can not be estimated until the timeout of initial inactivity timer, whose value is set to T_{init}, the UE is considered having ED traffic type (line 27).[8]

After the initial phase, for the PU traffic UE, APA updates the estimated traffic period and the estimated initial uplink packet reception time (line 35). For this purpose, the time when UE receives preamble in the successful *RAPID* procedure is stored to t'_{r-1} (line 33).

[8] Many MTC applications have periodic traffic with T values much longer than $T_{\text{init}}/R_{\text{th}}$. T could be in the order of hours, days, or even months. By the way, these applications have latency requirements that can be supported by contention-based random access [26].

2.4.3 Margin Value

Accurate period estimation is difficult due to channel errors. Therefore, a margin value is required and it is given by

$$\alpha_{R-1} = \frac{1}{R}\sum_{r=1}^{R}\left|t'_{r-1} - \left(\tilde{t}'_0 + (r-1)\tilde{T}_{R-1}\right)\right|, \quad R \geq 2, \tag{2.5}$$

where $\tilde{t}'_0$ is the output of (2.4) and R is the number of t'_{r-1} samples. If we have a total of R samples of t'_{r-1}, generalization of (2.3) is given by

$$t \geq\ t'_{R-1} + \tilde{T}_{R-1} - \alpha_{R-1}, \quad R \geq 2. \tag{2.6}$$

2.4.4 Offset Index Decision

As mentioned in Section 2.3.6, gNB determines a UE's pid using the traffic type obtained from APA. For ED traffic UEs, gNB randomly selects one of candidate values of T_{ind}. For PU traffic UEs, gNB should determine T_{ind} considering slot numbers receiving uplink packets. For this purpose, gNB needs to know a set of allowed slot numbers in which UE will transmit preamble. We refer to this set of allowed slot numbers as the most accessed set, denoted by s_{k^*}.

Algorithm 2 shows how to determine k^* using $\mathbf{t}$ in (2.4). Firstly, we define s_k (where $1 \leq k \leq T_p$) to investigate frequency of each set (line 2). In case of $T_p = 1$, because there is only one offset index, k^* is one (line 4). If $T_p = 2$ or 3, for all the values of t'_{r-1} in $\mathbf{t}$, we can obtain value k, which is the index of the set containing the received slot number (lines 6–25). Especially, when $T_p = 3$, if the received slot number is zero, i.e., *temp* is zero, $k = 1$ is appropriate to make waiting time minimum (line 16). Lastly, k^* is the value of k with the largest value s_k (line 27). We name this algorithm most accessed set (MAS), which is a procedure to find $T_{\text{ind}} = k^*$. In addition, a proper R_{th} value should be chosen to obtain k^* according to the value of T_p, e.g., when

Algorithm 2 MAS: procedure to determine k^*.

```
Input:
$\mathbf{t}$, $R_{\mathrm{th}}$
Initialize:
$s_k \leftarrow 0$ for $\forall k$ $(1 \leq k \leq T_p)$
Call by Algorithm 1
switch $T_p$ do
    case $1$ break
    end case
    case $2$
        for $r \leftarrow 1 ... R_{\mathrm{th}}$ do
            $k \leftarrow (t'_{r-1}/t_{\mathrm{TTI}} - 1) \bmod T_p + 1$
            $s_k \leftarrow s_k + 1$
        end for
    end case
    case $3$
        for $r \leftarrow 1 ... R_{\mathrm{th}}$ do
            $temp \leftarrow t'_{r-1}/t_{\mathrm{TTI}} - 1$
            if $temp = 0$ then
                $k \leftarrow 1$
            else
                $k \leftarrow temp \bmod T_p$
                if $k = 0$ then
                    $k \leftarrow 3$
                end if
            end if
            $s_k \leftarrow s_k + 1$
        end for
    end case
end switch
$k^* = \underset{k}{\operatorname{argmax}}\, s_k$
```

$T_p = 2$, R_{th} should be an odd number.

2.5 Random Access Load Analysis

We analyze random access load of two types of random access, i.e., contention-based random access (4-step RA) and *RAPID* as presented in Section 2.3.7. We define random access load as the number of scheduled signals including both the untransmitted

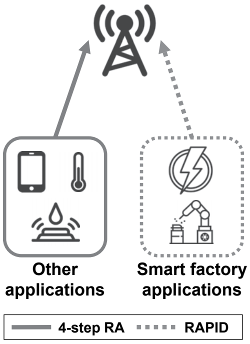


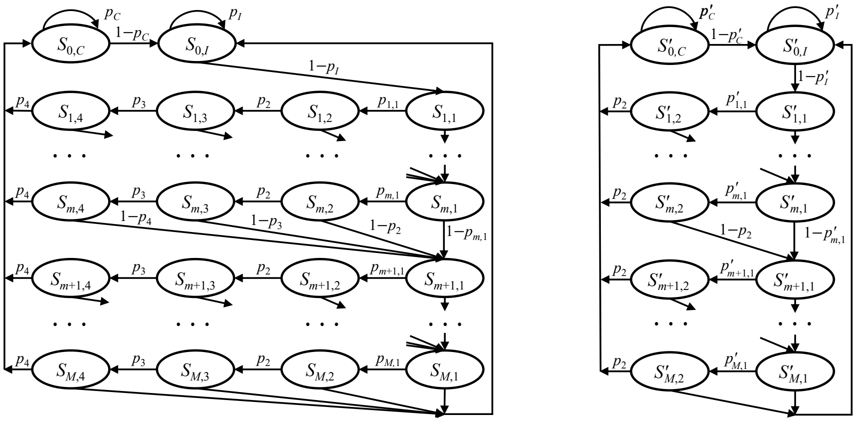
Figure 2.6: System model.

signals after unnecessary resource allocation and the transmitted signals. We also develop an optimization problem to determine the number of preambles for *RAPID*, i.e., $n(S_{\mathrm{cr}})$, using the analysis result.

2.5.1 System Model

We consider a scenario whereby many MTC UEs in RRC_INACTIVE state transmit uplink packets. Each UE has one application with one of two types of latency requirements, i.e., delay-sensitive or delay-tolerant [27], [28]. As shown in Fig. 2.6, UEs running smart factory applications are delay-sensitive devices [29]. The UEs with smart factory applications should use *RAPID* to meet the latency requirement. The latency requirements of other applications are delay-tolerant such that UEs with those applications perform 4-step RA. The UEs performing 4-step RA should succeed random access procedures within the maximum number of random access opportunities. Under this scenario, we analyze random access load of each random access procedure to find the optimal number of preambles for *RAPID*, i.e., $n(S_{\mathrm{cr}})$. We employ Markov chain to model each random access procedure. In the proposed model, the following assumptions are made.

- We assume packet arrival process follows Poisson [30].
- In 4-step RA, physical random access channel is available in every slot [30].



(a) Model for 4-step RA

(b) Model for *RAPID*

Figure 2.7: Markov chain models for random access procedures.

- In 4-step RA, collision of control messages, i.e., RRC connection request, occurs with constant and independent probability [31].
- Radio resources used for random access procedures are enough to serve the UEs trying each random access.[9]

2.5.2 Markov Chain Model for 4-Step RA

Fig. 2.7(a) shows the Markov chain model for 4-step RA to analyze random access load, where there are three types of states, i.e., RRC_CONNECTED, RRC_INACTIVE, and random access procedure states. The state $S_{0,C}$ represents RRC_CONNECTED state. A UE in this state transmits uplink packets and a gNB operates inactivity timer of which value is denoted by t_I for that UE. The state $S_{0,I}$ represents RRC_INACTIVE state, where the UE waits for the generation of uplink packet. Rest of the states, $S_{m,n}$'s (where $1 \leq m \leq M$ and $1 \leq n \leq 4$), represent 4-step RA procedure. Index m is the number of random access attempts, and M is the maximum number of random access

[9] In LTE system, downlink control channel resources for random access are limited [32]. For MTC application, however, enough resources for downlink control channel should be guaranteed to improve the random access performance [3, 33].

opportunities. Index n represents each step of 4-step RA presented in Section 2.2.2.

For a UE in the state $S_{0,C}$, if an uplink packet is generated before the timer value t_I becomes zero, the UE stays in that state. On the other hand, if nothing happens until t_I becomes zero, a state transition takes place from $S_{0,C}$ to $S_{0,I}$. For the UE in the state $S_{0,I}$, if an uplink packet is generated, the state is transferred from $S_{0,I}$ to $S_{1,1}$. Since the packet arrival process follows a Poisson process, inter-packet arrival time follows an exponential distribution, denoted by $X \sim \text{Exp}(\lambda_{\text{cb}})$, where λ_{cb} is the packet arrival rate of the UE performing 4-step RA. We can obtain transition probability values for $S_{0,C}$ and $S_{0,I}$ as $p_C = 1 - e^{-\lambda_{\text{cb}}(t_{\text{up}}+t_I)}$ and $p_I = e^{-\lambda_{\text{cb}} t_{\text{TTI}}}$, respectively. In the m-th random access trial, if the UE fails to transmit (or receive) message for 4-step RA, the state is transferred from $S_{m,n}$ to $S_{m+1,1}$. When $m = M$, if the UE fails, the next state is $S_{0,I}$.

We define transition probability in each step of 4-step RA as follows.

- $p_{m,1}$: This value means preamble detection probability, i.e., $1 - e^{-m}$, where m indicates the m-th preamble transmission [30], [32]. Even if multiple UEs transmit the same preamble and collision occurs, the gNB can detect the preamble if at least one preamble transmission succeeds without channel error.

- p_2 and p_4: These values represent successful downlink control messages reception ratio. For these messages, because gNB uses very low modulation and coding scheme with high transmission power, we consider $p_2 = p_4 \approx 1$ [30].[10]

- p_3: This value means successful transmission probability of RRC connection resume request message. $p_3 = (1-\rho_{\text{col}})(1-\rho_{\text{ch}})$, where ρ_{ch} is channel error, can be reduced to $1 - \rho_{\text{col}}$ because of $1 - \rho_{\text{ch}} \approx 1$ like p_2 and p_4.

[10] We validate this value is appropriate by comparing simulation using channel model [34] in Section 2.5.6.

The collision probability, denoted by ρ_{col}, is determined as

$$\begin{aligned}\rho_{\text{col}} = &\sum_{j=1}^{N_{\text{cb}}-1} \binom{N_{\text{cb}}-1}{j} \tau^j (1-\tau)^{N_{\text{cb}}-1-j} \\ &\times \left(1 - \left(1 - \frac{1}{n(S_{\text{cb}})}\right)^j\right),\end{aligned} \tag{2.7}$$

where N_{cb} is the number of UEs performing 4-step RA and τ means the probability that a UE successfully transmits a preamble in one slot. Eq. (2.7) represents the probability that when one UE selects a preamble, at least one of the other UEs successfully transmit the same preamble in the same slot.

We denote $\pi_{m,n}$ as stationary probability of state $S_{m,n}$. First, $\pi_{0,I}$ is obtained as

$$\begin{aligned}\pi_{0,I} = &\frac{(1-p_C)\pi_{0,C} + \pi_{M,1}(1-p_{M,1})}{1-p_I} \\ &+ \frac{\sum_{n=2}^{4} \pi_{M,n}(1-p_n)}{1-p_I}.\end{aligned} \tag{2.8}$$

We can simply calculate $\pi_{1,1} = \pi_{0,I}(1-p_I)$. Next, we define $f(m)$ and $g(m)$ to represent stationary probabilities of $\pi_{m,1}$ and $\pi_{m,n}$ (where $2 \le n \le 4$), respectively.

$$\begin{aligned}f(m) &= \pi_{m,1} \\ &= \begin{cases} \pi_{0,I}(1-p_I), & m = 1, \\ f(m-1)\Big(1 - p_{m-1,1} + p_{m-1,1} & \\ \times \big(1 - p_2 + \sum_{l=3}^{4} \prod_{j=2}^{l-1} p_j (1-p_l)\big)\Big), & m \ge 2, \end{cases} \\ g(m) &= \sum_{n=2}^{4} \pi_{m,n} \\ &= p_{m,1} f(m) + p_{m,1} \sum_{n=2}^{3} \prod_{j=2}^{n} p_j f(m) \\ &= p_{m,1} f(m) \left(1 + \sum_{n=2}^{3} \prod_{j=2}^{n} p_j\right).\end{aligned} \tag{2.9}$$

Because the sum of stationary probabilities of all states is one, we can obtain the following equation using (2.9).

$$\begin{aligned}
&\pi_{0,C} + \pi_{0,I} + \sum_{m=1}^{M}\sum_{n=1}^{4}\pi_{m,n} = 1, \\
&\pi_{0,C} + \pi_{0,I} + \sum_{m=1}^{M}\left(f(m) + g(m)\right) = 1, \\
&\pi_{0,C} + \pi_{0,I} + \sum_{m=1}^{M} f(m)\left(1 + p_{m,1}\left(1 + \sum_{n=2}^{3}\prod_{j=2}^{n} p_j\right)\right) = 1.
\end{aligned} \tag{2.10}$$

Eq. (2.10) is the function of ρ_{col} so that $\pi_{0,C}$ can be derived in terms of ρ_{col}. The other stationary probabilities of all states are also derived in terms of ρ_{col}. Meanwhile, τ also represents the proportion of time in successful preamble transmission of the states $S_{m,1}$ (where $1 \leq m \leq M$). Therefore, τ is obtained as

$$\tau = \frac{1}{T_{\text{tot}}}\sum_{m=1}^{M}\pi_{m,1}t_{TTI}p_{m,1}, \tag{2.11}$$

where T_{tot} is the average holding time for all states, i.e.,

$$T_{\text{tot}} = \pi_{0,C}T_{0,C} + \pi_{0,I}T_{0,I} + \sum_{m=1}^{M}\sum_{n=1}^{4}\pi_{m,n}T_{m,n}, \tag{2.12}$$

where $T_{m,n}$ is the holding time for each state $S_{m,n}$. $T_{0,C}$ is calculated by [35, Eq. (9)]

$$\begin{aligned}
T_{0,C} &= E\big[\min\left(X, t_{\text{up}} + t_I\right)\big] \\
&= \int_0^{\infty} P\big(\min\left(X, t_{\text{up}} + t_I\right) > x\big)dx \\
&= \int_0^{t_{\text{up}}+t_I} P(X > x)dx = \int_0^{t_{\text{up}}+t_I} e^{-\lambda_{\text{cb}}x}dx \\
&= \frac{1}{\lambda_{\text{cb}}}\left(1 - e^{-\lambda_{\text{cb}}(t_{\text{up}}+t_I)}\right),
\end{aligned} \tag{2.13}$$

where X is a random variable representing inter-packet arrival time. Because preamble transmission is possible in every slot, $T_{0,I}$ is t_{TTI}, which is the slot length.

The remaining state holding time values can be calculated using Table 2.1. Each state holding time consists of two components, i.e., when each step (index n) of the random access procedure succeeds (p_n) or fails ($1 - p_n$). Therefore, the states $S_{m,n}$ with same n have the same holding time value. The holding time of the state $S_{m,1}$ is

$$T_{m,1} = t_{\text{TTI}} p_{m,1} + (t_{\text{TTI}} + W_{\text{RAR}} + BW_{\text{avg}})\,(1 - p_{m,1}), \tag{2.14}$$

where W_{RAR} is the *RAR window* size, and BW_{avg} is the average value of backoff window size. The holding time of the state $S_{m,2}$ is written as

$$T_{m,2} = 1.5p_2 + (W_{\text{RAR}} + BW_{\text{avg}})\,(1 - p_2). \tag{2.15}$$

The value 1.5 is the time value of preamble detection and RAR transmission (No. 3 in Table 2.1). The holding time of the state $S_{m,3}$ is obtained as

$$T_{m,3} = 1.75p_3 + (1.75 + W_{\text{res}} + BW_{\text{avg}})\,(1 - p_3), \tag{2.16}$$

where W_{res} is the *contention resolution timer* value. The value 1.75 is the sum of UE processing delay and RRC message transmission time (No. 4–5 in Table 2.1). Lastly, the holding time of the state $S_{m,4}$ is

$$T_{m,4} = 4.5p_4 + (W_{\text{res}} + BW_{\text{avg}})\,(1 - p_4). \tag{2.17}$$

The value 4.5 contains gNB processing delay, RRC message transmission time, and UE processing delay (No. 6–8 in Table 2.1).

