Enabling mMTC and URLLC in 5G: Initial Access, Traffic Prediction, and User Availability

Thilina Nuwan Weerasinghe

UiA University of Agder

Doctoral Dissertations at the University of Agder 321

Thilina Nuwan Weerasinghe

Enabling mMTC and URLLC in 5G: Initial Access, Traffic Prediction, and User Availability

Doctoral Dissertation for the Degree *Philosophiae Doctor (Ph.D.)* at the Faculty of Engineering and Science, Specialisation in Information and Communication Technology

University of Agder
Faculty of Engineering and Science
2021

Doctoral Dissertations at the University of Agder 321
ISSN: 1504-9272
ISBN: 978-82-8427-026-5

Printed by 07 Media
Oslo

Abstract

The 5th generation (5G) mobile communication networks focus on three main technology pillars as enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC), and massive machine type communications (mMTC). Among them, URLLC and mMTC introduce many novel challenging tasks for system and protocol design. This is a result of the stringent service requirements associated with URLLC and the massive number of devices that constitute mMTC. In this dissertation, we investigate three topics towards enabling mMTC and URLLC in 5G networks as stated below.

The first topic which is the main focus of this dissertation lies in proposing initial access approaches in URLLC and mMTC scenarios. Initial access plays an essential role for end device to base station communications. In fourth generation, the grant based (GB) long term evolution-advanced (LTE-A) random access procedure was adopted for initial access. However, when a large number of devices compete for scarce radio resources, the LTE-A random access procedure causes higher device collisions resulting in thereby long latency. The proliferation of devices triggered by massive Internet of things (mIoT)/mMTC deployments and the stringent access requirements of futuristic applications further exaggerates this problem. Moreover, mMTC devices typically generate small packets making the initial control message exchange in LTE-A random access a major burden. As an effort to address these shortcomings in existing initial access procedures, this dissertation proposes both grant based and grant-free access schemes. The proposed GB schemes enable both device and resource grouping to give priority access for mMTC devices with URLLC requirements considering a large device population. Meanwhile, the proposed GF access scheme enables priority access for high priority devices based on a heterogeneous traffic arrival scenario. A two dimensional Markov model that aggregates both high and low priority traffic through a pseudo-aggregated process is developed.

Furthermore, prediction of traffic arrivals at a base station is of paramount importance for successful detection of congestion and to initiate proactive measures to minimize collisions. In reality, a base station only has information about the number of detections which does not naturally reflect the number of arrivals that caused it. Therefore, as the second topic, this dissertation presents a machine learning based traffic arrival prediction model for bursty traffic arrivals in an mMTC scenario. It relies on the detected data at the base station to learn about actual number of device arrivals and predict the occurrence of a bursty traffic arrival. Moreover, resource allocation strategies based on the predictions of the proposed model are also introduced.

On the other hand, 5G networks are envisaged to provide anytime and anywhere communications. Evaluating the availability and reliability of a network from an end

user perspective appears as a vital component for benchmarking service providers. As the third topic, this dissertation presents an end user availability evaluation metric and demonstrates its applicability in a heterogeneous network for a mobile user. The proposed approach applies a dependability theory perspective for evaluating user availability by considering factors that deteriorate reliability in both space and time domains.

The performance of the proposed schemes is evaluated through analyses and extensive discrete-event simulations. In a nutshell, we believe that the schemes and analyses presented in this dissertation will advance the state-of-the-art research within the studied topics and contribute positively towards solutions for enabling URLLC and mMTC in 5G NR and beyond 5G networks.

Preface

This dissertation is a result of the research work carried out at the Department of Information and Communication Technology (ICT), University of Agder (UiA), in Grimstad, Norway, from July 2017 to March 2020, on a full time basis and continued thereafter untill December 2020 on a part time basis from Sri Lanka. Professor Frank Y. Li, UiA, has been the main supervisor of this Ph.D. work and Dr. Indika A. M. Balapuwaduge, UiA, has been the co-supervisor. The research work presented in this dissertation has been funded by the Research Council of Norway through the Center for Research-Based Innovation (SFI) Offshore Mechatronics, project number 237896, and the Department of ICT at UiA.

Production note: LaTeX has been adopted as the tool for writing this dissertation, as well as the papers produced during the author's Ph.D. study. The mathematical calculations and simulation results are obtained by using MATLAB.

Acknowledgments

At the time of my Ph.D. dissertation submission, I would like to take this opportunity to express my gratitude to my supervisors, colleagues, friends, and family who gave me encouragement and support through all these years.

First and foremost, I am grateful to my main supervisor, Professor Frank Y. Li, for his support, guidance, and care through out my Ph.D. journey. I had the pleasure of working with him earlier during my Master's degree and that experience motivated me to pursue a Ph.D. study. I have benefited immensely from his expertise, guidance, encouraging thoughts, and great enthusiasm for pursuing high quality research. He was constantly pushing and encouraging me towards high quality publications and created a very pleasant working environment. He was also like a guardian to me even beyond the role of a supervisor. Hereby I express my heartfelt gratitude to Professor Frank for all the support and guidance without which I may have not come this far.

Furthermore, I express my sincere gratitude to my co-supervisor Dr. Indika Balapuwaduge. He has been a teacher, mentor, and a friend for me for almost 12 years and his presence as the co-supervisor was extremely helpful to my Ph.D. endeavor. He was always available for supervision and offered valuable advices to my Ph.D. work and writings. His insightful comments have significantly improved the quality of my research work.

Moreover, I also like to thank Prof. Vicente Casares-Giner with the Departamento de Comunicaciones, Universitat Politècnica de València (UPV), for all the generous support and guidance he provided during our collaborations. Prof. Vicente generously shared his knowledge and expertise with me and he was always helpful and available, not only during his research stay at UiA but also even with his busy schedule at the UPV. I was able to develop my knowledge and research skills by collaborating with him.

In addition, I like to express my gratitude to the SFI Center of Offshore Mechatronics for funding my Ph.D. study. A special acknowledgement to Prof. Geir Hovland, for his extensive support to me through all these years. A sincere word of thanks goes also to the Head of the Department of ICT, Folke Haugland, for all his support, especially during the turbulent times. Both of them were very considerate and took trouble in accommodating several requests of mine to ensure a benign working circumstance.

I appreciate especially the assistance provided by the Coordinators of the Ph.D. program at the Department of ICT, UiA, Emma Elisabeth Horneman and Kristine Evensen Reinfjord. Emma provided her full support and encouragement to me during challenging times and this was continued by Kristine as well. Together they took care of all the Ph.D. students by providing viable solutions for any administrative issues we faced.

Furthremore, I express my heartfelt gratitude to the Electronics and IoT group at UiA,

Especially, Associate Prof. Lei Jiao and Dr. Debasish Ghose for their friendly support and feedback on my work.

Moreover, I also like to thank the administration of the University of Ruhuna (UoR) Sri Lanka, for granting me leave and supporting me towards completion of my Master and Ph.D. studies. The partnership between UoR and UiA enabled me to come to Norway through the Quota Scheme for the Master's degree and paved the way for my Ph.D. study as well.

My gratitude and thanks go to all friends and colleagues at UiA for providing me support and companionship during my stay in Norway. Especially to all the Ph.D. fellow students and Postdoctoral fellows, who shared their experiences and common troubles during our Ph.D. lives. I made lots of new friends and great memories thanks to them and I am forever grateful for that. Furthermore, my heartfelt gratitude to the Sri Lankan community at UiA and Agder for providing me support, understanding, and friendship through all these years when ever required.

Last but not least, I like to express my deepest gratitude, love, and affection to my family. My parents, Rupawansa Weerasinghe and Chitra Somalatha, and my sister Dakshika Weerasinghe always encouraged me and provided me with love and support through all these years. Special mention to my wife Pabasara Weeraman who postponed her own carrier goals to provide love and support to me. She shared all the ups and downs in my Ph.D. life while motivating and encouraging me throughout this endeavor. And to our little daughter Anuki for giving me all the love and joy during the difficult times of the Ph.D. journey.

Thilina Nuwan Weerasinghe
Galle, Sri Lanka
18th December 2020

Publications

In the following, the publications related to the research conducted during the Ph.D. period are presented. Papers A-D are reproduced as Part II of this dissertation. Papers 5-8 which are listed as the second part of this list are not included in this dissertation. While the four included papers are sorted according to the topics studied in this dissertation, the other four papers are sorted in a chronological order according to the dates of the publication.