The value τ can be obtained by solving system of equations with unknown variables ρ_{col} and τ, i.e., (2.7) and (2.11). Specifically, the right-hand side of (2.11) is changed to the formula in terms of τ. Then, we select the intersection point with $y = \tau$, i.e., left-hand side of (2.11), to obtain τ. Finally, we can calculate ρ_{col} and all stationary probabilities using τ.

2.5.3 Average Random Access Load for 4-Step RA

In 4-step RA, random access load for each state transition is one except for four transitions, i.e., transitions from $S_{0,C}$ and $S_{0,I}$. For each state $S_{m,n}$, random access load is $1 \times \pi_{m,n} p_n + 1 \times \pi_{m,n}(1 - p_n) = \pi_{m,n}$. (When n is one, p_n is replaced by $p_{m,1}$.) Therefore, the average random access load for 4-step RA is obtained as

$$E[L_{\text{cb}}] = \frac{1}{T_{\text{tot}}} \sum_{m=1}^{M} \sum_{n=1}^{4} \pi_{m,n}. \tag{2.18}$$

2.5.4 Markov Chain Model for *RAPID*

Fig. 2.7(b) shows the Markov chain model for *RAPID*. We only investigate random access load for ED traffic UEs. This is because one of $n(S_{\text{cr}})$ is always allocated to PU traffic UEs, while $n(S_{\text{cr}}) - 1$ preambles are used for ED traffic UEs as mentioned in Section 2.3.7. Thus, changing $n(S_{\text{cr}})$ only affects the random access load for ED traffic UEs. The model for *RAPID* also has three types of states like the model for 4-step RA. Because *RAPID* procedure is completed within two steps, the difference from the model for 4-step RA is that the maximum value of n is two.

The inter-packet arrival time of ED traffic also follows an exponential distribution, i.e., $X' \sim \text{Exp}(\lambda_{\text{ed}})$, where λ_{ed} is packet arrival rate of the ED traffic UE. Therefore, the value of p'_C is $1 - e^{-\lambda_{\text{ed}}(t_{\text{up}}+t_I)}$ and p'_I is $e^{-\lambda_{\text{ed}} T'_p t_{\text{TTI}}}$. It should be noted that, in the exponent of p'_I, we use T'_p instead of T_p. T'_p is introduced to represent the average interval of state transitions from $S_{0,I}$ and given as[11]

$$T'_p = \begin{cases} T_p, & T_p = 1, 2, \\ 10/3, & T_p = 3. \end{cases} \tag{2.19}$$

The value of transition probability, p_2, is the same as the value for 4-step RA, but $p'_{m,1}$ is different from the value of $p_{m,1}$. In *RAPID*, we consider the number of UEs allocated to the same offset index and preamble. When k UEs transmit the same preamble in the same slot, if at least one preamble transmission is successful, the gNB can transmit one or more RARs. Therefore, in *RAPID*, preamble detection probability for a UE transmitting the m-th preamble is obtained as

$$\begin{aligned} p'_{m,1} &= 1 - \prod_{j=1}^{M} \left(e^{-j}\right)^{\overline{N}'_{\mathrm{UE}}(j)}, \\ \overline{N}'_{\mathrm{UE}}(j) &= \begin{cases} \overline{N}_{\mathrm{UE}}(j), & j = m, \\ \overline{N}_{\mathrm{UE}}(j) - 1, & j \neq m, \end{cases} \end{aligned} \tag{2.20}$$

where $\overline{N}_{\mathrm{UE}}(j)$ is the average number of UEs transmitting the j-th preamble at the same time. The value $\overline{N}_{\mathrm{UE}}(j)$ is

$$\overline{N}_{\mathrm{UE}}(j) = \sum_{k=1}^{\lceil \overline{N}_{\mathrm{RAR}} \rceil} p_k(j) k, \tag{2.21}$$

where $p_k(j)$ is the probability that k UEs transmit the same preamble simultaneously, i.e., transmit the j-th preamble transmission at the same time, when one UE transmits

[11] For the state $S_{0,I}$, a state transition occurs at the allowed slots. In *RAPID*, however, the number of slots between two consecutive allowed slots is not always equal to T_p. For instance, if the packet is generated in slot number seven when $T_p = 3$ and $T_{\mathrm{ind}} = 1$, preamble is transmitted in the next slot number one.

a preamble in a particular slot.

$$
\begin{aligned}
p_k(j) = &\binom{\lceil \overline{N}_{\mathrm{RAR}} \rceil - 1}{k-1} \left(\frac{T'_p}{\mathrm{slot}_{\mathrm{avg}}} (1 - p_{j-1,1}) \right)^{k-1} \\
&\times \left(1 - \frac{T'_p}{\mathrm{slot}_{\mathrm{avg}}} (1 - p_{j-1,1}) \right)^{\lceil \overline{N}_{\mathrm{RAR}} \rceil - k},
\end{aligned}
\tag{2.22}
$$

where $\overline{N}_{\mathrm{RAR}}$ is the average number of ED traffic UEs uniformly allocated to the combinations of offset indices and the number of preambles that ED traffic UEs use, which is

$$
\overline{N}_{\mathrm{RAR}} = \max\left(\frac{N_{\mathrm{ed}}}{T_p\big(n(S_{\mathrm{cr}}) - 1\big)}, 1 \right). \tag{2.23}
$$

$\mathrm{slot}_{\mathrm{avg}}$ is the average number of slots for which one packet is generated for a UE, so that $\mathrm{slot}_{\mathrm{avg}} = 1/(\lambda_{\mathrm{ed}} t_{\mathrm{TTI}})$.

$\pi'_{m,n}$ is the stationary probability of state $S'_{m,n}$. As we did in Section 2.5.2, $\pi'_{0,I}$ is obtained as

$$
\pi'_{0,I} = \frac{(1 - p'_c)\pi'_{0,C} + \pi'_{M,1}(1 - p'_{M,1}) + \pi'_{M,2}(1 - p_2)}{1 - p'_I}, \tag{2.24}
$$

and $\pi'_{1,1}$ is $\pi'_{0,I}(1 - p'_I)$. We also define $f'(m)$ to represent the stationary probability of state $S'_{m,1}$.

$$
f'(m) = \begin{cases} \pi'_{0,I}(1 - p'_I), & m = 1, \\ f'(m-1)\big(1 - p'_{m-1,1} & \\ +p'_{m-1,1}(1 - p_2)\big), & m \geq 2. \end{cases} \tag{2.25}
$$

We also derive the following equation using the property that the sum of all stationary probabilities is one.

$$
\begin{aligned}
&\pi'_{0,C} + \pi'_{0,I} + \sum_{m=1}^{M}\sum_{n=1}^{2} \pi'_{m,n} = 1, \\
&\pi'_{0,C} + \pi'_{0,I} + \sum_{m=1}^{M} \left(1 + p'_{m,1}\right) f'(m) = 1.
\end{aligned}
\tag{2.26}
$$

Because there is no unknown variable in (2.26), we can obtain $\pi'_{0,C}$. Also, the other stationary probabilities are calculated.

2.5.5 Average Random Access Load for *RAPID*

The average random access load for *RAPID* can be obtained as

$$
E[L_{\text{ed}}] = \frac{1}{T'_{\text{tot}}} \sum_{m=1}^{M} \pi'_{m,1} + \big(2\text{eff}(\overline{N}_{\text{RAR}}) - 1\big)\pi'_{m,2}, \tag{2.27}
$$

where T'_{tot} is the average holding time for all states, which is

$$
T'_{\text{tot}} = \pi'_{0,C}T'_{0,C} + \pi'_{0,I}T'_{0,I} + \sum_{m=1}^{M}\sum_{n=1}^{2} \pi'_{m,n}T'_{m,n}. \tag{2.28}
$$

$T'_{0,C}$ is calculated in the same way as $T_{0,C}$. In the result of (2.13), only λ_{cb} is changed to λ_{ed}. The holding time for the state $S'_{0,I}$ is $T'_{0,I} = t_{\text{TTI}}T'_{p}$. The values of holding time for states $S'_{m,1}$ and $S'_{m,2}$ are $T'_{m,1} = t_{\text{TTI}}p_{m,1} + (t_{\text{TTI}} + W_{\text{RAR}})(1 - p_{m,1})$ and $T'_{m,2} = 2.75p_2 + W_{\text{RAR}}(1 - p_2)$, respectively.

For the random access load for state transition from $S'_{m,1}$ is one, and for state transition from $S'_{m,2}$, we can exploit the result of (2.23), which is $\overline{N}_{\text{RAR}}$. It also represents the average number of RAR messages a gNB transmits after receiving a preamble. However, it is not accurate to use the value of $\overline{N}_{\text{RAR}}$ to obtain the average random access load. This is because when k UEs transmit the same preamble in the same slot, random access load due to multiple RAR transmissions is reduced from one UE's per-

Table 2.3: System parameters.

Parameter	Value	Parameter	Value
Carrier frequency	2 GHz	Avg. backoff window size	5 ms
System bandwidth	20 MHz	*Contention resolution timer*	24 ms
t_{TTI}	0.5 ms	T_{init}	5 s
$n(S_{\mathrm{cr}}) + n(S_{\mathrm{cb}})$	54	T_I	5 ms
RACH period (T_p)	3 slots	R_{th}	10
RAR window size	2.5 ms	δ_{th}	0.1

spective. Therefore, we calculate effective random access load, denoted by $\mathrm{eff}(\overline{N}_{\mathrm{RAR}})$, reflecting the probability that k UEs transmit the same preamble in the same slot, and it is calculated by

$$\mathrm{eff}(\overline{N}_{\mathrm{RAR}}) = \sum_{k=1}^{\lceil \overline{N}_{\mathrm{RAR}} \rceil} p'_k \frac{\overline{N}_{\mathrm{RAR}}}{k}, \tag{2.29}$$

where p'_k is the probability that one or more UEs transmit the same preamble simultaneously when one UE transmits a preamble in a particular slot, i.e.,

$$\begin{aligned} p'_k = &\binom{\lceil \overline{N}_{\mathrm{RAR}} \rceil - 1}{k-1} \left(\frac{T'_p}{\mathrm{slot}_{\mathrm{avg}}} \sum_{m=0}^{M-1} (1 - p_{m,1}) \right)^{k-1} \\ &\times \left(1 - \frac{T'_p}{\mathrm{slot}_{\mathrm{avg}}} \sum_{m=0}^{M-1} (1 - p_{m,1}) \right)^{\lceil \overline{N}_{\mathrm{RAR}} \rceil - k}. \end{aligned} \tag{2.30}$$

When the average random access load of *RAPID* is calculated, $\mathrm{eff}(\overline{N}_{\mathrm{RAR}})$ is doubled and then one is subtracted as in (2.27). This means that random access load includes not only all RAR transmissions but also unnecessary uplink resource allocation for RRC connection resume request messages except uplink resource allocation for a UE who transmits a preamble, i.e., this allocation is used for the transmission of an RRC connection resume request message.

2.5.6 Validation of Analysis

For the validation of analysis, we introduce scenarios that consist of two types of appli-

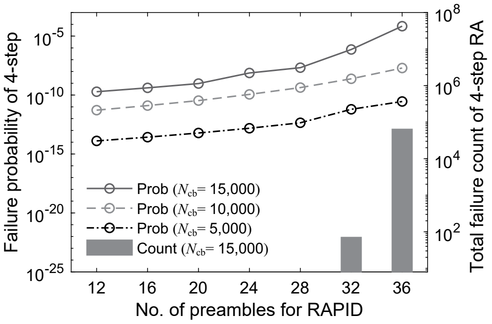
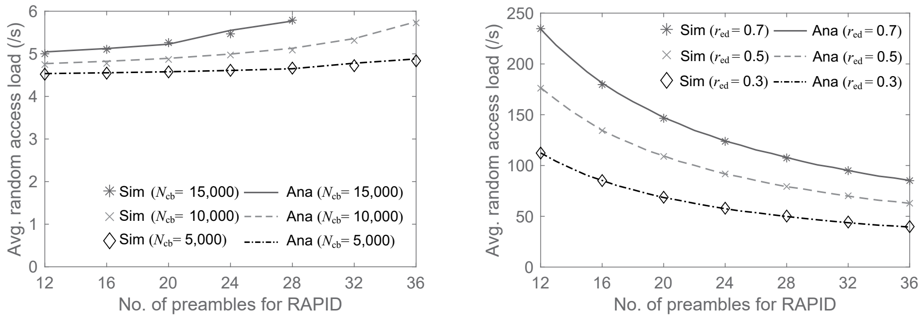


Figure 2.8: Failure probability of 4-step RA for UE in the left axis, and the total failure count of 4-step RA in the right axis (‘Prob’ and ‘Count’ represent failure probability and total failure count, respectively).



(a) Avg. random access load for 4-step RA

(b) Avg. random access load for *RAPID*

Figure 2.9: Average random access load for 4-step RA and *RAPID* (‘Sim’ and ‘Ana’ represent simulation and analysis results, respectively).

cations defined in Section 2.5.1. Table 2.3 summarizes system parameters for analysis and simulation [36], [37]. Specifically, the parameters related to random access are revised considering the reduced t_{TTI}, i.e., 0.5 ms. For the simulation, we create path loss and shadowing following 3D-urban macro (3D-UMa) model defined in [38]. We also consider fast fading channel model generated using ITU-R IMT UMa model in [34].

Since we have to find an appropriate number of preambles for *RAPID*, i.e., $n(S_{\mathrm{cr}})$, we first determine the upper bound of $n(S_{\mathrm{cr}})$. This is the same way that we find the minimum number of preambles for 4-step RA to guarantee the reliability of UEs performing 4-step RA. For this purpose, we observe two values: (i) the failure probability of 4-step RA for a UE in the analysis and (ii) the total failure count of 4-step RA in the simulation. The failure probability of 4-step RA is given by

$$P_{\text{fail,cb}} = \pi_{M,1}(1 - p_{m,1}) + \sum_{n=2}^{4} \pi_{M,n}(1 - p_n), \tag{2.31}$$

where M is the maximum number of random access attempts, which is 10 [26]. The total failure count of 4-step RA is the number of failures in all of 10 4-step RA attempts. Fig. 2.8 shows the above two values under the various $n(S_{\text{cr}})$ and the number of UEs performing 4-step RA, i.e., N_{cb}. Traffic arrival rate, i.e., λ_{cb}, is fixed at 1 packet/s. When N_{cb} is 5,000 or 10,000, the maximum value of total failure count is under ten, so that we do not present it in Fig. 2.8. When N_{cb} and $n(S_{\text{cr}})$ are 15,000 and 32, respectively, the value of total failure count starts increasing. That is why, in certain scenarios, if the number of preambles in 4-step RA, i.e., $n(S_{\text{cb}})$, is smaller than a certain value, the number of UEs to try random access continues to increase due to collision, and random access repeatedly fails. Therefore, we determine a threshold value of the failure probability as 10^{-7} to prevent the total failure count from starting to increase. We define the reliability of 4-step RA, which is $1 - P_{\text{fail,cb}}$. Therefore, we only consider $n(S_{\text{cr}})$ up to the value satisfying the reliability of 4-step RA above 99.99999%.

Fig. 2.9(a) shows the average random access load of 4-step RA for a UE according to $n(S_{\text{cr}})$. When the number of UEs performing 4-step RA is fixed, we can observe the average random access load slightly increases as $n(S_{\text{cr}})$ increases. Fig. 2.9(b) shows the average random access load of *RAPID* for an ED traffic UE according to $n(S_{\text{cr}})$. The number of UEs performing *RAPID*, denoted by N_{cr}, is fixed at 1,000. We consider various ratios of the number of ED traffic UEs to the total number of UEs, which is $r_{\text{ed}} = 0.3, 0.5, 0.7$. Traffic arrival rate of ED traffic UE, i.e., λ_{ed}, is 6.8 packet/s [29]. The random access load is reduced as r_{ed} decreases and $n(S_{\text{cr}})$ increases, because fewer ED traffic UEs are allocated to the same combination of preamble and allowed slot numbers.

2.5.7 Optimization Problem

We develop the optimization problem for determining $n(S_{\mathrm{cr}})$ to minimize the sum of random access loads of 4-step RA and *RAPID* for ED traffic UEs. The optimal number of preambles for *RAPID* is

$$\begin{gathered}\underset{n(S_{\mathrm{cr}})}{\operatorname{argmin}}\ E[L_{\mathrm{cb}}]N_{\mathrm{cb}} + E[L_{\mathrm{ed}}]N_{\mathrm{ed}} \\ \text{subject to } n(S_{\mathrm{cr}}) = j\ (0 \leq j \leq 54) \\ P_{\mathrm{fail,cb}} < 10^{-7},\end{gathered} \tag{2.32}$$

where 10 preambles are assigned for contention-free random access. Therefore, $n(S_{\mathrm{cr}})$ can have the value which is from zero to 54. Also, as observed in Section 2.5.6, $P_{\mathrm{fail,cb}}$ is less than 10^{-7} to satisfy the reliability of 4-step RA above 99.99999%.

Since we cannot represent the objective function by a closed form, the above optimization problem should be solved using an exhaustive search. Accordingly, we calculate the values of the objective function depending on $n(S_{\mathrm{cr}})$, i.e., zero to 54. That is, the value of the objective function should be calculated up to 55 times. Objective function consists of two terms. The second term which is (2.27) can be easily obtained from solving (2.26), which is a linear equation. The first term can be a bottleneck in terms of computational complexity. As mentioned in Section 2.5.2, the system of equations composed of (2.7) and (2.11) should be solved. At this time, it can be a high-order equation depending on N_{cb}. We can obtain computational complexity as $O(n^3)$ by using subdivision algorithm to compute isolating intervals for the real roots of a n-th order polynomial [39], [40].

The value of the first term in the objective function increases as the number of preambles for *RAPID* increases because of increase of the collision probability. On

Table 2.4: Simulation scenario.

Applications	Smart factory		Other types
Traffic type	PU	ED	ED
T or λ	50 ms	6.8/s	0.5/s
The number of UEs	1,000		23,000
r_{ed}	0.3, 0.5, 0.7		-

the other hand, the value of the second term in the objective function decreases as the number of preambles for *RAPID* increases because of decrease of the unnecessary RAR transmission. Also, the value of the second term increases as r_{ed} increases because of increase of unnecessary RAR transmission. The optimal number of preambles for *RAPID* is the value of $n(S_{\mathrm{cr}})$ that yields the minimum value of the objective function.

2.6 Performance Evaluation

In this section, we evaluate the performance of *RAPID* via MATLAB simulation. 4-step RA, PRADA [2], and LC-MTC random access (LC-MTC RA) [3] presented in Section 2.2.4 are used as comparison schemes.

2.6.1 Simulation Setup

Table 2.4 shows the parameters of the scenario in order to evaluate *RAPID*. The total number of UEs connected with a gNB is 24,000 [36], and the number of UEs with smart factory application is 1,000. For smart factory application, period of PU traffic UEs is 50 ms and traffic arrival rate of ED traffic UEs is 6.8 packets/s [29]. Traffic arrival rate of UEs with the other applications is 0.5 packet/s. The packet sizes of PU traffic and ED traffic are 125 bytes and 10 bytes, respectively [36]. We use the system parameters defined in Table 2.3.

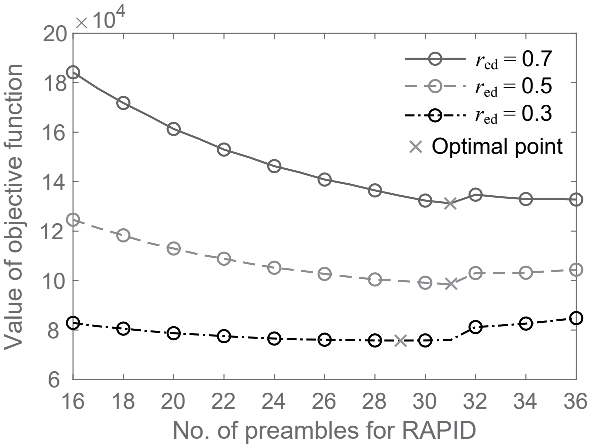


Figure 2.10: The optimal number of preambles for *RAPID*.

2.6.2 Number of Preambles for *RAPID*

We now provide the optimal value of $n(S_{\text{cr}})$ for each scenario in Table 2.4 by solving the optimization problem in (2.32) using MATLAB. Fig. 2.10 shows the value of objective function according to $n(S_{\text{cr}})$ under various ratio of ED traffic UE. First, as mentioned in Section VI-F, we determine the minimum number of preambles for 4-step RA ensuring reliability above 99.99999%. In case of our simulation scenario, the minimum number of preambles for 4-step RA is 18. Thus, we observe the value of $n(S_{\text{cr}})$ until 36. The appropriate $n(S_{\text{cr}})$ minimizes the random access load while ensuring the reliability of 4-step RA. In the rest of simulation for *RAPID*, therefore, we set the values of $n(S_{\text{cr}})$ to 29, 31, and 31 when r_{ed} is 0.3, 0.5, and 0.7, respectively.

2.6.3 Performance of *RAPID*

Latency and Reliability: In general, reliability is defined as the probability that a certain amount of data from a user device will be successfully transmitted to another peer within a predetermined time [41]. The predetermined time is uplink latency requirement value, which is denoted by L_{rq}. Accordingly, the reliability can be expressed

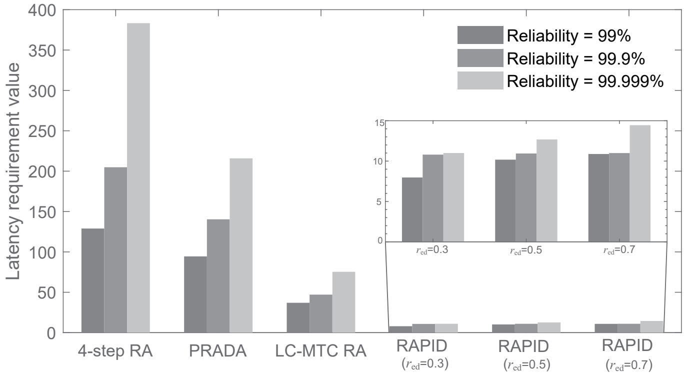


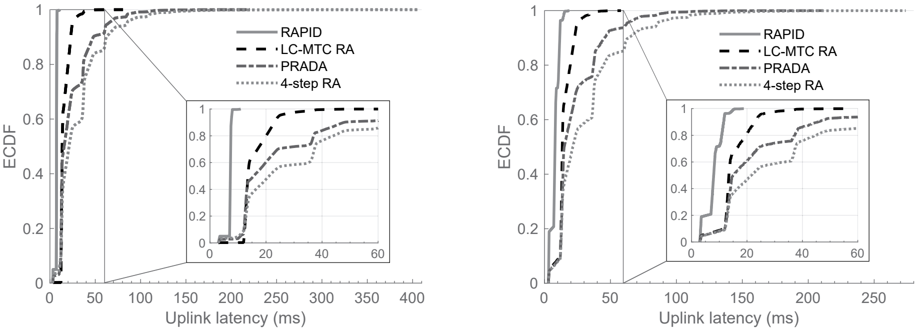
Figure 2.11: Satisfiable latency requirements with various reliability values of PU traffic UEs.

as

$$\text{Reliability} = P(D \leq L_{\text{rq}}), \tag{2.33}$$

where D is the measured uplink latency. As the reliability is a function of the uplink latency requirement, we can obtain the corresponding latency requirement value for a given reliability. We define the value as satisfiable latency requirement with the given reliability.

Fig. 2.11 shows satisfiable latency requirements with different reliability values of PU traffic UEs for each scheme. In 4-step RA, we can find the latency requirement with 99% reliability is larger than 100 ms. In this scenario, it is hard to satisfy the latency requirement of delay-sensitive UEs by 4-step RA. In the case of PRADA and LC-MTC RA, the satisfiable latency requirement increases sharply as the reliability value to be satisfied increases. In *RAPID*, when r_{ed} is 0.7, the satisfiable latency requirement with 99.999% reliability is 14.44 ms, which is 80.8% smaller than that of LC-MTC RA. The reason for the latency decrease in *RAPID* is the higher random access success probability in virtue of multiple RAR transmissions after a successful preamble transmission from at least one UE among the UEs transmitting the same preamble. In *RAPID*, as r_{ed} increases, the satisfiable latency requirement with each



(a) PU traffic UEs

(b) ED traffic UEs

Figure 2.12: ECDF of uplink latency for smart factory applications when $N_{\mathrm{cb}} = 23{,}000$, $N_{\mathrm{cr}} = 1{,}000$, and $r_{\mathrm{ed}} = 0.3$.

reliability slightly increases. This is because as the value of r_{ed} increases, fewer PU traffic UEs transmit the same preamble on average, and thus the probability of a successful preamble transmission from at least one UE decreases.

Figs. 2.12(a) and 2.12(b) shows empirical cumulative distribution function (ECDF) of the uplink latency for smart factory applications. In Fig. 2.12(b), uplink latency values are more distributed between 3 ms and 4 ms than in Fig. 2.12(a) for all schemes. In case of *RAPID*, ED traffic UEs stay longer in the initial phase than PU traffic UEs. On the other hand, in the other schemes, it is possible to have the uplink packet delivered before the inactivity timer times out for ED traffic UEs. In Fig. 2.12(b), for ED traffic UEs, the satisfiable latency requirement with 99.999% reliability is 18.85 ms. This value is larger than the value of the PU traffic UE because fewer UEs transmit the same preamble on average, and thus the probability of a successful preamble transmission from at least one UE decreases.

Random access load: The bars in Fig. 2.13 show the sum of random access load for all applications. For each scheme, there are two types of random access load, i.e., necessary and unnecessary. The necessary random access load is the number of signals used for successful random access. In contrast, unnecessary random access load includes three types: (i) the signals when random access fails, (ii) the unnecessary RAR

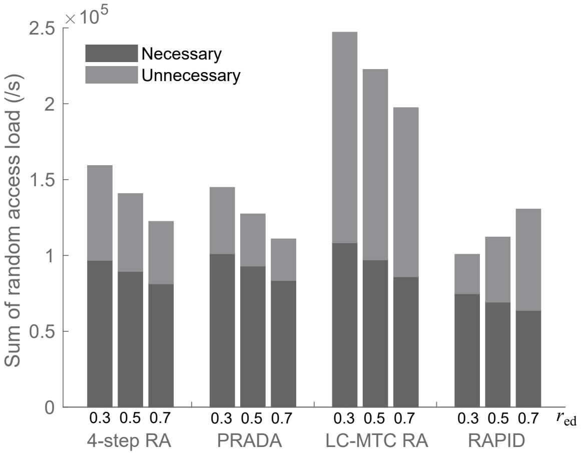


Figure 2.13: The right-side graph shows the sum of random access load for all applications.

transmissions to the UEs who do not transmit preamble but have the same preamble ID and offset index for *RAPID*, and (iii) the unnecessary uplink resource allocation for RRC connection resume request messages for the UEs mentioned in (ii). In cases of 4-step RA, PRADA, and LC-MTC RA, because the ratio of PU traffic UEs who generate more packets than ED traffic UEs is reduced, the sum of random access load decreases as r_{ed} increases. However, when *RAPID* is used for smart factory applications, the sum of random access load increases as r_{ed} increases because of increasing of unnecessary RAR transmission.

The necessary random access load of PRADA is higher than that of 4-step RA, and the unnecessary random access load of PRADA is lower than that of 4-step RA. This is because PRADA allocates different random access resources to each access class, and hence, random access success ratio is higher than that of 4-step RA. LC-MTC RA further increases random access success ratio by decreasing the collision probability, thus yielding higher necessary random access load than PRADA and 4-step RA. On the other hand, unnecessary random access load of LC-MTC RA is the highest among all schemes. This is because LC-MTC RA reduces collisions at the cost of making a gNB transmit separate RARs through multiple downlink control channels.

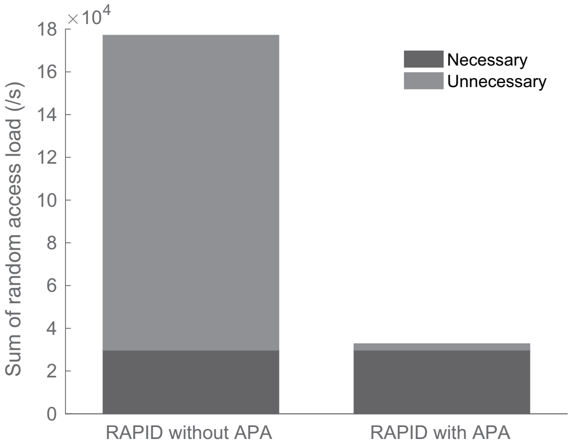


Figure 2.14: Sum of random access load for smart factory applications with and without APA when $N_{\mathrm{cr}} = 1,000$ and $r_{\mathrm{ed}} = 0.3$.

In case of *RAPID* where $r_{\mathrm{ed}} = 0.3$, necessary random access load is reduced by 26.1% compared with PRADA. This is because *RAPID* requires only two message exchange procedures. Moreover, unnecessary random access load is decreased by 40.6% compared with PRADA. This is because UEs who perform *RAPID* do not suffer from random access failures due to collision. Also, with the help of APA, unnecessary RAR transmission can be minimized. Compared with PRADA, *RAPID* reduces the sum of random access load by 30.5% and 11.9% when r_{ed} is 0.3 and 0.5, respectively. As r_{ed} increases, the random access load reducing gain of *RAPID* decreases due to the increased unnecessary random access load, because more UEs share the same preamble and offset index. When r_{ed} is 0.7, unnecessary random access load is increased further and the sum random access load of *RAPID* becomes comparable with that of 4-step RA. *RAPID*, however, reduces the latency requirements that cannot be satisfied with comparison schemes.

2.6.4 Performance of APA

For the validation of APA performance, we consider a scenario consisting of smart factory application UEs where the number of UEs is 1,000 and $r_{\text{ed}} = 0.3$. We allocate $n(S_{\text{cr}})$ to 54, which is the maximum value defined in (2.32). In the initial phase, APA perfectly distinguishes the traffic type of UEs, i.e., PU or ED traffic. In case of PU traffic UEs, the average of estimated period is 49.99 ms, and the variance is 0.015. It can be seen that APA accurately predicts the traffic type of UEs and the period of the PU traffic UEs. Fig. 11 shows the sum of random access load for smart factory applications and without APA. When APA is not applied, all UEs regardless of traffic types are uniformly allocated to the combinations of preambles and allowed slot numbers. When APA is applied, however, the UEs can be classified according to traffic types. Especially, traffic characteristics of PU traffic UEs can be grasped by APA and unnecessary random access load can be reduced by 98% compared with *RAPID* without APA.