Papers Included in the Dissertation

Paper A: T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Priority-based initial access for URLLC traffic in massive IoT networks: Schemes and performance analysis," *Computer Networks*, vol. 178, Article 107360, Sep. 2020.

Paper B: T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Transactions on Communications*, early access article, Jan. 2021.

Paper C: T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Supervised learning-based arrival prediction and dynamic preamble allocation for bursty traffic," in *Proc. IEEE International Conference on Computer Communications (INFOCOM) Workshops*, Apr. 2019.

Paper D: T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Time-Space domain availability analysis under reliability impairments," *IEEE Networking Letters*, vol. 1, no. 3, pp. 103-106, Sep. 2019.

Other Publications Not Included in the Dissertation

Paper 5: T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Per-user availability for ultra-reliable communication in 5G: Concept and analysis," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018.

Paper 6: T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Preamble reservation based access for grouped mMTC devices with URLLC requirements," in *Proc. IEEE International Conference on Communications (ICC)*, May 2019.

Paper 7: T. N. Weerasinghe, I. A. M. Balapuwaduge, F. Y. Li, and V. Casares-Giner, "MDP-based resource allocation for uplink grant-free transmissions in 5G new radio," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, May 2020.

Paper 8: A. Søraa, T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Preamble transmission prediction for mMTC bursty traffic: A machine learning based approach," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec. 2020.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two Dimensional
3GPP	Third Generation Partnership Project
4G	Fourth Generation mobile communication systems
5G	Fifth Generation mobile communication systems
ACB	Access Class Bearing
ACK	Acknowledgement
APPA	Arrival Prediction based Preamble Allocation
BS	Base Station
CDF	Cumulative Distribution Function
CP	Cyclic Prefix
CR	Coverage Region
C-RNTI	Cell Radio Network Temporary Identifier
D2D	Device to Device communications
DGDP	Device Grouping with Dedicated Preambles
DRX	Discontinues Reception
DSA-GF	Dynamic Slot Allocation - Grant-Free
DT	Down Time
DTMC	Discrete Time Markov Chain
EAB	Enhanced Access Bearing
eMBB	Enhanced Mobile Broadband
eNB	Evolved NodeB
FR1	Frequency Range 1
FR2	Frequency Range 2
FSA	Frame Slotted ALOHA
GB	Grant Based
GF	Grant-Free
gNB	Next Generation NodeB
HARQ	Hybrid Automatic Repeat Request
HPT	High Priority Traffic
HS	Hybrid Scheme
HTC	Human Type Communications
IFT	Intermediate First Transmission
IoT	Internet of Things
IRSA	Irregular Repetition Slotted ALOHA
LPT	Low Priority Traffic

LSTM	Long Short Term Memory
LTE	Long Term Evolution
LTE-A	Long Term Evolution-Advanced
MDT	Mean Down Time
MIB	Master Information Block
MIMO	Multiple Input Multiple Output
mIoT	Massive Internet of Things
mMTC	Massive Machine Type Communications
mmWave	Millimeter Wave
MTBF	Mean Time Between Failures
MTC	Machine Type Communications
MTFF	Mean Time to First Failure
MU	Mobile User
MUT	Mean Up Time
NACK	Negative Acknowledgement
NB-IoT	Narrowband Internet of Things
NOMA	Non Orthogonal Multiple Access
NR	New Radio
NUD	Non-URLLC Device
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PAWG	Preamble Allocation Without Grouping
PDCCH	Physical Downlink Control Channel
P-IRSA	Priority Irregular Repetition Slotted ALOHA
PPP	Poisson Point Process
PRACH	Physical Random Access Channel
PR-ALOHA	Priority Reservation ALOHA
QoS	Quality of Service
RA	Random Access
RACH	Random Access Channel
R-ALOHA	Reservation ALOHA
RAR	Random Access Response
RAUG	RA Slot based URLLC Grouping
RI	Reliability Impairment
RMSE	Root Mean Square Errors
RRC	Radio Resource Control
SDU	Service Data Unit
SG	Savitzky-Golay filtering
SGPA	Static Group based Preamble Allocation
SIB	System Information Block
SIB2	System Information Block 2
SIC	Successive Interference Cancellation
TFF	Time to First Failure
TTI	Transmission Time Interval

UAC	Unified Access Control
UD	URLLC Device
UE	User Equipment
UR	Ultra-Reliable Communications-Region
URC	Ultra-Reliable Communications
URLLC	Ultra-Reliable and Low Latency Communications
UT	Up Time

PART I

Chapter 1

Introduction

The rapid evolution of mobile communication technologies facilitates various emerging applications with high-demanding requirements, paving the way towards a networked society of human-beings and devices. With each generation of mobile communication systems, novel technological directions emerge and propelling them requires innovative approaches in design and implementation of hardware, software, and protocols.

In this chapter, we first present a few fundamental elements that are related to this dissertation. Thereafter, the main research questions, how these questions are addressed in the dissertation, and the main contributions are summarized. The chapter also gives an outline of the thesis organization.

1.1 5G Technology Pillars

The 5^{th} generation (5G) mobile communication networks introduce three main technology pillars or use cases comprising of enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC), and massive machine type communications (mMTC) [1]. The eMBB use case is a natural evolution from the existing 4^{th} generation (4G) networks with the objective of providing much faster data rates and better user experience. Through eMBB, 5G networks are envisaged to deliver data transfer rates up to 20 Gbps providing high capacity while also supporting high speed user mobility.

URLLC is an emerging direction which was not a main focus in previous generations of mobile communications. URLLC applications include many futuristic scenarios like remote surgery, factory automation, and vehicle to everything communications. Accordingly, to meet the stringent reliability and latency requirements expected under URLLC, revolutionary approaches are required. Depending on different applications requirements, the 3^{rd} generation partnership project (3GPP) has defined various reliability and latency threshold values for URLLC traffic. For some applications, this requirement could be $1-10^{-6}$ reliability and a user plane latency of 1 ms [2]. However, ultra-reliability and low latency are generally two contradictory requirements. To achieve higher reliability, it takes longer latency. Therefore, achieving both high reliability and low latency simultaneously is a highly challenging task which expects a tradeoff between these two requirements [3].

On the other hand, the concept of machine type communications (MTC) has been

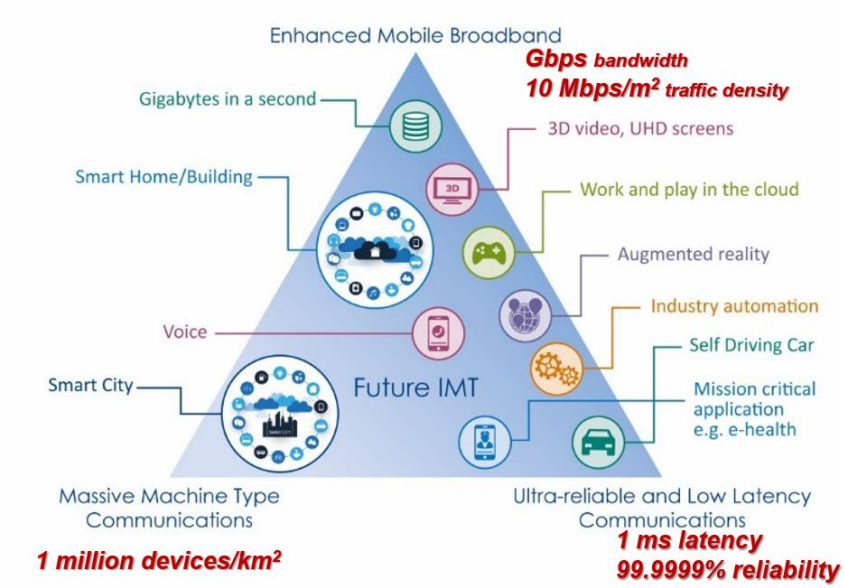


Figure 1.1: 5G main technological directions [1].

developed over the past decade. With the the rapid deployment of the Internet of things (IoT), a huge number of devices are being connected to the IoT through ubiquitous wireless connections especially via cellular networks. It is envisioned that 5G will accommodate more than 1 million devices per km^2 and thereby creating massive IoT (mIoT) networks. The requirement to deal with such a massive number of MTC devices has fostered the concept of mMTC. However, allocating scarce resources especially radio resources to accommodate the needs and limitations for such a higher number of devices encounters a major challenge for the success of mMTC in 5G networks [4].

1.2 Elements Related to This Thesis Work

In this section, we discuss a few fundamental elements upon which the dissertation work is built. The topics that are discussed herein are further explored in the chapters related to each paper in Part II.

1.2.1 5G Numerologies and Frame Structure

5G new radio (NR) is built up based on the same principles of orthogonal frequency division multiple access (OFDMA) technology which was adopted in long term evolution (LTE) and in LTE-advanced (LTE-A). NR is designed to support deployment across a wide range of frequencies. In NR, two operating frequency ranges are defined by the 3GPP [5] with frequency range 1 (FR1) corresponding to 450 MHz – 6000 MHz and frequency range 2 (FR2) corresponding to 24250 MHz – 52600 MHz. Distinct from LTE-A, NR is enabled to support multiple subcarrier spacings. Subcarrier spacing is equal to the reciprocal of the symbol time and can be represented as the width of one subcarrier in the frequncy domain. In LTE, only a 15 kHz subcarrier spacing was adopted. Similar to the LTE frame structure, 5G NR also has the same time duration for a radio frame

Table 1.1: 5G numerologies and related parameters [20]

Numerology index μ	Number of symbols per slot	Number of slots per frame	Number of slots per subframe
0	14	10	1
1	14	20	2
2	14	40	4
3	14	80	8
4	14	160	16

(10 ms) and a subframe (1 ms).

However, in NR each subframe may contain one or multiple number of slots based on the numerology in use. Correspondingly, a numerology represents a different type of subcarrier spacings and symbol length. Tab. 1.1 gives an overview of 5G NR numerologies and the corresponding number of slots per frame and subframe. Specifically, 5G NR consists of five distinct OFDMA numerologies that support radio operations in both FR1 and FR2. Correspondingly, the subcarrier spacing index μ may have five possible values each mapping to a specific subcarrier spacing value according to the formula $\Delta f = 2^{\mu} \times 15$ kHz. The corresponding subcarrier spacings for different values of $\mu = 0$ to 4 are obtained as 15 kHz, 30 kHz, 60 kHz, 120 kHz, and 240 kHz, respectively.

Having multiple numerologies offers additional flexibility for scheduling of different services. With shorter duration of slots, transmissions can be scheduled much faster than a traditional LTE based network. Furthermore, NR enables both uplink and downlink transmissions within a slot, making it possible to support low latency traffic. In addition, different numerologies support multiple deployment scenarios from sub-1 GHz range to millimeter wave (mmWave) applications. The higher numerologies $\mu = 3$ and $\mu = 4$ support high frequencies in the mmWave range defined in the range of FR2. Furthermore, since symbol length and subcarrier spacing are inversely proportional to each other, wider subcarrier spacings reduce the cyclic prefix (CP) length which is an overhead to a system. This is especially useful for smaller cells where delay spread is low. For applications which tolerate longer delay spread, narrower subcarrier spacings are preferable.

1.2.2 Reliability and Availability in Mobile Networks

Reliability and availability are two most important factors when assessing the operation quality of a given mobile network. When a system is available, users require reliable services with a sufficient level of quality. Therefore, both reliability and availability are fundamental performance attributes for dependability analysis.

The reliability of a system is defined as the probability that a system will perform its intended functions without failure for a given interval of time under specified operating conditions [7]. The availability of a system can be analyzed mainly in two ways, i.e., instantaneous (point) availability and steady state availability. Instantaneous availability is the probability that the system is operating properly and is available to perform its functions at a specified time. The steady state availability can be obtained as the average

availability over a sufficiently long period of time.

1.2.3 Initial Access Mechanisms

Initial access is an essential procedure for connectivity establishment in wireless networks. Due to the scarcity of radio resources and the possibility of collisions in wireless access, it is important to employ efficient medium access protocols in order to ensure high performance in terms of reliability and latency. Initial access approaches can be classified into two categories. The first one is grant based (GB) access. In GB, devices first need to obtain an access grant before transmitting their data packet to a base station. To this end, numerous random access (RA) protocols exist and all of them involve an initial message exchange procedure between a transmitting device and its base station in order to gain access grant. The data packet will only be transmitted upon the successful completion of the initial control message exchange and thereby after an obtained grant.

The other approach is grant-free (GF) access where a device could directly transmit its data packet without obtaining a grant from the base station [8]. Although the data packet is directly transmitted without prior communications, we still consider GF as an initial access scheme since it represents the initial point of communication between the device and the base station. GF is preferable for small data transmissions which are common for mMTC applications. The elimination of initial message exchange reduces the resulting latency of communications for GF communications. However, its performance could quickly deteriorate depending on the offered traffic load. In general, a higher traffic arrival rate leads to higher collisions and more packet losses.

1.3 Research Questions

In this section, we explore potential research directions for enabling URLLC and mMTC in NR networks and then identify a few research questions that are related to this thesis work.

1.3.1 Research Directions for URLLC and mMTC

In order to provide URLLC and mMTC, research in diverse fields is required [10]. Initial access mechanisms play a major role for meeting URLLC service requirements. This is specially applicable when we consider URLLC in an mMTC scenario. When a huge number of devices are served inside a cell as envisioned in many 5G scenarios, it is well understood that this situation could lead to complications for initial access. It is therefore vital to analyze how existing GB approaches like the LTE-A RA procedure could be adopted to cater such requirements. Especially, considering bursty traffic arrivals when a huge number of traffic arrivals within a short period of time, the performance of GB initial access could be further deteriorated.

On the other hand, in NR networks, more focus is attained by GF based protocols due to its low latency capabilities and lesser control overhead at lower traffic arrival rates [9]. So far, little work has been done on providing priority based access in a GF setting for

NR networks. This is however appealing as devices in 5G are heterogeneous in nature with different service requirements.

In addition, machine learning techniques may be utilized to improve the performance of URLLC traffic by dynamic traffic predictions and resource allocations. Such approaches also require further research.

Furthermore, it is important to define proper metrics to measure URLLC satisfaction provided by a given network. Generally, the reliability level is measured based on the number of successful packets versus the total number of packet transmissions during a considered period. Moreover, it is imperative to consider metrics that inherit the availability aspect of communication networks and analyze the performance of various reliability measures. From a dependability perspective, the availability of a system in time domain can be defined as the system up time over the total observation time. Moreover, for a mobile network operator, it is also important to consider the space domain availability when defining the overall availability. Therefore it is interesting to evaluate URLLC considering both time and space domains.

1.3.2 Research Questions

In this dissertation, we make efforts to answer the following research questions.

- *Question 1:* Following the principle of traditional GB initial access, how to develop access schemes that can meet the URLLC requirements in an mMTC context with the presence of huge device density? How to utilize the scarce preamble resources efficiently while enabling URLLC?
- *Question 2:* For GF data transmissions, how to efficiently allocate radio resources with respect to different service requirements given the existence of heterogeneous traffic?
- *Question 3:* How to predict traffic arrivals at a gNB and use such information to reduce RA congestion during a bursty traffic arrival?
- *Question 4:* How to evaluate URLLC of an end user from a dependability theory perspective? How to analyze the reliability and availability of an end user and achieve better performance when both spatial and temporal domain parameters are considered?

1.4 Contributions and Methodology

1.4.1 Contributions

To address the above mentioned research questions, the following contributions have been made during this Ph.D. work.

- Three RA schemes that are built upon the LTE-A/NR initial process are developed. The proposed schemes envisage an mMTC scenario where a massive number of devices access a base station simultaneously. For such an arrival, the proposed schemes

attempt to reduce RA congestion by device and/or resource based grouping considering the URLLC level requirements of devices. Their performance is evaluated based on several metrics and compared with traditional LTE-A and GF schemes.

- A GF initial access scheme for hybrid traffic which dynamically allocates radio slot resources between high priority and low priority devices is proposed. A two dimensional (2D) Markov chain integrated with a pseudo-aggregated process is developed to analytically model the proposed scheme. Furthermore, the performance of the model is evaluated via simulations.