2.7 Summary

We propose *RAPID*, which is a new random access procedure for delay-sensitive UEs in RRC_INACTIVE state introduced in 5G. *RAPID* completes the random access procedure by exchanging two messages using AS context ID of UE in RRC_INACTIVE state. We also develop APA for reducing random access load caused by unnecessary RAR transmission. We then develop an optimization problem to obtain the number of preambles for *RAPID* based on random access load analysis. We also validate the analysis via comparison with simulation results. Through simulations and mathematical analysis considering mMTC devices, we demonstrate that *RAPID* can support delay-sensitive UEs by satisfying more strict latency requirement compared with the state-of-the-art schemes.

Chapter 3

EsTA: Self-Uplink Synchronization in 2-Step Random Access

3.1 Introduction

The 5G system will be the foundation technology for business innovation in various vertical industries such as smart factories, cars, and smart cities. In June 2018, 3GPP finalized Release 15 specifications, which are the first 5G NR standard including non-standalone and standalone modes. In April 2019, mobile operators in South Korea and the United States launched commercial 5G services. In June 2020, the 3GPP completed Release 16 including not only technical improvements of Release 15 specifications but also the introduction of new features. In Release 17, 3GPP is working on the new features for a wide variety of industry verticals and non-terrestrial access systems to build out to be substantially more versatile than 4G LTE.

One of the new features in Release 16 is the use of two-step random access channel (2-step RACH). The 3GPP RAN working groups specify 2-step random access covering both physical layer and higher layer. 2-step random access potentially offers benefits in the following two scenarios [4]. First, for burst transmission of small packets, simple random access is attractive for reducing the significant overhead of

RRC connection setup and resume procedures [14]. Second, for the NR Unlicensed spectrum (NR-U), reducing the steps of random access helps decrease the latency for connecting a UE to a gNB since they perform a listen-before-talk procedure for connection step.

In this article, we present the details of 2-step random access in 5G NR, namely 2-step contention-based random access (CBRA), which suffers from *preamble collision* when many UEs try channel access. This is because many UEs compete for the limited number of preambles on the same time-frequency resource called PRACH resource. We briefly introduce the existing 2-step random access schemes proposed to solve the *preamble collision* problem [42], [43]. Each scheme has pros and cons, so we present the challenges of the random access schemes. As a means to solve the *preamble collision* problem, we focus on TA command estimation of a UE for self-uplink synchronization with the gNB.

Therefore, we propose estimation of timing advance (EsTA), a framework that helps a UE determine its own TA value. In the proposed framework, an edge RAN controller trains a simple DNN model on large data consisting of features (Reference Signal Received Power values) and labels (TA commands). Each UE estimates the TA command using machine learning, and determines a TA value.

In summary, this chapter includes:

- a comprehensive overview of 2-step CBRA defined in the 5G NR and recently studied 2-step random access schemes

- a discussion of the challenges of 2-step random access schemes

- a framework for TA command estimation using a DNN model and TA value determination for each UE to resolve the *preamble collision* problem.

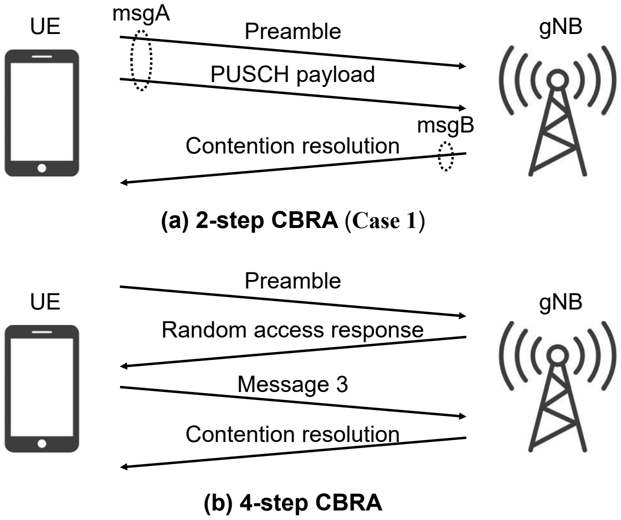


Figure 3.1: CBRA in 5G NR.

3.2 Background

In the 5G NR Release 15, CBRA has the same steps as 4-step CBRA in 4G LTE. To reduce latency, the 5G NR Release 16 and recent literature simplify the existing CBRA procedures from four steps to two steps. We overview the 2-step CBRA procedures defined in the 5G NR and two random access procedures in [42], [43].

3.2.1 Overview of 2-Step CBRA

The RAN working group addresses a simplified 2-step random access as a work item. This item proposes a common design for unlicensed spectrum as well as licensed spectrum. This article focuses on the 2-step CBRA in the licensed spectrum. As shown in Fig. 3.1, 2-step CBRA has the advantage of reducing latency by simplifying the existing 4-step CBRA that carries preamble and payload separately. Message A (msgA) contains a preamble on PRACH and a payload on physical uplink shared channel (PUSCH). The payload corresponds to the message 3 in 4-step CBRA, which is the first scheduled uplink transmission [44]. After transmitting a msgA with a preamble, a UE waits for a message B (msgB) from the gNB on physical downlink shared channel (PDSCH) for the configured window. The gNB takes different actions depending on its reception status of msgA.

- **Case 1** The gNB detects the preamble from the UE and successfully decodes the payload. It notifies the UE of contention resolution by sending a successful RAR with a TA command which is an integer value greater than or equal to zero.
- **Case 2** The gNB detects a single preamble but fails to decode the payload. Using the preamble reception time, it sends back a fallback RAR to the UE with the TA command and an uplink grant for the payload retransmission.
- **Case 3** The gNB detects multiple identical preambles from UEs. There is no fallback RAR because the gNB is unable specify the preamble reception time of each UE. Therefore the gNB transmits a backoff indication to UEs that will attempt random access again.
- **Case 4** The gNB fails to detect the preamble. There is no RAR to the UE.

The UE upon receiving the RAR successfully completes the 2-step CBRA. Upon receiving the fallback RAR, the UE falls back to 4-step CBRA with message 3 transmission (i.e., payload retransmission). In **Case 3**, the UE performs a backoff procedure and retransmits msgA after waiting for the length of the configured window called *RAR window*. In **Case 4**, the UE retransmits msgA after waiting for the length of *RAR window*. If 2-step CBRA could not succeed even when the UE trasmits the msgA 'M' times, the UE would fall back to 4-step CBRA that starts from the preamble transmission.

3.2.2 Channel Structure for msgA

Different from the first step in the 4-step CBRA, msgA in 2-step CBRA contains payload as well as preamble. Therefore, the channel structure for msgA should be

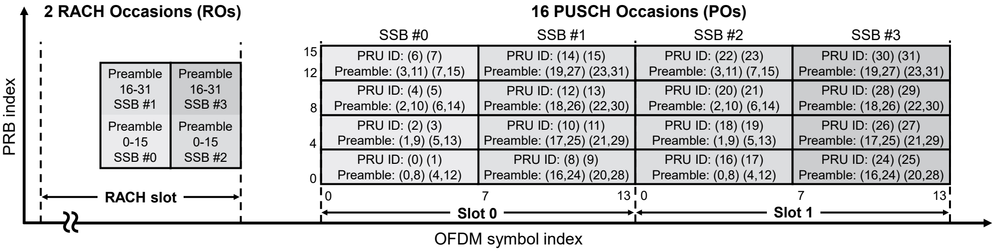

2 RACH Occasions (ROs)
16 PUSCH Occasions (POs)
PRB index
Preamble 16-31 SSB #1
Preamble 16-31 SSB #3
Preamble 0-15 SSB #0
Preamble 0-15 SSB #2
RACH slot
SSB #0
SSB #1
SSB #2
SSB #3
15
12
8
4
0
PRU ID: (6) (7) Preamble: (3,11) (7,15)
PRU ID: (14) (15) Preamble: (19,27) (23,31)
PRU ID: (22) (23) Preamble: (3,11) (7,15)
PRU ID: (30) (31) Preamble: (19,27) (23,31)
PRU ID: (4) (5) Preamble: (2,10) (6,14)
PRU ID: (12) (13) Preamble: (18,26) (22,30)
PRU ID: (20) (21) Preamble: (2,10) (6,14)
PRU ID: (28) (29) Preamble: (18,26) (22,30)
PRU ID: (2) (3) Preamble: (1,9) (5,13)
PRU ID: (10) (11) Preamble: (17,25) (21,29)
PRU ID: (18) (19) Preamble: (1,9) (5,13)
PRU ID: (26) (27) Preamble: (17,25) (21,29)
PRU ID: (0) (1) Preamble: (0,8) (4,12)
PRU ID: (8) (9) Preamble: (16,24) (20,28)
PRU ID: (16) (17) Preamble: (0,8) (4,12)
PRU ID: (24) (25) Preamble: (16,24) (20,28)
0
7
13
0
7
13
Slot 0
Slot 1
OFDM symbol index

Figure 3.2: Example of the channel structure for msgA: mapping between preamble IDs and PRU IDs, and resource size of each PO.

newly defined, which includes mapping between a preamble on a PRACH resource and a time-frequency resource for the PUSCH payload, time-frequency resource size of PUSCH, and so on [45].

2-step CBRA uses the preamble format specified in Release 15 [46]. The msgA preamble set is different from the 4-step CBRA preamble set, but the both preamble sets can be transmitted through the same RACH occasion (RO) or separate ROs. Fig. 3.2 represents an example of the channel structure for msgA. Two ROs exist in a RACH slot, and each RO uses 32 preambles for 2-step CBRA. The beam association rule between synchronization signal block (SSB) and RO of 4-step CBRA [46] is used for 2-step CBRA as follows. The gNB associates an SSB index with its own RO and/or preamble. Different SSB indexes indicate different downlink beams of the gNB or different reception beams of the gNB in the case of uplink transmission. Fig. 3.2 shows that a set of 16 preambles out of 32 preambles is used to represent a specific beam. For example, if the UE transmits preamble 5 in the second RO, the gNB uses the beam corresponding to SSB #2 to receive the payload of the msgA.

The payload transmission consists of PUSCH occasions (POs) which span multiple orthogonal frequency division multiplexing (OFDM) symbols and physical resource blocks (PRBs). Each PO consists of multiple PUSCH resource units (PRUs), and each of which contains the following fields:

- PRU ID

- Multiple OFDM symbols and PRBs for uplink transmission
- Association with preamble(s) of a PRACH resource
- Modulation and coding scheme
- Uplink power control related parameters
- Demodulation reference signal (DMRS) port and DMRS sequence.

Fig. 3.2 shows that one PO includes two PRUs, and occupies four RBs in the frequency domain and seven symbols in the time domain. One PRU is associated with one or more preambles of a PRACH resource, i.e., preamble ID(s) of a specific RO. There are two types of resource mapping between preambles of PRACH resource and a PRU: many-to-one and one-to-one mapping. Fig. 3.2 shows an example of many-to-one mapping and the mapping ratio, where the number ratio of preamble IDs to a PRU is two. Preambles 0 and 8 transmitted in an RO for SSB #0 are mapped to PRU ID 0.

3.2.3 TA Handling for the Payload

TA value (N_{TA}) is a negative time offset for the UE to control uplink transmission timing. An adequate TA value makes uplink transmission better aligned with the symbol timing at the gNB. In 2-step CBRA, the gNB determines a TA command (T_{A}) based on the reception timing of the msgA preamble. Then, the UE calculates a TA value using the TA command and its numerology (μ), i.e., $N_{\mathrm{TA}}(T_A, \mu)$ [47].

On the other hand, the TA value for the payload of the msgA is set to zero [45]. The gNB receives uplink transmissions with different delays depending on the distance between the UE and the gNB. Basically, OFDM systems mitigate multipath interference with the help of the cyclic prefix (CP) between two adjacent OFDM symbols. The case (a) of Fig. 3.3 shows that the sum of round trip delay (RTD) and delay spread (τ) for each UE is less than or equal to the CP. So the gNB can decode OFDM symbols from different UEs transmitted at the same time. Otherwise, inter-symbol interference

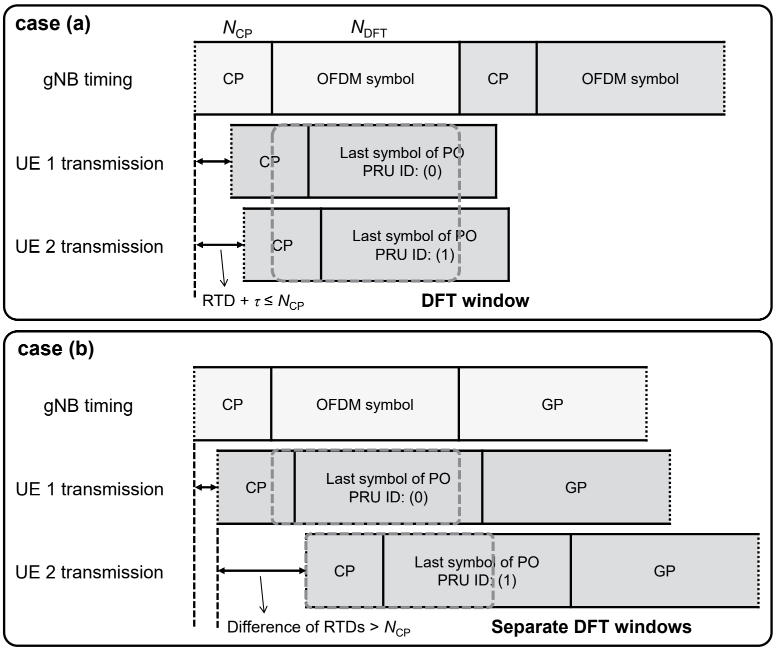


Figure 3.3: Two cases of the payload reception timing at the gNB.

(ISI) occurs due to delayed OFDM symbols. Thus, the RAN working group added optional guard period (GP) to the end of each PO illustrated in the case (b) of Fig. 3.3. The GP ranges from 0 to 3 symbols [45].

Fig. 3.4 represents the number of delayed samples according to the distance from the gNB and numerologies. When the numerologies are zero, one, and two, the subcarrier spacing values are 15, 30, and 60 kHz, respectively. We easily calculate the number of delayed samples by rounding up the sum of RTD and τ divided by the sampling time, i.e., $\lceil(\text{RTD}+\tau)/\text{sampling time}\rceil$ where τ is 93.325 ns, which is the root mean square delay spread of 3D-UMa model used in [42]. The sampling time is the inverse of the sampling rate, i.e., 1/($N_{\mathrm{DFT}} \times$subcarrier spacing) where N_{DFT} is discrete fourier transform (DFT) size, i.e., 4096. The CP length (N_{CP}) is 288 in the unit of the number of samples, indicated by the red line in Fig. 3.4. As the subcarrier spacing values increase, we can find that environments where the maximum distance values

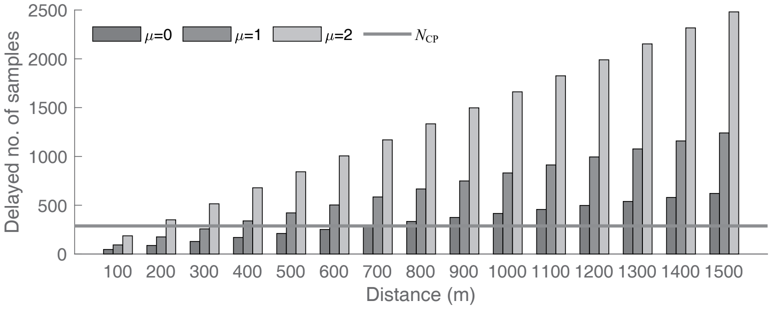


Figure 3.4: Number of delayed samples vs. distance from the gNB and numerologies.

from the gNB (or coverages) are over 700, 400, and 200 m, need the GP.