- A machine learning based scheme for predicting arrival traffic in a bursty arrival scenario is proposed. Based on the predicted traffic arrivals, dynamic resource allocation can be accommodated.

- A per-user URLLC evaluation metric is defined from a dependability theory perspective. The proposed metric combines both spatial and temporal factors that could affect the availability of a service.

1.4.2 Methodology

The following methodological workflow is pursued.

- Literature review was performed for the purpose of exploring clues to address the identified research questions. State-of-the-art efforts from both standardization body (especially the 3GPP) and research community were investigated.

- Protocol/scheme design: Different schemes have been developed as our efforts to resolve the identified problems under various scenarios.

- Model development: Mathematical modeling using Markov chains and implementing the developed analytical models.

- Implementation of the the proposed schemes and performing extensive simulations to validate the proposed schemes and to evaluate their performance.

- Comparison with existing schemes.

The results of this study have been published in three journals and five conferences. Among them the four papers which are published in the *IEEE Transactions on Communications*, *Computer Networks, IEEE Networking Letters, Proceedings of IEEE International Conference on Computer Communications (INFOCOM 19) workshops* respectively are included in Part II of the thesis.

1.5 Thesis Outline

The dissertation is organized into two parts. Part I contains an overview of the work carried out throughout this Ph.D. study and Part II includes a collection of four published

or submitted papers, which are mentioned in the list of publications. In addition to the introduction chapter presented above, the following chapters are included.

- Chapter 2 initially presents the background about initial access in wireless networks. Thereafter, an outline of the GB schemes which are originally proposed in Paper A is presented.
- Chapter 3 first presents an overview about grant-free communications followed by an outline of the priority enabled GF scheme for heterogeneous traffic arrivals presented in Paper B.
- Chapter 4 emphasizes the importance of arrival prediction at the base station with relevant background. Thereafter, it introduces an overview of the proposed machine learning based bursty traffic arrival prediction scheme in Paper C.
- Chapter 5 presents the concepts on evaluating user availability in wireless networks. It further introduces the study in Paper D about end user availability from a dependability perspective.
- Chapter 6 concludes the dissertation, reemphasizes the contributions, and highlights a few potential future research directions relevant to the topic addressed in this dissertation.

Chapter 2

Grant based Initial Access for mMTC Traffic Arrivals

In this chapter, we first discuss the basic principles of GB initial access based on the legacy LTE-A random access procedure. Then the main issues associated with this legacy LTE-A RA process are discussed and existing solutions in the literature are summarized. Thereafter, the main ideas of the proposed group based access schemes are presented.

2.1 Grant based Random Access

GB random access is predominantly employed for initial access in OFDMA based access networks, e.g., for the LTE-A RA process. For GB access, a device first obtains permission from a base station (evolved NodeB (eNB) for 4G and next generation NodeB (gNB) for 5G) before transmitting its data packets. The process of acquiring the required permission consists of the exchange of four control messages prior to a data transmission.

2.1.1 LTE-A Random Access Procedure

The most well known GB access scheme used in LTE-A RA [7] is a four step handshake procedure as depicted in Fig. 2.1. A device which intends to transmit data first randomly selects a preamble from a set of preambles available in a given slot. Preambles in 4G and 5G are generated using the Zadoff-Chu sequences. These preambles are constant amplitude zero auto-correlation sequences with periodic auto-correlation function, facilitating accurate preamble detection and timing. Furthermore, the Zadoff-Chu sequences exhibit excellent cross-correlation properties. However, the Zadoff-Chu sequences are difficult to generate in real time and require a large amount of memory for code storage due to the orthogonality requirement [12]. Thus, the number of available orthogonal preambles is limited and it is advertised by the eNB/gNB through the physical random access channel (PRACH) configuration common message [13] [9]. Generally, there are 64 preambles per RA slot in a given cell and 10 of them are reserved for contention-free access while the other 54 preambles are shared among competing devices.

The four step access procedure illustrated in Fig. 2.1 is an RA procedure. Prior to data transmission, each device will randomly select one of the available preambles and

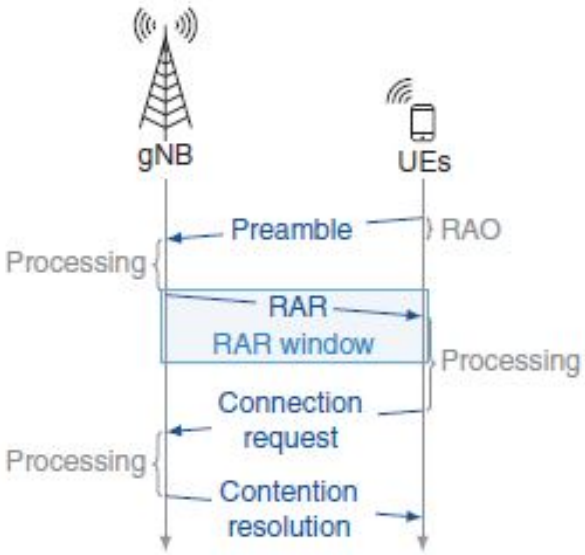


Figure 2.1: Illustration of the four step contention based access procedure for LTE-A [9].

transmits the preamble in the next available random access slot defined according to the PRACH configuration information.

The gNB will broadcast information about the set of available preambles, preamble formats, and the PRACH configuration index information to all devices in the cell periodically through a system information block 2 (SIB2) message. If two or more devices select the same preamble and transmit in the same RA slot, a collision occurs. In such a situation, often all involved devices need to retransmit preambles after a randomly selected backoff time chosen from the provided backoff window. If no collision is detected, the gNB will then reply with a random access response (RAR) message. The third and fourth messages consist of contention resolution and once all these steps are successfully completed, the initial access process is considered to be successful.

SIB and master information block (MIB) messages are broadcasted by the gNB periodically. The MIB message follows a fixed schedule with periodicity of 40 ms [13]. The SIB1 message follows a fixed schedule with periodicity of 80 ms. SIB2 provides the necessary details about the RA process to all devices in a cell. Such information includes the PRACH configuration index which defines the frequency in which an RA slot is assigned in the time domain and the preamble type. According to [12], there are 32 such PRACH configuration indexes. In Paper A, we consider two PRACH configurations where an RA slot is available in every fifth subframe or every subframe of a radio frame respectively as illustrated in Fig. 2.2. Furthermore, we have taken advantage of the flexibility provided by the NR frame structure to accommodate two RA slots within a subframe in order to satisfy mMTC requirements as explained in Sec. 2.4.

2.2 RACH Congestion and Existing Solutions

Random access channel (RACH) congestion is one of the major problems associated with the LTE-A RA procedure and it is mainly caused by the limited number of orthogonal preambles available per slot per cell [14] [15]. The existence of a large number of devices

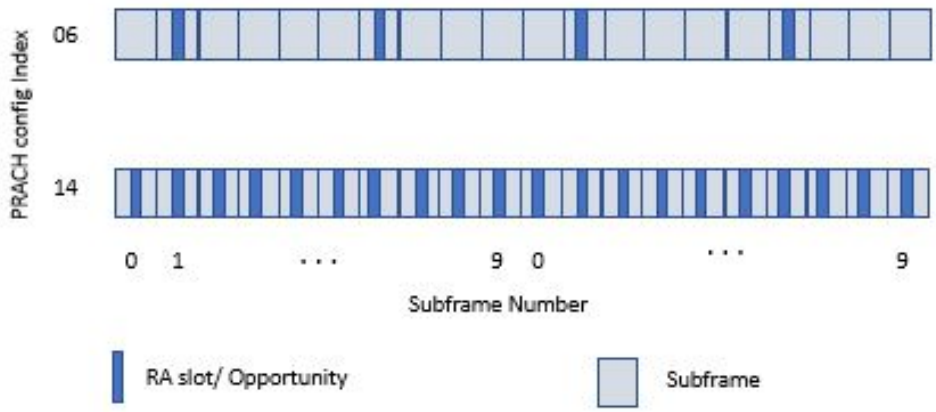


Figure 2.2: RA slot/opportunity distribution for different PRACH config indexes in LTE-A.

in mMTC scenarios can cause a high number of collisions among the competing devices at a given RA slot. This will result in retransmissions and additional delay. In bursty traffic scenarios where a large number of devices suddenly transmit over a short time duration, it causes extra congestion leading to even worse situations in terms of delay and successful access [16]. To overcome such problems, the standardization body and research community have come up with different proposals [17], [18].