3.2.4 2-Step Random Access in Recent Literature

In 2-step CBRA, when multiple UEs try channel access, the probability that different UEs use the identical preamble in the same PRACH resource increases, resulting in the *preamble collision* problem. We divide this problem into two cases. First, if more than two UEs at a similar distance from the gNB transmit the same preamble using the same RO, the payload transmission would fail even if their preamble transmissions are successful (**Case 2**). We refer to this case as the *undetected collision* problem. Second, if multiple UEs at different distances from the gNB transmit the same preamble using the same RO, the gNB detects multiple identical preambles. In this case, the gNB fails to determine the TA command for each UE (**Case 3**). This case is called the *detected collision* problem.

We have proposed two types of 2-step random access to resolve the *preamble collision* problem. First, in the contention resolution-based random access [42], the gNB allocates a unique context ID to each UE. The UE selects and transmits a preamble from a specific preamble set, using a specific PRACH resource mapped to the context ID (one-to-many mapping between a subset of PRACH resources and context IDs). Upon receiving the preamble, the gNB transmits multiple RARs to candidate UEs that could have sent the preamble. This scheme completes random access by exchanging only two messages. In this way, we can solve the *undetected collision problem*. In addition, when a *detected collision* occurs, the gNB uses the TA command for each UE that was stored during the UE's first access. However, the shortcoming is that unnecessary RAR transmissions to all candidate UEs increase with the number of UEs. Also, the scheme considers only fixed UEs.

Second, the proposed scheme in [43] enables 2-step random access by representing a preamble ID by bits and dividing the bits into two parts: ID part and information part. ID part is dedicated to the single UE, and information part conveys the purpose of random access and buffer status of the UE. As the existing 64 preambles (corresponding to 6 bits) cannot cover both information and unique IDs for many UEs, we have proposed a preamble sequence generation method. This scheme does not suffer from both the *undetected collision* and *detected collision* problem because the gNB can not detect the same preamble at the same time. The limitation is that the preamble generation is possible only in an ultra-dense network scenario where the cell density is higher than 1000 cells/km^2 [43].

3.3 Challenges of 2-Step Random Access

In this section, we present the challenges of 2-step random access introduced in this article.

3.3.1 Preamble Allocation

For the preamble of msgA, 2-step CBRA uses a disjoint set of preambles from 4-step CBRA out of 64 preambles, so a preamble allocation problem arises. For preamble allocation, we consider the incidence ratio between 2-step CBRA and 4-step CBRA in a specific cell. Similar preamble allocation problems have been studied in our previous work. In [42], a subset of preambles is allocated to contention resolution-based random access considering the number of UEs and their traffic characteristics. In [43], each UE is assigned a unique ID part and multiple preambles have the same ID.

3.3.2 Resource Mapping for msgA

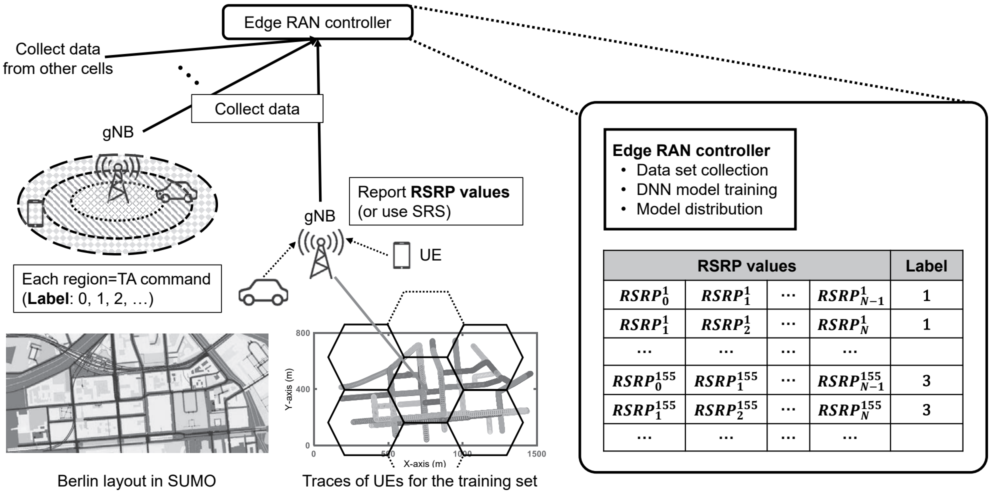
5G NR has two types of resource mapping between preamble ID(s) of a specific RO and a PRU: many-to-one and one-to-one. Many-to-one mapping maps two or more preamble IDs of a specific RO to one PRU ID, while one-to-one mapping maps one preamble ID of the specific RO to one PRU ID. We select one of two configurations according to the number of UEs simultaneously attempting 2-step CBRA. If the number is small, many-to-one mapping is suitable for efficient use of PUSCH resources. Otherwise, one-to-one mapping is appropriate. The use of many-to-one mapping increases the probability that different UEs use the same PUSCH resource to send their payloads, resulting in a collision.

3.3.3 DFT Operation in gNB

2-step CBRA uses the TA value of zero for the msgA payload. Thus, the GP may exist at the end of the PUSCH payload according to the parameters such as the numerology, delay spread, and coverage of the gNB. The gNB determines where to locate DFT windows in the payload including the GP using the preamble reception timing. For instance, in the case (b) of Fig. 3.3, the gNB measures the difference in timing of preamble reception between UE 2 and UE 1 greater than the CP length, and so performs additional DFT operation. Meanwhile, if the UE could transmit the msgA payload by well determining the TA value for the area it is located at, the GP would be unnecessary. Then, the gNB takes advantage of resources to transmit and receive other data.

3.3.4 Detected Collision Problem

The *detected collision* problem occurs when UEs at different distances from the gNB send the same preamble. There are two works that solve the *detected collision* prob-



RSRP values				Label
$RSRP_0^1$	$RSRP_1^1$	⋯	$RSRP_{N-1}^1$	1
$RSRP_1^1$	$RSRP_2^1$	⋯	$RSRP_N^1$	1
⋯	⋯	⋯	⋯	
$RSRP_0^{155}$	$RSRP_1^{155}$	⋯	$RSRP_{N-1}^{155}$	3
$RSRP_1^{155}$	$RSRP_2^{155}$	⋯	$RSRP_N^{155}$	3
⋯	⋯	⋯	⋯	

Figure 3.5: Overview of the training procedures with the mobility model and topology.

lem [42], [48]. In [42], however, there is a limitation that the approach cannot be applied to mobile UEs. This limitation can be avoided as each UE estimates the TA command, and selects and transmits a preamble from a specific set of preambles corresponding to the estimated TA command. Upon receiving the preamble from the specific preamble set, the gNB implicitly notices the TA command of the UE. The authors in [48] resolve the *detected collision* problem by transmitting separate RARs under the assumption that all UEs calculate their own TA value. Applying the approach in [48] to 2-step CBRA, we can reduce the preamble collision probability. Therefore, we propose the self-uplink synchronization framework, aiming to overcome these limitations in the next section.

3.4 EsTA: Proposed Self-UL Synchronization Procedure

The self-uplink synchronization framework that we propose in this section helps a UE determine a TA value within the timing error limit defined in 5G NR [49].

3.4.1 Overview

In the proposed framework, the UE estimates the TA command using a DNN model and reference signal received power (RSRP) values. The input of the DNN is RSRP values and the output is an estimated TA command ($\hat{T_A}$). As shown in Fig. 3.5, the edge RAN controller collects labeled data sets from gNBs, consisting of UEs' RSRP values and corresponding TA commands. After training the DNN with data sets, the edge RAN controller distributes the DNN model to UEs through the gNBs.

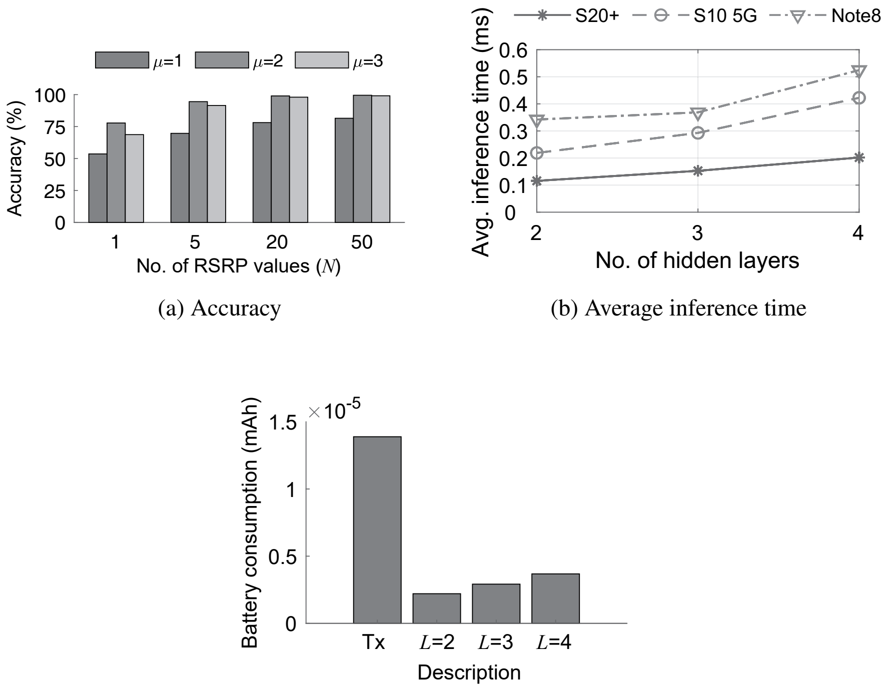
3.4.2 Overall Procedures

The procedures for estimating the TA command based on RSRP values are as follows:

- A UE in the RRC connected state [14] periodically reports its RSRP values to the connected gNB. Then, the gNB periodically sends a set of N RSRP values and the TA command (label) for each UE to the edge RAN controller. The label is the same as the TA command reported when the gNB received the latest RSRP value for a specific set. In Fig. 3.5, $RSRP_j^i$ represents the j-th RSRP value of UE i.

- With sufficiently collected labeled data, the edge RAN controller trains the DNN model. We consider a network with three hidden layers and one output layer. Each hidden layer is 200-way fully-connected. The outcome of each hidden layer is followed by *ReLU* activation function. The output layer uses *softmax* activation whose output is a probability vector for the TA command. We select a cost function with cross-entropy, and apply the Adam optimization algorithm for training.

- The edge RAN controller notifies each UE of the information about this model such as number of layers, weight matrix, and so on. It also updates the DNN model regularly or when needed.



(a) Accuracy

(b) Average inference time

(c) Battery consumption

Figure 3.6: Performance of the DNN model.

- After receiving the model information, the UE obtains the estimated TA command from the DNN model and N RSRP values.

We consider the gNB to collect RSRP values directly by measuring the received power of the sounding reference signal (SRS) using channel reciprocity, i.e., without RSRP reporting from UEs .

3.4.3 Performance Evaluation

Simulation environments: We use OpenStreetMap (OSM) provided by simulation of urban mobility (SUMO), linked to the actual map information [50]. SUMO can reflect real environments including vehicle movements and traffic lights. The total time of tracing is one hour, and the number of UEs is 155. For the channel model, we create the path loss and shadowing following 3D-UMa model, and consider the fast fading channel model generated using ITU-R IMT UMa model used in [42]. Channel environments, i.e., line-of-sight (LOS) or non-LOS (NLOS), are stochastically determined, depending on the distance between UEs and the gNB. We use the hexagrid topology for the seven cell deployment as shown in Fig. 3.5. The innermost cell is placed in the center of the trace map of UEs and the inter-site distance is 500 m. We consider frequency range 1 [46] because an application using small packets is appropriate for 2-step random access.

Measurement model: As assumed in [51], we simplify the RSRP formula by considering the assumption of the physical layer that the channel is flat within a PRB. This means that all resource elements within the PRB have the same power. The RSRP value is calculated as the sum of all PRBs' powers for the synchronization signal divided by the number of PRBs [52]. The period for the RSRP update is 5 ms, which is the shortest period for synchronization signals [46].

TA granularity: When the subcarrier spacing is 15 kHz, the TA granularity is 0.52 μs [47]. Because the TA value should be twice the propagation delay, the distance corresponding to the TA granularity is 78 m ($= 0.52 \times 3 \times 10^2/2$). When the subcarrier spacing values are 30 kHz and 60 kHz, the corresponding distances are 39 m and 19.5 m, respectively. When the DFT size is 4096, the TA granularity is given by 32 time samples regardless of subcarrier spacing. The value 32 is obtained by dividing the TA granularity by the sampling time for subcarrier spacing.

Accuracy: When measuring the accuracy of TA command estimation, we consider the UE initial transmission timing error less than or equal to $\pm T_e$, where T_e is the timing error limit value [49]. When the subcarrier spacing values are 15, 30, and 60 kHz, the corresponding timing error limit values are 24, 40, and 80 time samples, respectively. For the subcarrier spacing values of 30 and 60 kHz, because the timing error limit values are greater than the TA granularity, the use of TA values corresponding to different TA commands from the notified TA command is tolerated. Fig. 3.6(a) shows the accuracy of TA command estimation according to the subcarrier spacing and the number of RSRP values (N) for TA command estimation. As the number of RSRP values increases, estimation accuracy increases. We find that past RSRP values help with classification, and estimation accuracy converges when N is 50. The estimation accuracy increases to 99.6% and 99.1% for 30 and 60 kHz subcarrier spacing values, respectively, while it is 81.5% for 15 kHz due to its short timing error limit value.

Inference time: We observe the inference time according to the number of hidden layers to see how the complexity of DNN model affects mobile devices. In Fig. 3.6(b), as the number of hidden layers (L) increases, the average inference time increases. In the case of the most recent device, i.e., Galaxy S20+ released in 2020, the average inference time is 0.15 ms when L is set to three in the proposed framework. Because this value is smaller than the slot length when the subcarrier spacing is 60 kHz (0.25 ms), the UE can estimate the TA command in one slot after receiving the latest reference signal.

Battery consumption: Using "Batterystats" tool included in the Android framework, we observe the battery consumption (mAh) of the CPU over time for an application that runs the DNN model. We measure the battery consumption of Galaxy S20+ during 100,000 iterative inference operations. Through the measurement, we obtain the discharge current using battery consumption and application end time, which is (battery consumption)$\times$3600/(application end time). The discharge current for inference operation is 65.66–68.76 mA. Then, we can obtain the battery consumption for one inference operation using the discharge current and average inference time of Fig. 3.6(b). Fig. 3.6(c) shows the battery consumption for transmission of UE and one inference operation according to L. The discharge current of the UE for transmission is 100 mA [53]. When L is set to three in the proposed framework, the UE can estimate the TA command by consuming about 21% of the battery consumption for transmission.

Network performance: We perform system-level simulation considering the in-

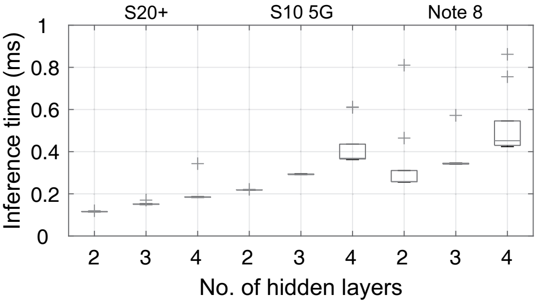


Figure 3.7: Distribution of inference time.