One representative category of approaches is access control based methods. While access class barring (ACB) and enhanced access barring (EAB) are the most popular mechanisms in LTE-A [19] [15], 5G introduces unified access control (UAC) [21]. In ACB approaches, devices are classified based on their quality of service (QoS) levels. Each device will receive an access probability rate indicator from its eNB based on their traffic class priority level. For GB access, each device will randomly select a value between 0 to 1. If the selected value is greater than the given access probability, the transmission is barred. This approach is further extended in EAB which introduces a larger number of classes. Accordingly, when a high priority device is transmitting, all other device types are barred from transmissions. 3GPP Release 15 specifications introduce the concept of UAC which is based upon the use of access identities and access categories and it applies to 5G NR [21]. In UAC, each access attempt from a device is categorized with both one or more access identities and access category. A UE at each access attempt will assess whether an actual access attempt can be made based on the access control information applicable to the selected access identities and access category. Factors like operator policies, deployment scenarios, subscriber profiles, and available services are associated with access identities and access categories. UAC also provides flexibility to include additional access identities and access categories which allow operators to define access categories based on their own criteria [21].

Apart from access control based approaches, there are several other approaches investigated by the 3GPP and in related literature [16] [17]. Resource separation is another approach where human type communications (HTC) and MTC devices are treated independently with separate resources and priority levels. In other efforts, devices are grouped or clustered and the cluster-head will act as a relay node between an eNB and its covered end devices. However, since all data goes through the cluster-node, the overall latency of

communication could increase. In eNB initiated pull based approaches, the eNB schedules when each device can transmit. Although this would work well for periodic traffic, such an approach is not suitable for event driven traffic with bursty nature. While these approaches improve the performance of devices, they still need improvements to tackle with the requirements of URLLC in mMTC scenarios, especially when bursty traffic is concerned.

2.3 Impact of mMTC Traffic Arrivals on Achieving High Reliability and Low Latency

As discussed earlier, a massive number of devices can be active in mMTC scenarios causing unique problems in achieving URLLC [10]. One of the main issues lies in the initial access procedure. Consider a scenario where a large number of devices attempt to access within a short period of time. There is a high probability for collisions. Such arrivals are considered as bursty traffic arrivals. Bursty traffic arrivals and its impact on LTE-A RA process have been analyzed in [16, 22]. However, the analysis therein was carried out without considering a massive number of devices. Furthermore, the analysis was based on the LTE-A frame structure which does not have the flexibility offered by 5G NR. Consider the occurrence of an emergency event that causes many sensors to transmit their measurements simultaneously. Collision could become much worse in mMTC scenarios with a massive number of devices deployed inside a cell. Hence, it is imperative to explore approaches to provide ultra high reliability while minimizing latency in bursty traffic scenarios as well.

2.4 Outline of the Proposed Approach

The traditional four step LTE-A RA scheme has numerous limitations when confronted with URLLC requirements in the presence of mMTC. In Paper A of Part II of this thesis, we consider a worst-case scenario that can be experienced at a gNB, where a huge number of devices become active in a bursty manner. In such instances, the presence of a massive number of devices attempting network access would cause high congestion, low access success probability, and higher latency. Therefore, it is beneficial to provide higher access privilege for those devices that require URLLC. Considering these aspects, we propose in Paper A two main schemes for initial access approaches of mMTC devices that have URLLC requirements and a hybrid scheme containing the merits of both schemes.

2.4.1 Device Grouping based Dedicated Preambles

The first scheme proposed in Paper A is referred to as device grouping with dedicated preambles (DGDP). This scheme focuses on a scenario where a set of URLLC devices which form a device group are located in the immediate vicinity of a point of common interest monitoring the same natural or physical phenomenon. This scenario is introduced as the location-bounded URLLC services. There could be several groups inside a cell and

the number of groups is limited by the number of preambles allocated for such grouping activity. The group formation and device registration with the gNB are performed at the establishment stage and one device is selected as the leader of the group beforehand based on different factors like the proximity to other sensors, previous uplink and downlink quality. Each group member's observation needs to be reported to the gNB for instance in the following sense. Consider a set of sensors located in close proximity within an industrial plant monitoring critical parameters like temperature, motor vibration, and pressure. When a triggering event such as a fire occurs, all devices in the group will attempt to transmit its sensor measurements to the gNB. The data from group devices should be collected (instead of relying on the group leader to report everything which introduces additional delays).

In the DGDP scheme, each group consists of a group leader and multiple group members and it is the group leader's responsibility to transmit the preamble on behalf of the group. When a triggering event occurs, the group leader will transmit the dedicated preamble. From the preamble, the gNB identifies the group and its members (from the beforehand registration process) and then allocates uplink radio resources for all members within that group. The initial control exchange process is thereby shrunk to two steps. Once the grant is received, the data transmission process can happen contention free since the preamble for each group is dedicated. Here contention free is only meant for preamble transmissions. For data transmissions, we assume that there are a sufficient number of resource blocks. Therefore, this scheme increases the reliability and latency of grouped devices compared to what is obtained in traditional LTE-A RA schemes.

2.4.2 RA Slot based URLLC Grouping

In the second scheme, an RA slot based URLLC Grouping (RAUG) is proposed. This mechanism is facilitated by the novel 5G NR frame structure and numerology concepts. The envisaged network scenario is considered as location-spread URLLC devices where unlike in the first scheme the devices are not bounded by the location. A device that is attempting a URLLC based transmission is categorized as a URLLC device (UD) while the other devices are regarded as non-URLLC devices (NUDs). Utilizing the flexibility of slot wise scheduling combined with different preamble formats and PRACH configurations available in 5G NR, the RAUG scheme proposes to have two RACH transmission slots within every subframe each for UDs and NUDs separately. With this approach, UDs obtain opportunities to transmit preambles in a dedicated RACH slot without having to compete with NUDs. Since no dedicated preambles are reserved for UDs or NUDs, all devices in this case have to follow the four step LTE-A RA access procedure.

2.4.3 Hybrid Scheme

Additionally, Paper A also proposes a hybrid scheme. Therein, a scenario which involves both location-bounded and location-spread URLLC devices is envisaged. Accordingly, both device grouping and resource grouping concepts are utilized to provide better access opportunities for URLLC devices. This combination of beneficial properties in both

DGDP and RAUG schemes enables to provide flexible service requirements for different priority level users and also improves the performance of both non-URLLC and non-grouped devices.

2.4.4 Performance Evaluation and Comparison

The proposed schemes are analyzed by considering three performance metrics as access success probability, preamble collision probability, and average delay for successfully accessed devices. These metrics are evaluated both analytically and via simulations for different configurable parameters. A bursty traffic arrival is modeled according to the 3GPP recommendation in [16]. Much higher than the device population considered in a benchmark study which considered 30000 devices [22], we configure a total number of devices up to 300000, reflecting a bursty arrival process in an mMTC scenario. The performance of different types of devices is evaluated and compared with that of traditional LTE-A based access. By employing the proposed schemes, grouped devices and URLLC devices obtain better performance. In addition, the performance of non-grouped and non-URLLC devices is also improved when compared with the results obtained from the traditional LTE-A RA procedure.

2.5 Chapter Summary

In this chapter, we initially discuss the LTE-A RA process and its limitations due to RACH congestion. Then we review briefly the existing work that attempts to improve the performance of various types of devices. Furthermore, we emphasize the imperativeness for further research in the area of initial access for combined URLLC and mMTC use cases in NR networks. In such a context, our proposed schemes represent a novel effort towards the beyond state-of-the-art research within the topic of GB access for 5G NR networks.

Chapter 3

Grant-Free Access with Traffic Priority in 5G NR Networks

With the proliferation of diverse application requirements, grant-free access has attracted a surge of research interests. Different from GB access, in GF access, devices transmit their data packets directly without the need for obtaining a prior grant from the gNB. GF access works on a mechanism similar to slotted ALOHA [8]. For providing access to heterogeneous traffic, enabling priority based access within grant-free plays a vital role for the performance of access schemes. In this chapter, we look first at background and literature about GF transmissions for LTE-A and NR networks. Then we outline the proposed scheme which is tailored to GF access based on NR numerologies.