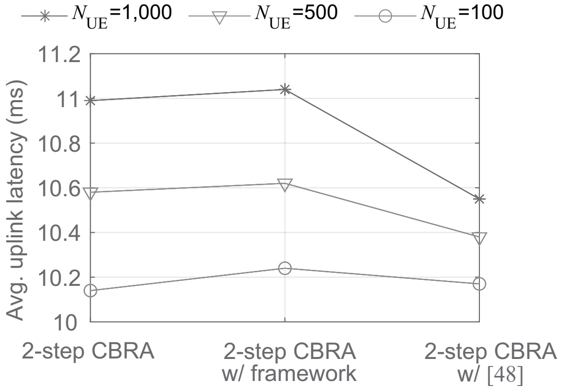


Figure 3.8: Average uplink latency.

ference time as well as the estimation error of the TA command. We use measurements of the inference time for S20+ (L=3) as shown in Fig. 3.7. Simulation parameters such as event-driven traffic arrival rate and uplink latency components follow our previous work [42].

We consider three cases depending on whether to apply the proposed framework and spatial group-based reusable preamble allocation [48]: i) 2-step CBRA, ii) 2-step CBRA with the proposed framework (2-step CBRA w/ framework), and iii) 2-step CBRA with [48] (2-step CBRA w/ [48]). Fig. 3.8 represents the average uplink latency according to the number of UEs (N_{UE}) for each case. When we apply the proposed framework, i.e., 2-step CBRA w/ framework, the average uplink latency value increases further, compared to 2-step CBRA due to errors in the estimation of TA command and inference time. In the case of 2-step CBRA w/ [48], when the number of UEs are 100, 500, and 1,000, the average uplink latency values decrease by 0.7%, 2.3%, and 4.4%, respectively. This is because the *preamble collision* problem occurs more frequently as the number of UEs attempting 2-step CBRA increases.

3.4.4 Future Research Perspectives

We summarize future research directions as follows.

Performance enhancement: To further reduce symbol errors caused by ISI at the gNB when multiple UEs transmit the msgA payload with their estimated TA values, we can apply a rule-based approach combined to the DNN model. This approach helps a UE determine its TA value more precisely by taking the probability distribution of the TA command obtained from the DNN model into account.

Generalization: To make the model easily adapt to new environments, we may adopt meta-learning [54], also known as "learning to learn". Specifically, by collecting different data sets for diverse network environments and applying a meta-learning method, e.g., model-agnostic meta-learning (MAML) [54], we can quickly initialize the model to a state trainable with a few data sets.

Real-world environments: The Android application programming interface (API) provides the RSRP and TA command (*CellSignalStrengthLte* class), so we can collect data sets through measurements in real-world environments. We should consider whether a data set is useful, taking into account its characteristics such as the time interval for updating RSRP values during measurements.

3.5 Summary

In 5G NR, the 3GPP included 2-step CBRA to further improve latency for channel access compared with 4-step CBRA. We introduced the newly defined messages and corresponding channel structure for 2-step CBRA. We presented 2-step random access schemes proposed in the recent literature to tackle the *preamble collision* problem that occurs when many UEs try 2-step and 4-step CBRA. We listed the challenges of the 2-step random access schemes, and proposed the self-uplink synchronization framework that allows a UE to determine its TA value to solve the *preamble collision* problem, using the DNN model. Lastly, we summarized the research directions to improve the performance of the proposed framework and generalize it for real world environments.

Chapter 4

IBA: Interference-Aware Beam Adjustment for 5G mmWave Networks

4.1 Introduction

In the last few years, mmWave communication attracts considerable interest, thanks to the high data rates enabled by its large available bandwidth [55], [56]. This makes mmWave a key technology for 5G NR systems. Most of the previous work focused on developing beamforming technologies between transmitter and receiver. In particular, research has been conducted to overcome NLOS environment because mmWave communications is the severe propagation attenuation caused by high path loss, shadowing and blockages [57]. That is, gNBs perform beam selection based on their surrounding environment.

Meanwhile, as the mmWave network becomes denser, co-channel interference becomes a factor limiting performance [7]. To overcome this, many studies have been conducted to overcome interference through scheduling by dividing frequency resources or time resources [7], [8]. However, since this approach uses time-frequency resources separately, it may not be possible to obtain the maximum throughput performance. We investigate whether the performance of the beam pair currently in use can be improved by simply changing the beam pair when the throughput performance deteriorates due to interference.

In this chapter, we propose interference-aware beam adjustment (IBA), a method of beam adjustment to coordinate interference in 5G mmWave networks. Basically, in 5G NR, when determining the beam pair of the link between the transmitter and the receiver by measuring the received power of beam candidates. Since this process called beam adjustment takes only the desired link into consideration, it may be vulnerable to interference. To reduce interference, transmitter or receiver can select other beams to overcome degradation by interference. The complexity of the IBA depends on the number of beams used in the beam pair. Because mmWave uses a number of beams, it is practically impossible to overcome interference by changing all of the beam pair candidates. Therefore, we introduce reducing search space for finding new beam pair to coordinate interference.

Rest of this chapter, in Section 4.2, we introduce beam management defined in 5G NR. In Section 4.3, we notice throughput degradation in simple topology in mmWave network. Next, in Section 4.4, we present IBA and observe whether the beam search space can be reduced. Through the observation, in Section 4.5, we demonstrate the performance of IBA by measuring throughput in an environment with interferers.

4.2 Background

4.2.1 Beam Management in 5G NR

Multi-antenna precoding should realistically allow fine-grained control, including both phase adjustment and amplitude scaling, of the different antenna elements. For this purpose, multi-antenna processing at the transmitter side is carried out in the digital domain before digital-to-analog conversion. However, in the case of operation at higher frequencies with a large number of closely space antenna elements, fully digital precoding for each antenna appears to be infeasible [58]. Therefore, the antenna

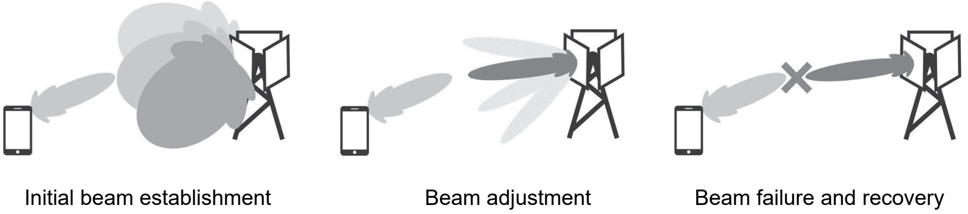


Figure 4.1: Beam management in 5G NR.

processing will rather be carried out in the analog domain with focus on beamforming [46].

As analog antenna processing will be carried out on a carrier basis, this also implies that beamformed transmission can only be done in one direction at a time. Downlink transmissions to different devices located in different directions relative to gNB must therefore be separated in time. Likewise, in the case of analog-based receiver-side beamforming, the receive beam can only focus in one direction at a time. Under these conditions, beam management is to establish and retain a suitable beam pair, that is, a transmitter-side beam direction and a corresponding receiver-side beam direction.

Fig. 4.1 represents beam management defined in 5G NR. Initial beam establishment includes the procedures by which a beam pair is initially established between a gNB and a UE. During initial cell search, the UE receives multiple SS blocks being transmitted in sequence within different downlink beams. Each such SS block, in practice the different downlink beams, is associated with a corresponding random access occasion and preamble. Through the subsequent preamble transmission in random access, the gNB identifies the downlink beam acquired by the UE.

Once an initial beam pair has been established, there is a need to regularly adjust beam pair due to movements and rotations of the UE. Furthermore, even for stationary UEs, movements of other objects in the environment could block the current beam pairs. This beam adjustment also include refining the beam shape, for example making the beam more narrow compared to a relatively wider beam used for initial beam establishment [46]. Beam adjustment can be divided into two separate procedures:

- Transmitter beam adjustment: transmitter-side beam adjustment aims at refining the gNB transmit beam, given the receiver beam currently used at the UE side.

- Receiver beam adjustment: receiver-side beam adjustment aims at finding the best receive beam, given the current transmit beam.

There are some cases where a currently established beam pair being rapidly blocked without sufficient time for the beam adjustment to adapt. The 5G NR includes specific procedures to handle such beam-failure events, also referred to as beam failure recovery. In beam failure recovery, the UE and gNB reselect a beam pair through random access.

4.2.2 System-Level Simulation and 3D Beamforming for 5G NR

We use up-to-date network simulator-3 (ns-3)-based system level simulator for 5G NR [59]. This simulator includes not only NR physical layer abstraction model but also NR frame structure. However, the three-sector model is not implemented in the antenna array model, so we added that model function by referring [60]. In this chapter, we apply 3D beamforming consisting of vertical (θ) and horizontal (ϕ) angles. The number of vertical angles for both Tx and Rx is 4. The number of horizontal angles for Tx and Rx is 32 and 8, respectively. Therefore, the number of all combinations is 4096 (=2^{12}).

4.3 Motivation

4.3.1 Throughput Degradation by Interference

As shown in Fig. 4.2, we first investigate interference affection under simple two cells scenario. Inter-site distance (ISD) between two cells is 60 m, and the distance between

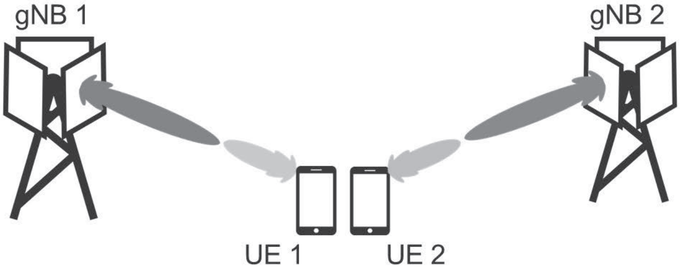
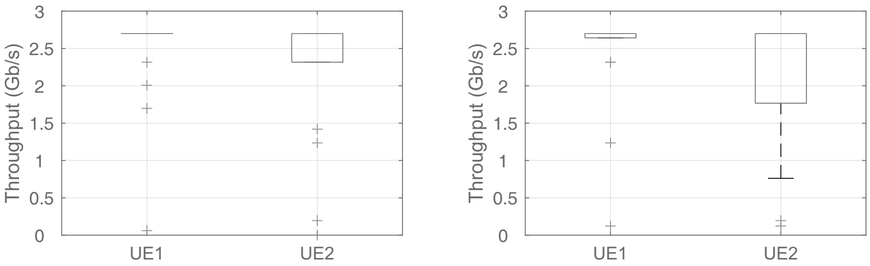


Figure 4.2: Scenario for interference affection.



(a) Single cell

(b) Two cells

Figure 4.3: Throughput of each UE in two cells.

gNB and UE is 20 m, respectively. Carrier frequency is 28 GHz and the system bandwidth is 800 MHz. Each gNB performs downlink transmission with source rate 3 Gb/s. We observe throughput of each UE while changing the channel condition, i.e., LOS or NLOS. Fig. 4.3 shows the throughput distribution of each UE with or without interferer. The top line of box plot represents median, which is middle of data set. The bottom line of box plot is lower quartile, which is 25% of data less than this value. The bottom of the dashed line is minimum, which is (lower quartile)-1.5×(inter-quartile range). The cross mark is outlier, which is less than the minimum.

Fig. 4.3(a) shows the throughput of each UE without interferer. If the channel condition is NLOS, throughput performance is degraded, but 75% of UE 1's throughput data represent 2.7 Gb/s. In the case of UE 2, the lower quartile is 2.32 Gb/s. If the interferer exists, throughput of UE 2 is more degraded than that of UE 1. Fig. 4.3(b) shows the lower quartile is 1.76 Gb/s. We notice that although gNBs and UEs use beamforming technology, UEs located at the cell edge can be affected by interference from neighboring cells.

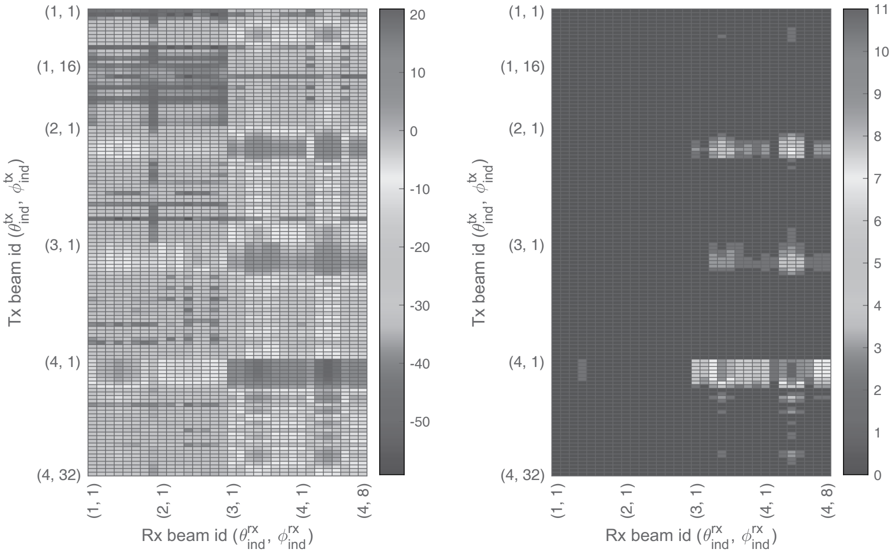
4.4 IBA: Proposed Interference Management Scheme

4.4.1 Overall Procedure

After performing the beam adjustment, a UE determines whether to perform additional beam adjustment according to the change in the channel quality indicator (CQI) value and the interference level. If the CQI value does not drop even when the interference level increases, the throughput performance is maintained, so both values should be observed at the same time. When the throughput performance degrades by observing the change in the CQI value and interference level, it is necessary to find a new beam pair candidate. Therefore, the performance of IBA depends on how quickly it finds the beam pair to recover CQI value. We need to set an appropriate search space and apply a new beam pair. In order to get the best performance, it is necessary to apply beam pairs in all cases (4096), but this is practically impossible. For instance, if we consider the shortest CQI update period, i.e., 4 slots, it takes more than 4 s to apply and check all beam pairs when the slot length is 0.25 ms. Therefore, we need to find the local optimal rather than the global optimal.

4.4.2 Reduction of Search Space

The most efficient way to reduce the search space is to adjust the beam pair of the desired link. The reason why it is inefficient to adjust the beam pair of the interference link is that when the number of interference link increases to N, the complexity becomes 4096^{N+1}. Therefore, we figure out if the throughput of UE 2 in Fig. 4.2 can be improved by adjusting the beam pair of the desired link. While changing the desired



(a) SINR

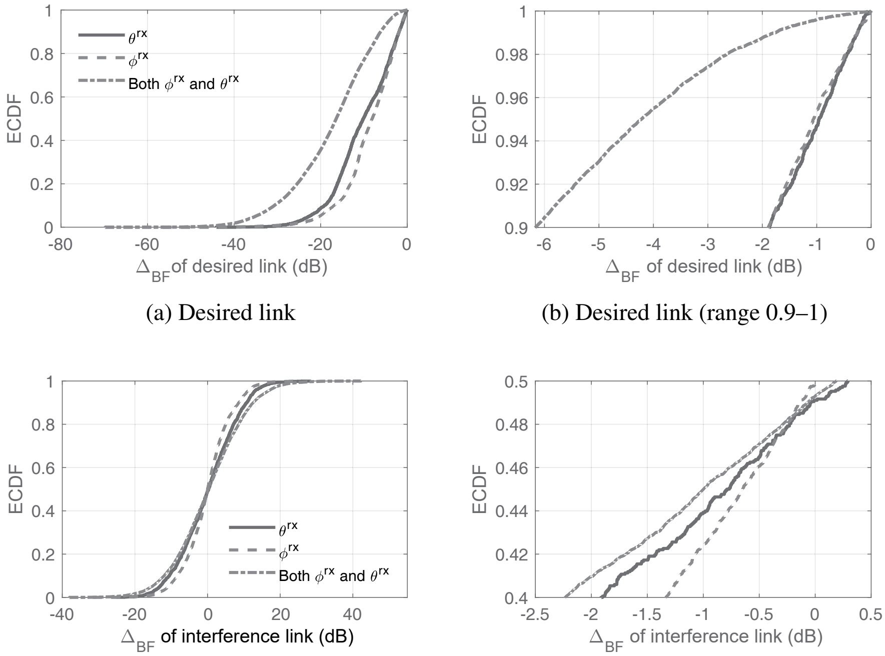
(b) CQI

Figure 4.4: SINR and CQI of UE 2 when interference link is fixed.

link of the UE 2 that is affected by the interferer, we observe the SINR value of the UE 2 and corresponding CQI.

Figs. 4.4(a) and 4.4(b) represent UE 2's SINR and CQI according to beam pair of desired link. The optimal beam pair derived from beam adjustment is ($\theta_{\mathrm{ind}}^{\mathrm{tx}}$, $\phi_{\mathrm{ind}}^{\mathrm{tx}}$)=(4,4) and ($\theta_{\mathrm{ind}}^{\mathrm{rx}}$, $\phi_{\mathrm{ind}}^{\mathrm{rx}}$)=(3,4). The corresponding CQI is 9. If we change $\theta_{\mathrm{ind}}^{\mathrm{rx}}$ from 3 to 4, CQI changes to 11. Changing $\theta_{\mathrm{ind}}^{\mathrm{rx}}$ and $\phi_{\mathrm{ind}}^{\mathrm{tx}}$ at the same time can also make CQI 11. Changing both the Tx beam and the Rx beam still takes a lot of time because 4096 beam pairs have to be compared. The UE should inform the gNB to change the Tx beam. This operation takes more time to apply a new beam pair. Also, the new Tx beam of the gNB can provide a higher interference level to the UEs of adjacent cells. So, we observe that changing the Rx beam can generally improve the throughput performance.

Since the change of the CQI value is determined according to the SINR, we separately observe the gain of the desired link and the interference link determining the SINR. There are two parameters in Rx beams, which are θ^{rx} and ϕ^{rx}. We observe how



(a) Desired link

(b) Desired link (range 0.9–1)

(c) Interference link

(d) Interference link (range 0.4–0.5)

Figure 4.5: Beamforming gain difference for desired and interference links.

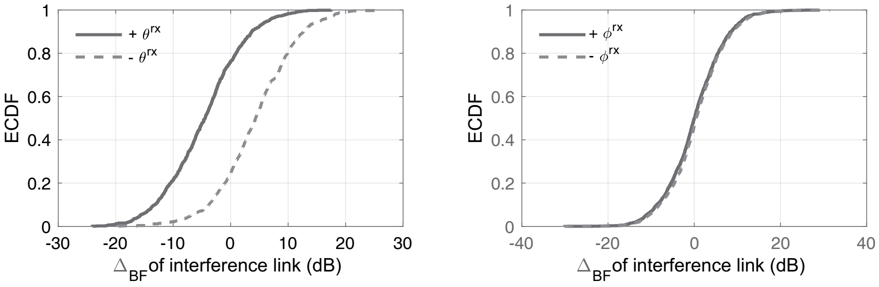
each link gain changes as the above angle changes. There are three cases: i) θ^{rx} ii) ϕ^{rx}, and iii) both θ^{rx} and ϕ^{rx}.

Fig. 4.5(a) represent the beamforming gain difference (Δ_{BF}) of desired link when Rx beam changes. The beamforming gain difference is represented by

$$\Delta_{\mathrm{BF}} = G(\theta_i, \phi_i) - G(\theta_{\mathrm{fix}}, \phi_{\mathrm{fix}}), \tag{4.1}$$

where $G(\theta, \phi)$ is beamforming gain when the vertical and horizontal angles are θ and ϕ, respectively. For the desired link, θ_{fix} and ϕ_{fix} are the angles of Rx beam determined by beam adjustment. Through Fig. 4.5(b), we can notice that case iii) is not appropriate due to significant decrease than other cases. In desired link, desired link reduction trend is similar for both case i) and ii).

Fig. 4.5(c) represent Δ_{BF} of interference link when Rx beam changes. At this time,



(a) θ

(b) ϕ

Figure 4.6: Beamforming gain difference according to beam change direction.

θ_{fix} and ϕ_{fix} are angles of randomly selected Rx beam. This is because interference gain is randomly determined according to the link between the UE and the interferer. Through Fig. 4.5(d), we can notice that case ii) is more proper than case i) due to more decrease in interference. Therefore, we apply the IBA algorithm to change θ^{rx} first before changing ϕ^{rx}. As shown in Fig. 4.6, we also observe the beamforming gain difference of interference link according to beam change direction. In the case of θ^{rx}, changing the direction in which the index increases is more likely to reduce interference than the direction in which the index decreases. In the case of ϕ^{rx}, the trend is similar regardless of direction.

4.4.3 Algorithm for IBA

Through the observation in Section 4.4.2, we design the IBA algorithm considering some options. Table 4.1 shows six options for IBA algorithm. First, the condition under which the IBA is executed are the interference level is increased and the CQI value should be lower than before. After the IBA is executed, if the current CQI value is less than or equal to the previous CQI value, the UE changes the index of the angle corresponding to the next step. Options 1–3 only change θ and options 4–6 change not only θ but also ϕ (or both of them in option 6). Other options, except option 2, work by reusing the Rx beam of the previous step ((n-1)th) in the next step ((n+1)th), unless

Table 4.1: Options for IBA algorithm.

Option	Angle	Step				
		1	2	3	4	5
1	θ	+1	-1	+2	-2	-
2	θ	+1	+2	+3	+4	-4
3	θ	+1	-1	-	-	-
4	θ	+1	-1	-	-	-
	ϕ	-	-	+1	-1	-
5	θ	+1	-1	-	-	-
	ϕ	-	-	-1	+1	-
6	θ	+1	-1	+1	-1	-
	ϕ	-	-	+1	-1	-

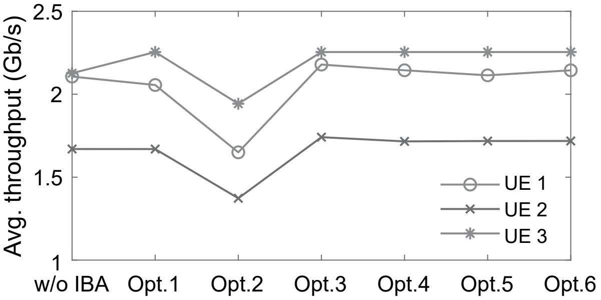


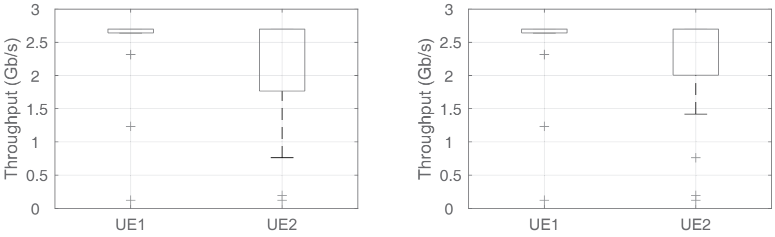
Figure 4.7: Average throughput according to options for IBA.

performance improves in a particular step (n-th).

To evaluate the options for IBA algorithm, we consider three cells scenario. The topology follows 3D-UMi model [61], and the IDS is 60 m. Fig. 4.7 represents average throughput of each UE. We can find that opt. 3 the highest gain, i.e., 3.4–6%, in terms of average throughput. Therefore, in Section 4.5, we adopt opt. 3 for IBA algorithm.

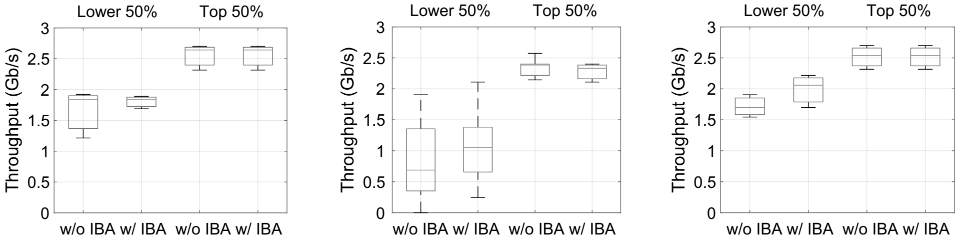
4.5 Performance Evaluation

Two cells: Fig. 4.8(b) shows the throughput distribution of each UE when applying IBA. As UE 2 performs IBA, the lower quartile of throughput data is improved by 13.6% compared with only applying beam adjustment. The minimum of throughput



(a) Beam adjustment (b) IBA

Figure 4.8: Throughput of each UE in two cells.



(a) UE 1 (b) UE 2 (c) UE 3

Figure 4.9: Throughput of each UE in three cells.

data is improved 86.5% compared with only applying beam adjustment.

Three cells: Fig. 4.9 represents the throughput distribution of each UE. For each UE, the leftmost two box plot is the throughput distribution of the lower 50% and the rest is the throughput distribution of the top 50%. In the case of UE 1, minimum value of lower 50% increased by 40%. In the case of UE 2, the median value decreased by 2% in the top 50% throughput, but the median value increased by 54% in the lower 50% throughput. The reason that the throughput of the top 50% decreases slightly is IBA searches for new rx beam in each step and returns to the original rx beam when performance deteriorates. UE 3 also increases median value by 21% for lower 50% throughput.

4.6 Summary

We propose IBA, which is a interference-aware beam adjustment method to coordinate interference in 5G mmWave networks. It is important to reduce search space to find new beam pair for applying IBA. To this end, we observed the change in gain of the desired link and interference link while changing the Tx and Rx beam pairs. Through the simulation, it has been shown that it is appropriate to adjust the Rx beam of the desired link to control interference.

Chapter 5

Concluding Remarks

5.1 Research Contributions

In this dissertation, we have addressed

In Chapter 2, we have proposed *RAPID*, a two-step random access procedure for delay-sensitive UEs in RRC_INACTIVE state introduced in 5G. *RAPID* completes the random access procedure by exchanging two messages using AS context ID of UE in RRC_INACTIVE state. The proposed scheme can play important roles in satisfying the latency requirements of various applications targeted in 5G.

In Chapter 3, we have proposed EsTA, a framework that helps UE to estimates TA command and determine TA value. Through EsTA, we can solve preamble collision problem. By applying EsTA to *RAPID*, it can support mobile UEs.

In Chapter 4, we have proposed IBA which adjust beam pair to coordinate interference in 5G mmWave networks. To this end, we observed the change in gain of the desired link and interference link while changing the Tx and Rx beam pairs. Through the simulation, it has been shown that it is appropriate to adjust the Rx beam of the desired link to control interference.

5.2 Future Work

As further improvement on the results of this dissertation, there are several research items as follows.

First, regarding EsTA, we need to enhance the estimation accuracy when SCS is 15 kHz.

Second, regarding network performance derived from EsTA, we need to observe link-level performance affected by inter-symbol interference.

Lastly, regarding beam adjustment, we need to perform more simulation in various scenarios where the interference links are more dynamic. Also, we will consider reinforcement learning in how to determine the angle and index to adjust when changing the beam.

Bibliography

[1] 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 Commun. Surveys & Tuts.*, vol. 16, no. 1, pp. 4–16, 2014.

[2] 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 Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, 2014.

[3] J. S. Kim, S. Lee, and M. Y. Chung, “Efficient random-access scheme for massive connectivity in 3GPP low-cost machine-type communications,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6280–6290, 2017.

[4] 3GPP RP-190711, “3GPP Work item description, 2-step RACH for NR,” Sep. 2019.

[5] S. Mukherjee, A. K. Shinha, and S. K. Mohammed, “Timing advance estimation and beamforming of random access response in crowded TDD massive MIMO systems,” *IEEE Trans. on Commun.*, vol. 67, no. 6, pp. 4004–4019, 2019.

[6] R. Keating, M. Säily, J. Hulkkonen, and J. Karjalainen, “Overview of positioning in 5G new radio,” in *Proc. IEEE International Symposium on Wireless Communication Systems (ISWCS)*, 2019, pp. 320–324.

[7] W. Feng, Y. Wang, D. Lin, N. Ge, J. Lu, and S. Li, “When mmWave communications meet network densification: A scalable interference coordination perspective,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1459–1471, 2017.

[8] Z. Sha, Z. Wang, S. Chen, and L. Hanzo, “Early-late protocol for coordinated beam scheduling in mmWave cellular networks,” in *Proc. GLOBECOM*. IEEE, 2019, pp. 1–6.

[9] T. Taleb and A. Kunz, “Machine type communications in 3GPP networks: potential, challenges, and solutions,” *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, 2012.

[10] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021,” *White paper*, 2017.

[11] 3GPP TR 22.891, “Feasibility study on new services and markets technology enablers; stage 1,” ver. 14.2.0, Sept. 2016.

[12] I. L. Da Silva, G. Mildh, M. Säily, and S. Hailu, “A novel state model for 5G radio access networks,” in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2016.

[13] J. Kim, D. Kim, and S. Choi, “3GPP SA2 architecture and functions for 5G mobile communication system,” *ICT Express*, vol. 3, pp. 1–8, 2017.

[14] 3GPP TS 38.331, “NR; radio resource control (RRC); protocol specification,” ver. 15.8.0, Jan. 2020.

[15] 3GPP R2-1800214, “RRC_INACTIVE context ID,” Jan. 2018.

[16] 3GPP TS 36.321, “E-UTRA; medium access control (MAC) protocol specification,” ver. 15.6.0, Jan. 2020.

[17] 3GPP R2-166236, “NR control plane latency in new RRC state,” Oct. 2016.

[18] 3GPP TS 38.211, “NR; physical channels and modulation,” ver. 15.1.0, Apr. 2018.

[19] 3GPP TS 36.912, “Feasibility study for further advancements for E-UTRA (LTE-Advanced),” ver. 14.0.0, Mar. 2017.

[20] 3GPP R1-1700652, “On 2-step random access procedure,” Jan. 2017.

[21] 3GPP R1-1700892, “NR 2-step random access procedure,” Jan. 2017.

[22] J. Zhang, L. Lu, Y. Sun, Y. Chen, J. Liang, J. Liu, H. Yang, S. Xing, Y. Wu, and J. Ma, “5G PoC of SCMA-based uplink grant-free transmission UCNC framework,” in *Proc. IEEE VTC-Spring*, June 2017.

[23] 3GPP TS 36.331, “E-UTRA; radio resource control (RRC); protocol specification,” ver. 15.1.0, Apr. 2018.

[24] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, “Traffic models for machine type communications,” in *Proc. VDE International Symposium on Wireless Communication Systems (ISWCS)*, Aug. 2013.

[25] W. W. Piegorsch, *Statistical data analytics: foundations for data mining, informatics, and knowledge discovery*. John Wiley&Sons, 2016.

[26] J.-P. Cheng *et al.*, “Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks,” in *Proc. IEEE GLOBECOM Workshops*, Dec. 2011.

[27] E. Dutkiewicz *et al.*, “Massive machine-type communications,” *IEEE Network*, vol. 31, no. 6, pp. 6–7, 2017.