3.1 Grant-Free Access

Grant-free approaches have recently gained increasing attention for two main reasons. First, as already discussed in earlier chapters, an initial handshake procedure for grant based random access introduces considerable latency. In addition, mMTC devices typically generate small data packets in a sporadic manner. Therefore, the control overhead introduced by GB approaches is significantly high with respect to the data packet size that needs to be transmitted. Consequently, GF traffic appears to provide a viable alternative for reducing such latency and resolving control overhead issues associated with GB under low to medium traffic arrival rates. In LTE-A networks, semi-persistent scheduling or configured-grant based approaches have been proposed as latency reduction methods [23]. In such approaches, a gNB will schedule the transmission slots for periodic traffic so that those devices do not need to obtain grants for data transmission. While this can be considered as a grant-free approach, such methods are not practical for mMTC devices where traffic arrivals are generally event driven and sporadic.

3.1.1 Existing Work on GF Access

Several GF based approaches have been introduced in the literature. In [8], GF schemes are categorized into four main approaches as illustrated in Fig. 3.1. These categories are reactive, K repetitions where K is an integer with $K > 1$, proactive, and power boost

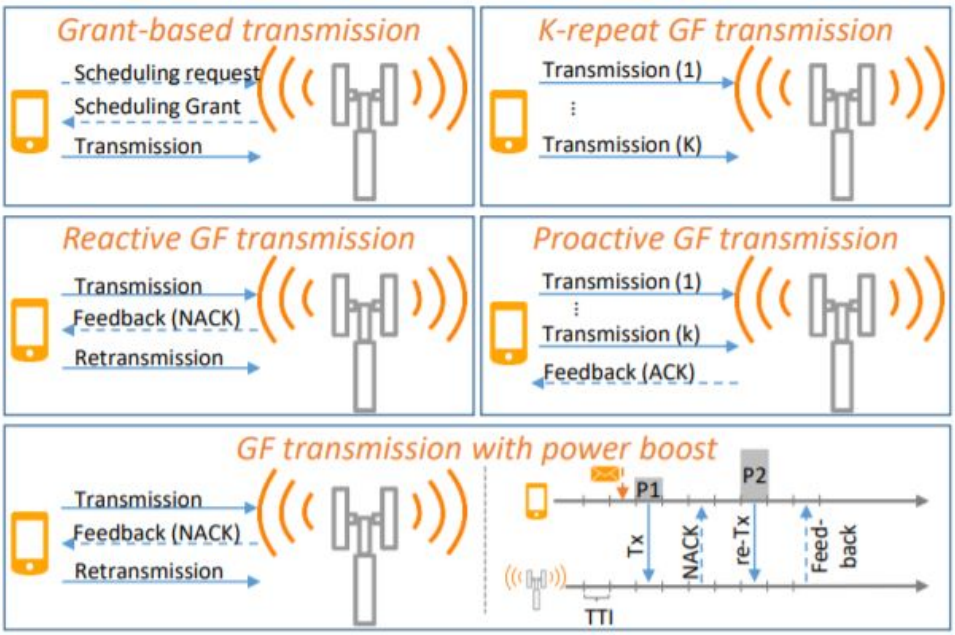


Figure 3.1: Different categories of initial access [8].

based. In reactive schemes, a device will transmit its data packet and wait for a hybrid automatic repeat request (HARQ) based ACK/NACK message. If no ACK message is received (or NACK is received), the device will retransmit the same packet again up to a certain retransmission limit. Such HARQ schemes are often referred to as reactive since retransmissions are triggered based on the knowledge about the previous transmission [24].

In the K repetition scheme, a device will transmit K replicas of the same packet before receiving a possible ACK. Since the gNB only sends an ACK after K repetitions, there could be redundant transmissions which are unnecessary especially when considering the slot resource constraint. However, this approach may provide higher access probability for a URLLC packet since it is transmitted multiple times given that the traffic arrival rate is low.

In proactive schemes, the gNB will send feedback for each packet transmission. A device will transmit its packets on selected GF slots. Unlike reactive schemes, the device will not wait for the feedback after the first transmission. Instead, it will keep transmitting its data packets until an ACK for a previously transmitted message is received. Therefore, the number of repetitions can be less than K. This scheme is computationally heavy for end devices, as they need to monitor feedback from the gNB. However, it is also likely to be more resource efficient than K repetitions with overestimated K and more reliable than K repetitions with underestimated K.

In a power boosting scheme, retransmissions mentioned in reactive schemes are performed with power boosting at each retransmission. Such an approach can be implemented as the retransmitted packets require priority in order to have successful access within a lower latency.

There are many studies in the literature focusing on these four types of approaches, their performance evaluations, and possible variants [25–29]. As grant-free transmissions inherit the ALOHA/slotted ALOHA transmission mechanism, many prior research related

to slotted ALOHA can be applied to grant-free based access mechanisms. For example, the irregular repetition slotted ALOHA (IRSA) [30] can be viewed as another interesting variant of K repetition mentioned above. In IRSA, each device will transmit an irregular number of replicas randomly in the available slots. Different from K repetition, this scheme aims to minimize collisions among retransmissions due to the irregularity.

3.1.2 Single and Multi-User Detection Consideration

Single user detection and multi-user detection methods are employed to uniquely identify individual users at a given slot upon successful packet reception. In traditional communication systems, if two or more packets transmit within the same slot, it will be considered as a collision. However, when multi-user detection is possible, more than one user which transmitted within the same slot can be detected successfully. Multi-user detection is possible thanks to advanced signal processing available at receivers. When multiple users are involved in the same time slot on the same frequency, non-orthogonal multiple access (NOMA) based interference cancellation methods such as successive interference cancellation (SIC) and other techniques like multiple input multiple output (MIMO) can be applied to separately identify individual users.

Grant-free based approaches can be employed for both single user detection and multi-user detection. While there are a plethora of research related to NOMA based GF access [31], the 3GPP has decided to discontinue NOMA as a work item for 5G NR at the current phase [32]. However, it remains still as a study item for beyond 5G systems. For the work performed in Paper B of Part II, we decide to focus is on single user detection based GF schemes. In Sec. 3.5 below, we explore some draft ideas on how to enhance this work with the help of multi-user detection.

3.2 Priority based Access

Providing priority for devices with higher QoS requirements in initial access has been considered in many studies [33] [34]. Among them, ACB and ECB [19] used in GB random access were discussed in Chapter 2. In what follows, we summarize a few popular approaches to provide priority access from the literature.

A priority based access for ALOHA based scheme was considered in [33] by introducing priority to the traditional reservation ALOHA (R-ALOHA) scheme [35]. R-ALOHA enables reserving specific time slots and the proposed priority reservation-ALOHA (PR-ALOHA) scheme reserves them exclusively for high priority traffic. This scheme enables higher priority traffic to compete in the reserved set of slots. In priority IRSA (P-IRSA) proposed in [34], access control is introduced and users with high priority obtain higher channel access probability under increased traffic load in an IRSA based system. In [36], a priority-based channel allocation mechanism for wireless cellular networks was presented. The main idea of their work is to prioritize call requests and available channels. Therein, both users and channels are categorized into high and low priority groups. Furthermore, a pseudo-Bayesian ALOHA algorithm with multiple priorities was proposed in [37]. In that work, the authors intended to reduce the waiting time of delay sensitive request packets

in a multimedia environment. The transmission probability of each device is adjusted based on both channel history and a priority parameter relevant to their priority class. Access control for cellular and device to device (D2D) communications in a single cell was studied in [38]. The access of both cellular and D2D users is controlled by assigning different priority levels for each category. In addition, game theoretic approaches were also proposed in the literature for multiple access control in wireless systems [39] [40].

Although access control in a general sense is a widely explored topic, little work has been done regarding enabling priority based access for heterogeneous traffic for 5G NR based GF communications. In the following, we present an outline of our proposed approach with regard to the scheme presented in Paper B of Part II.

3.3 Outline of the Proposed Scheme

In Paper B, we present a priority enabled grant-free access scheme for dynamic slot allocation referred to therein as DSA-GF. DSA-GF has been proposed based on the 5G NR frame structure and is applicable to different numerologies. As mentioned earlier in Chapter 1, a subframe which has a fixed length of 1 ms could consist of one or more slots. Considering small data transmissions, mMTC traffic is facilitated for grant-free access within all these slots. The main idea of DSA-GF is to differentiate heterogeneous traffic based on their QoS requirements and provide higher priority to devices with higher QoS requirements without starving low priority traffic.

3.3.1 Definition of Terminologies

The following terminologies are used throughout this chapter and in Paper B.

Active device: A device that has at least one packet ready to transmit. Active devices are composed of new arrivals and backlogged devices.

Postponed device/transmission: A device that does not select to transmit in the current subframe due to the restriction imposed by the permission probability.

Collided device: A device whose transmission in the current subframe failed due to a collision with other transmission(s) in the same slot.

Backlogged devices: The sum of devices that postpone their transmission to the next subframe and those devices that were involved in collisions in the current subframe.

New arrivals: Devices that have generated a new packet during the current subframe. These devices follow the immediate first transmission (IFT) principle for possible transmission (constrained by the permission probability) in the next subframe.

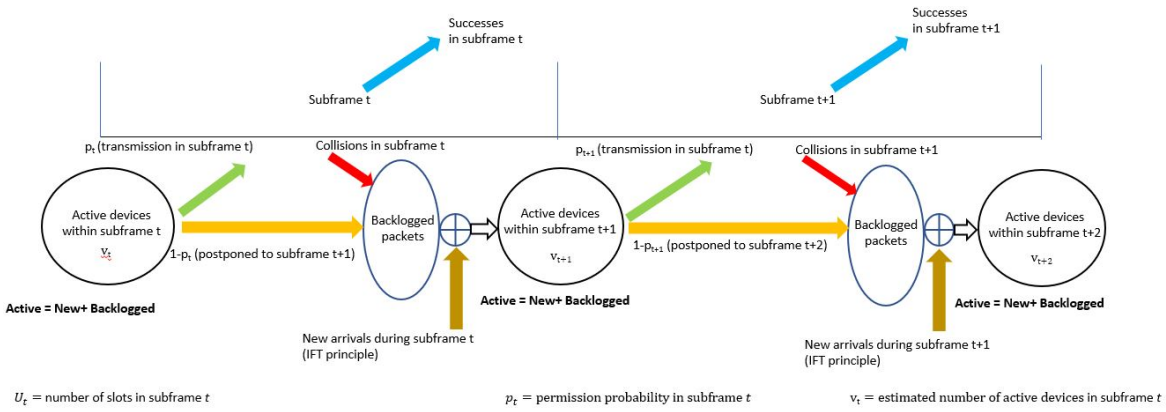


Figure 3.2: Relationships among new arrivals, postponed transmissions, backlogged devices, and active devices.

Collided slot: A collided slot is a slot which is selected by two or more devices.

Successful slot: A successful slot is a slot occupied by only a single user.

Idle slot: An idle slot is an empty slot which is not occupied by any user.

Fig. 3.2 illustrates the relationships among new arrivals, postponed transmissions, backlogged devices, and active devices. The formation of the set of active devices in each subframe is also illustrated.

3.3.2 Network Scenario and Assumptions

When different priority devices intend to perform GF access, it is preferable to have dynamic control over the resources that can be allocated for each priority type. In this work, we consider two traffic classes with different priorities referred to as high priority traffic (HPT) and low priority traffic (LPT) respectively. The priority level is given considering the reliability and latency requirements of each traffic class. We further utilize the novel concept of 5G NR numerology when deciding the resource allocation process. For example if numerology three is adopted, each subframe will consist of 8 slots. Our aim is to dynamically allocate these slots among HPT and LPT devices based on the predefined criteria. We assume that the packet size of each device is small so that a packet can be transmitted within a slot of the adopted numerology (i.e., the packet transmission duration $<$ 14 OFDMA symbols). Such an assumption is reasonable for small packet transmissions under mMTC scenarios [9] [24]. Accordingly, all available slots within a subframe of a given numerology are regarded as available for GF traffic. Note however that the proposed scheme is not restricted to a fixed packet size.

To represent the number of active devices within each slot, we consider a Bernoulli process for packet generation with a combination of different number of devices and activation probabilities. Only the devices which are active will attempt to transmit their data packets in a randomly selected slot. The transmission decision will be taken based on the permission probability devices receive from the gNB. If a device decides to transmit

based on the permission probability, a GF slot specific for the device type will be selected randomly for its packet transmission.

3.3.3 The Proposed Dynamic Slot Allocation Scheme

The detailed description of the proposed scheme with an algorithm for active device estimation is presented in Sec. III of Paper B. In the following, we summarize the main components of the proposed scheme.

- **Observation:** At the end of each subframe, the gNB observes the number of holes, collided slots, and successful slots for each traffic type.
- **Estimation:** According to its observations, the gNB estimates the number of backlogged devices at the end of the subframe using an algorithm based on the Bayes rule. Thereafter, the gNB estimates the number of new arrivals during each subframe. This process is carried out for each traffic type independently.
- **Slot allocation:** Based on the backlogged and new arrival estimations, the gNB decides the number of slots that would be allocated HPT. Once the HPT slots are allocated, the remaining is allocated to LPT. The HPT is prioritized in this manner. Note that the minimum and the maximum number of slots that can be allocated to a specific traffic class is predefined.
- **Permission probability and data transmission:** Both HPT and LTP devices will receive a permission probability which the gNB calculated based on the number of available slots and the estimated number of active devices. With this probability, each active device randomly selects one of the allocated slots within the current subframe to transmit its packet. Otherwise, the device postpones its transmission to the next subframe.

The number of slots that can be allocated for HPT and LPT traffic is controlled by the maximum and minimum numer of slots for each traffic type. These parameters are configured considering factors like traffic load over a long period of time (which is much longer than the subframe or frame duration). By configuring the maximum number of slots for HPT which is less than the total number of slots available in the subframe, we provide at least one slot for LPT traffic. As such, the starvation of LPT traffic has been avoided.

3.4 Analytical Model and Simulations

The performance the proposed DSA-GF scheme is modeled using a discrete time Markov chain (DTMC). More specifically, a 2D Markov chain that integrates both HTP and LPT traffic through a pseudo-aggregated process is developed. During each subframe, every active device generates a packet with a certain probability according to a Bernoulli process. We observe the system at the boarder of two consecutive subframes and such a

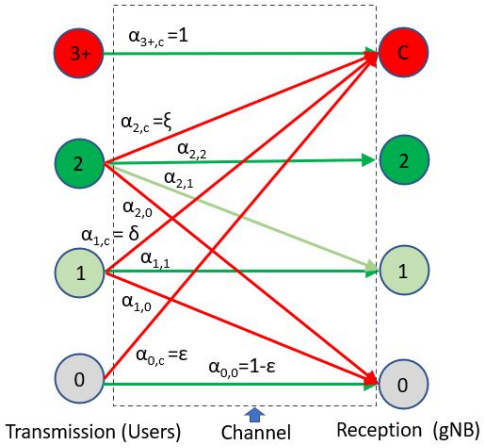


Figure 3.3: Illustration of a multi-user detection scenario with channel impairments.

time instant is regarded as a transition instant. The developed Markov model is defined by a set of three random variables. They correspond to the number of estimated device arrivals at the current subframe, the number of actual device arrivals at the current subframe, and the number of slots allocated in the current subframe. The model is built upon the state transitions of these parameters from one subframe to the next for both HPT and LPT traffic. Four metrics for performance evaluation including throughput per subframe, throughput per slot, access delay, and packet loss probability are defined.

In addition to the analytical model, the proposed scheme is also simulated. Here we consider different combinations of offered traffic for HPT and LPT, and the impact on each other's performance is evaluated. In addition, the impact of maximum and minimum number of slots that can be allocated to HPT on the performance of both traffic types is also evaluated. While the main body of the study is carried based on the NR frame structure with numerology index three, we also demonstrate the applicability of the proposed scheme to other numerologies as well. The simulation results demonstrate a close match with the results obtained through the analytical model.

Another salient feature of our model is that the number of states in our pseudo-aggregated 2D Markov chain model is much less than what is needed in conventional 2D Markov models. It is well known that when the number of states in one dimension increases, the size of the state space in a 2D model would grow dramatically. In this work, we adopt a pseudo-aggregated process which performs state aggregation to address the problem of how to deal with large state spaces in a Markov chain through a much reduced state space.