[28] C. Anton-Haro and M. Dohler, *Machine-to-machine (M2M) communications: architecture, performance and applications*. Elsevier, 2014.

[29] A. Kumar *et al.*, “An online delay efficient packet scheduler for M2M traffic in industrial automation,” in *Proc. IEEE SysCon*, Apr. 2016.

[30] K. Zhou and N. Nikaein, “Low latency random access with TTI Bundling in LTE/LTE-A,” in *Proc. IEEE ICC*, May 2015.

[31] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, 2000.

[32] 3GPP TR 37.868, “Study on RAN improvements for machine-type communications,” ver. 11.0.0, Sept. 2011.

[33] P. Osti, P. Lassila, S. Aalto, A. Larmo, and T. Tirronen, “Analysis of PDCCH performace for M2M traffic in LTE,” *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4357–4371, 2014.

[34] J. Meinila, P. Kyosti, L. Hentila, T. Jamsa, E. Suikkanen, E. Kunnari, and M. Narandzic, “D5.3: WINNER+ final channel models,” Wireless world initiative new radio WINNER, June 2010.

[35] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, and L. Chen, “Performance analysis of power saving mechanism with adjustable DRX cycles in 3GPP LTE,” in *Proc. IEEE VTC 2008-Fall*, Sept. 2008.

[36] A. Weber, P. Agyapong, T. Rosowski, G. Zimmerman, M. Fallgren, S. Sharma, A. Kousaridas, C. Yang, I. Karls, S. Singh *et al.*, “Performance evaluation framework,” *METIS*-II *Deliverable 2.1*, 2016.

[37] G. C. Madueño, Č. Stefanović, and P. Popovski, “Efficient LTE access with collision resolution for massive M2M communications,” in *Proc. IEEE GLOBECOM Workshops*, Dec. 2014.

[38] 3GPP TR 36.873, “Study on 3D channel model for LTE,” ver. 12.7.0, Jan. 2018.

[39] M. Sagraloff and K. Mehlhorn, “Computing real roots of real polynomials,” *Journal of Symbolic Computation*, vol. 73, pp. 46–86, 2016.

[40] A. Kobel, F. Rouillier, and M. Sagraloff, “Computing real roots of real polynomials... and now for real!” in *Proc. ACM ISSAC*, Jul. 2016.

[41] S. E. E. Ayoubi, S. Jeux, F. Marache, F. Pujol, M. Fallgren *et al.*, “Refined scenarios and requirements, consolidated use cases, and qualitative techno-economic feasibility assessment,” *METIS*-II *Deliverable 1.1*, 2016.

[42] J. Kim, S. Kim, T. Taleb, and S. Choi, “RAPID: Contention resolution based random access using context ID for IoT,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7121–7135, 2019.

[43] S. Kim, S. Kim, J. Kim, K. Lee, S. Choi, and B. Shim, “Low latency random access for small cell toward future cellular networks,” *IEEE Access*, vol. 7, pp. 178 563–178 576, 2019.

[44] 3GPP TS 38.300, “NR; NR and NG-RAN overall description; Stage 2,” ver. 16.0.0, Jan. 2020.

[45] 3GPP R1-2000151, “Final report of 3GPP TSG RAN WG1 #99,” ver. 1.0.0, Nov. 2019.

[46] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2018.

[47] 3GPP TS 38.213, “NR; physical layer procedures for control,” ver. 16.0.0, Jan. 2020.

[48] T. Kim, H. S. Jang, and D. K. Sung, “An enhanced random access scheme with spatial group based reusable preamble allocation in cellular M2M networks,” *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1714–1717, 2015.

[49] 3GPP TS 38.133, "NR; Requirements for support of radio resource management," ver. 16.2.0, Jan. 2020.

[50] Simulation of urban mobility. [Online]. Available: http://sumo.dlr.de

[51] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simulation Modelling Practice and Theory*, vol. 96, p. 101933, 2019.

[52] The Network Simulator-3. [Online]. Available: http://www.nsnam.org

[53] V. Rath and T. Shilpa, *Towards 5G: Applications, requirements and candidate technologies*. John Wiley & Sons, 2017.

[54] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. International Conference on Machine Learning*, Aug. 2017, pp. 1126–1135.

[55] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin *et al.*, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, 2017.

[56] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37 328–37 348, 2018.

[57] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast machine learning for 5G mmWave vehicular communications," in *Proc. IEEE INFOCOM*. IEEE, 2018, pp. 1961–1969.

[58] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[59] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, “An E2E simulator for 5G NR networks,” *Simulation Modelling Practice and Theory*, vol. 96, p. 101933, 2019.

[60] M. Rebato, M. Polese, and M. Zorzi, “Multi-sector and multi-panel performance in 5G mmWave cellular networks,” in *Proc. GLOBECOM*. IEEE, 2018, pp. 1–6.

[61] 3GPP TR 38.913, “Study on scenarios and requirements for next generation access technologies,” ver. 15.0.0, Jun. 2018.

초록

최근 사업자는 제조, 자동차, 헬스 케어 등 다양한 분야에서 5G 시스템을 사용하여 서비스를 만들고 있다. 5G 사용 사례에는 IoT 장치를 이용한 작은 패킷 전송에서 고화질 비디오 스트리밍과 같은 고속 데이터 전송까지 포함된다. 대규모 IoT 장치가 작은 패킷을 전송하는 경우 전력 소모 절약이 중요하므로 기지국과의 연결을 끊은 다음 랜덤 액세스를 통해 다시 기지국과 연결하여 데이터를 전송해야한다. 그러나 기존의 랜덤 액세스 절차는 다양한 지연시간 요건을 만족시키기 어렵다. 한편, 높은 데이터 전송 속도를 위해 넓은 대역폭의 밀리미터파 대역을 사용한다. 이때, 밀리미터파 대역 채널 특성을 극복하기 위해 빔포밍 기술이 적용된다. 그러나 현재 5G 표준에서 송신기와 수신기 사이의 빔 쌍을 결정할 때, 간섭은 고려되지 않는다. 이 논문에서는 5G 및 그 이후의 네트워크에서 다양한 사용 사례를 지원하기 위해 다음 세 가지 개선 사항을 고려한다. (i) 지연에 민감한 장치를 위한 2 단계 랜덤 액세스 절차, (ii) 프리앰블 충돌 문제를 해결하기 위한 자체 상향링크 동기화 프레임 워크, 그리고 (iii) 간섭을 줄이기 위한 간섭 인식 빔 조정이다. 첫째, 지연에 민감한 장치를 위한 2 단계 랜덤 액세스인 RAPID는 특정 신뢰도를 만족시키기 위한 지연시간을 줄이기 위해 제안되었다. RAPID와 경합 기반 랜덤 액세스를 수행하는 장치가 공존할 경우 RAPID가 랜덤 액세스 부하를 줄이기 위해 RAPID를 위해 할당되는 프리앰블 수를 결정하는 것이 중요하다. 시뮬레이션 결과에 따르면 RAPID는 99.999%의 신뢰도를 만족시키는 지연시간을 최신 기술에 비해 80.8% 줄이면서, 랜덤 액세스 부하를 30.5% 줄인다. 둘째, 프리앰블 충돌 문제를 해결하기 위해 자체 상향링크 동기화 프레임워크인 EsTA를 개발한다. 프리앰블 충돌은 여러 장치가 동일한 프리앰

블을 전송할 때 발생한다. 구체적으로, 단말이 심층 신경망 모델을 사용하여 timing advance(TA) command를 추정하고 TA값을 결정하는 프레임 워크를 제안한다. 네트워크 시스템의 부반송파 간격이 30 및 60 kHz 일 때, TA command 추정 정확도는 98–99%를 달성 할 수 있다. 마지막으로, 밀리미터파 네트워크에서 간섭을 줄이기 위한 간섭 인식 빔 조정 방법인 IBA를 제안한다. 시간과 주파수 자원을 다르게 예약하여 간섭을 줄이는 기존의 방법과 달리 IBA는 빔 조정을 통해 간섭을 제어한다. 이 때, 간섭을 줄이기 위해 새로운 빔 쌍을 찾는 검색 공간을 줄이는 것이 중요하다. 현실적으로 모든 빔 쌍의 조합을 검색하는 것은 불가능하다. 따라서 IBA는 Monte Carlo 방법을 통해 검색 공간을 축소하여 local optimum을 달성하도록 설계되어야한다. IBA는 5G 표준의 빔 조정 방법과 비교했을 때, 하위 50% throughput의 중간값이 최대 50%까지 향상된다. 요약하면, 우리는 5G 및 그 이후의 다양한 사용 사례를 위해서 2 단계 랜덤 액세스, 자체 상향링크 동기화 프레임 워크, 그리고 간섭 인식 빔 조정 방법을 제안한다. 이 연구를 통해 최신 기술에 비해 지연시간 및 처리량과 같은 네트워크 성능이 향상된다.

주요어: 5세대(5G), 랜덤 액세스, 부하 분석, 지도학습, 밀리미터파, 빔 관리

학번: 2014-21637

감사의 글

다사다난했던 6년 반의 연구실 생활을 마무리하며, 여기까지 올수 있게 도움을 주신 분들께 감사인사를 드립니다.

대학원 생활동안 훌륭한 세 분의 지도교수님들의 가르침을 받을 수 있었기에 더 성장할수 있었고, 인생을 살아가는데에도 많은 동기부여가 되었습니다. 먼저 작년 여름 힘들었던 시기를 잘 이겨낼 수 있게 도와주신 박세웅 교수님께 깊은 감사를 드립니다. 두 배로 커진 연구실을 이끌며, 항상 학생들과 소통하시고 한결같은 모습을 보여주신 교수님은 저에게 큰 귀감이 되었습니다. 학위논문을 준비하면서 부족하지만 항상 제가 진행하는 연구에 지지를 보내주시고, 대내외적으로 바쁘신 와중에도 논문 작성에 많은 도움을 주셔서 감사드립니다. 교수님과 함께한 시간이 짧아서 아쉽지만, 사회에 나가서도 교수님의 가르침대로 항상 진실되게 살아가는 제자가 되겠습니다. 졸업 후에도 좋은 소식 전해드리기 위해 노력하면서 학교에 와서 교수님과 함께 축구할 날을 기대하고 있겠습니다.

긴 시간동안 날카로운 통찰력과 열정으로 연구를 지도해주신 최성현 전무님께 깊은 감사를 드립니다. 현재의 위치에 안주하지 않고 더 성장하기 위해 노력하고 지도하시는 전무님의 모습을 보면서 대학원 생활의 어려움을 극복하는데 큰 힘이 되었습니다. 전무님과 함께 코드를 한줄 한줄 보면서 진행했던 미팅부터 학교를 떠나시기 전 논문 관련 마지막 미팅까지 모든 순간이 저를 성장시키는 계기가 되었습니다. 매순간 배웠던 것들을 잘 간직하여 사회에 나가서 잘 실천할 수 있는 제자가 되겠습니다.

대학원에 들어와서 논문과 친해질수 있게 도움을 주신 이병기 교수님께도 깊은 감사의 인사를 드립니다. 교수님과 함께한 논문세미나를 통해서 논문 읽는법을 알게

되었고, 세미나 후 점심식사를 함께하면서 교수님과 얘기를 나눌수 있어서 즐거웠습니다. 연말마다 교수님 그리고 연구실원들과 함께 사진도 찍고, 식사를 했던것도 좋은 추억으로 남아있습니다. 따뜻했던 기억들을 잘 간직하여 사회에 나가서도 베풀수 있는 제자가 되겠습니다.

짧은 시간이었지만 MWNL에서 함께 세미나에 참석하여 좋은 코멘트와 질문을 해주신 임종한, 최지웅 교수님께도 감사드립니다. 다음으로 바쁘신 시간 내어 학위논문 심사위원장을 맡아주신 심병효 교수님께 깊은 감사를 드립니다. 학생들에게 친근하게 다가가시고, 위트 있는 말씀으로 긴장을 풀어주시는 모습은 언제나 기억에 남을것 같습니다. 교수님께서 말씀하신대로 사회에 나가서도 의미있는 일을 하는 사람이 되도록 하겠습니다. 올해 3월부터 짧은 시간이었지만, 수 차례의 미팅에 흔쾌히 응해주신 이경한 교수님께도 깊은 감사를 드립니다. 교수님의 날카로운 분석과 통찰력을 통해서 연구에 좀 더 흥미를 끌어올릴수 있었고, 학위논문을 잘 마무리할 수 있었습니다. 마지막으로 학위논문 심사간 날카로운 질문과 사소한 실수도 잘 찾아내서 고쳐주신 오정석, 백정엽 교수님께도 깊은 감사를 드립니다. 두 분 교수님이 있어 저의 학위논문의 완성도를 한 단계 더 높일수 있었습니다.

세 연구실에 있으면서 같이 연구하고, 얘기할 수 있었던 연구실 구성원들 모두에게 감사인사를 전합니다. TSP에 지원을 하면서 동기인 호영이와 함께 선욱이형과 연철이에게 면접을 보고 카페에서 커피를 마셨던게 엊그제 같은데, 벌써 졸업학기가 되어 NETLAB의 신입생인 도균, 예린이를 보고 있으면, 정말 시간이 빠르게 지나간것 같습니다. TSP에 입학하여 연구실에 잘 적응할 수 있도록 많은 도움을 준 TSP 구성원분들께 감사 인사를 드립니다. MWNL에서 열정적이고 똑똑한 연구실 구성을들을 보면서 연구에 대한 동기부여를 더 높여 열심히 연구에 매진할 수 있었습니다. 연구뿐만 아니라 운동, 취미생활도 함께 하면서 더욱 돈독해 질수 있어서 좋았습니다. MWNL을 졸업한 그리고 같이 지냈던 모든분들께 감사 드립니다. 마지막 연구실인 NETLAB에서 새로운 연구실 구성원들과 1년 정도의 시간동안 친해지고, 서로 알아갈 수 있어서 좋았습니다. 모두 따뜻한 마음을 가지고 대해줘서 연구실에 잘 적응할 수 있었습니다. 연구실 구성원들 모두 원하는 바를 이루고 졸업을 하길 응원하겠습니다.

중앙대학교 전자전기공학부 선후배 그리고 요트부 구성원들께도 감사 인사드립니다. 특히 요트부 생활을 하면서 학교에 잘 적응하고, 새로운 활동을 하면서 좋은 사람들을 만날 수 있었던 계기가 되었습니다. 지금 사회 구성원으로서 역할을 하고 있는 111 ROTC 49기 동기들에게도 감사인사를 전합니다. 다들 각자의 위치에서 열심히 생활하는 모습을 지켜보면서 대학원 생활을 잘 해올 수 있었습니다.

다음으로 소중한 가족, 부모님께 감사드립니다. 항상 무한한 사랑을 주시며, 옆에서 묵묵히 응원하시는 부모님이 있었기에 여기까지 올 수 있었습니다. 진심으로 감사드립니다. 또 호담 공방의 대표인 동생 수정이도 고맙습니다. 앞으로 더 잘 돼서 돈도 많이 벌고 원하는 바 이루도록 옆에서 응원하겠습니다.

마지막으로 영혼의 단짝인 혜린이에게 감사드립니다. 항상 내가 선택하는 길을 존중해주고, 배려해 주었기에 힘든시간 잘 버텨내며 여기까지 올수 있었습니다. 지금까지 늘 그래왔듯 웃음이 넘치고, 행복이 넘치는 삶을 함께 살아가도록 노력하겠습니다.

긴 학위과정을 마치고, 사회에 첫발을 내딛는 시점에서 저에게 도움을 주고 응원해 주신 모든분들에게 감사드립니다.

2020 년 8월

김준석 올림