3.5 Potential Extensions for Multi-Packet Reception

A potential extension of the proposed work is to consider multi-user detection along with the effects of channel impairments like interference and fading. For such possible extension, a preliminary scenario can be explored as illustrated in Fig. 3.3. It represents a scenario where multi-user detection is enabled such that a number of simultaneous users can possibly transmit in the same slot without causing a collision.

In Fig. 3.3, $\alpha_{x,y}$ with $x \in (0, 1, 2, 3+)$ and $y \in (0, 1, 2, C)$ denotes the transition probability from number of transmissions to number of receptions. ξ, δ, and ε represent the values of these transition probabilities to the collided reception state from number of transmissions $2, 1$, and 0 respectively and these values are typically very small. For example, if the number of transmitting users equals to two, there are in total four receiving possibilities at the gNB. With probability $\alpha_{2,2}$, ideal detection occurs meaning that both packets are correctly received. With probability $\alpha_{2,1}$, only a single packet out of the two may be successfully detected. On the other hand, due to the effects of interference and fading, there is a possibility of gNB detecting the transmission as either a collision with $\alpha_{2,c} = \xi$ or as zero reception with probability $\alpha_{2,0}$. Accordingly, the proposed scheme can be extended considering these aspects of multi-user detection and channel conditions.

3.6 Chapter Summary

In this chapter, we first introduce the concepts related to grant-free communications followed by different categories of GF access existing in the literature. Then the state-of-the-art on different GF access schemes is summarized. Afterwards, an outline of the proposed scheme in Paper B of Part II is presented.

Chapter 4

Traffic Prediction and Resource Allocation for mMTC Traffic

In a heterogeneous network, devices will have different service requirements and transmission periodicity. While some devices transmit routine periodic traffic, other devices generate event driven aperiodic data traffic. For the latter type, traffic prediction and resource allocation are challenging due to the non-deterministic nature of packet arrivals. This problem intensifies in mMTC networks where a large number of devices coexist. Furthermore, some of these devices may have URLLC requirements. Thus providing such devices with required QoS remains as an interesting research topic. In this chapter, we propose a machine learning based approach for traffic arrival prediction and resource allocation which is applicable to mMTC networks especially under bursty traffic conditions.

4.1 Arrival Types and Modelling of mMTC Traffic

4.1.1 Types of Arrivals

In an mMTC scenario, different types of traffic arrivals can be identified as follows [2].

- Periodic traffic: In periodic traffic, the data transmission interval is repeated. For example, a sensor device may send periodic updates of a position or the repeated monitoring parameter. Therefore such traffic arrivals can be predicted and once the periodicity is known, device transmissions can be scheduled by the gNB via configured grants.

- Aperiodic traffic: In aperiodic traffic, packet transmissions are triggered by an event. It is also referred to as event driven traffic which is not easily predictable. For example, such events could be process or environment related where certain parameters for instance, temperature or pressure have exceeded or fallen below the predefined threshold.

- Hybrid traffic: Hybrid traffic is a combination of both aforementioned types of traffic. Generally, the incoming traffic experienced by the gNB is expected to be mixed. Bursty traffic, where a huge number of devices attempt channel access at the

Table 4.1: Traffic models for MTC devices [16]

Characteristics	Traffic Model 1	Traffic Model 2
Number of MTC devices	1000, 3000, 5000, 10000, 30000	1000, 3000, 5000, 10000, 30000
Arrival Distribution	Uniform distribution over T	Beta distribution over T
Distribution period (T)	60 seconds	10 seconds

same time or within a short time duration can be categorized under hybrid traffic as the involved devices could generate both periodic and aperiodic traffic.

4.1.2 Modeling Arrivals for mMTC Traffic

Generally, mMTC devices are considered to have sporadic traffic arrivals. However, traffic arrivals could be composed of a hybrid traffic type. To model traffic arrivals, there are several approaches commonly used in the literature. In [16], the 3GPP recommends two reference models when simulating MTC traffic arrivals. These two models include uniform arrivals and bursty traffic arrivals. In uniform arrivals, it is assumed that a constant number of devices perform channel access within the observation time period and their arrivals are modeled using the uniform distribution. On the other hand, the bursty traffic scenario represents a situation where a large number of MTC devices attempt to access the channel within a very small duration. This type is modeled using Beta distribution as also mentioned in Chapter 2. Tab. 4.1 presents the two arrival models and other respective parameters for each model.

In addition, Poisson arrivals are commonly adopted to model arrival processes in wireless communication systems. In a Poisson arrival process, the inter-arrival times between consecutive arrivals vary according to an exponential distribution. The arrival rate λ represents the expected number of device arrivals during a given interval.

4.2 Traffic Prediction: Motivation

Traffic prediction is a powerful tool in order to have a better overview about the number of devices that are competing for limited resources. In a GB initial access scheme like the LTE-A RA procedure, such limited resources are represented by the limited number of preambles that devices compete for. In a GF transmission, devices compete for a number of GF slots which is often limited. If the gNB can predict an increment in device arrivals, it can dynamically adjust system parameters to adapt to such an update. For example, if a higher number of device arrivals are expected in an LTE-A RA procedure, the eNB may configure EAB and ACB parameters to bar certain traffic classes in order to allow high priority devices to transmit. Similarly, parameters like PRACH configuration index and backoff interval can be updated accordingly. In a GF scenario, if a certain traffic class has higher priority, the gNB can allocate a larger number of GF slots to that traffic type when the predicted traffic rate is high. In some situations where GF traffic and GB traffic

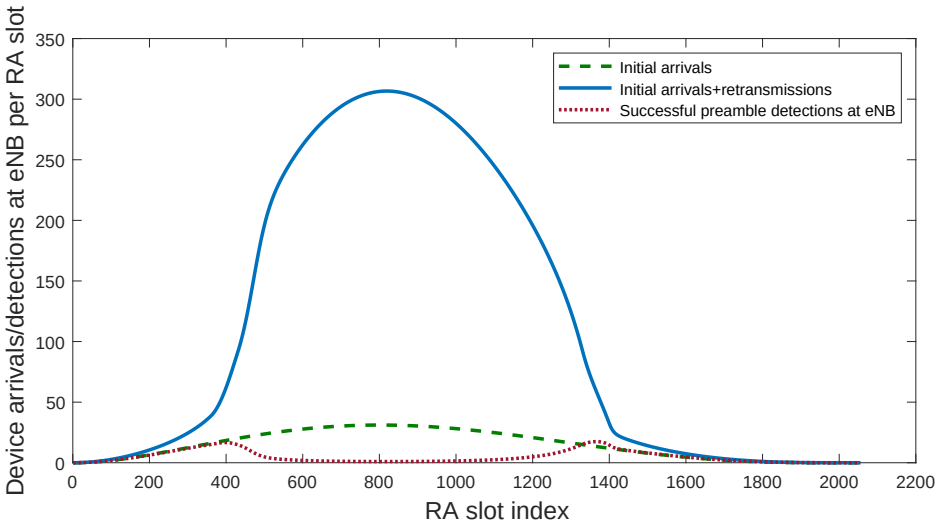


Figure 4.1: Number of arrivals and detections per RA slot in a bursty arrival of 30k devices [41].

co-exist, it is also beneficial to have the knowledge about the overall traffic intensity in order to properly allocate radio resources among GB and GF users.

In other words, traffic prediction may play a pivotal role in case of bursty traffic arrivals. In case of a triggering event like a power reset in a smart grid, or an industrial emergency, a large number of devices may transmit simultaneously or within a short duration. Considering a GB RA procedure dealing with such bursty traffic, the performance of the GB procedure deteriorates rapidly as the bursty arrival reaches its peak. This effect is illustrated in Fig. 4.1. As the number of arrivals increases, the numbers of collisions and retransmissions also increase causing more collisions and retransmissions. This cascading effect ultimately results in low access success probability and higher delay for competing devices. With the envisaged mMTC use case in 5G, there is an even higher possibility of bursty arrival occurrence. An intensive device density coupled with diverse application scenarios could result in many mMTC devices requiring high reliability and low latency performance. Under such circumstances, providing stringent QoS guarantees is impossible based on the traditional LTE-A RA procedure. As presented in Paper A, when compared to GB schemes, GF schemes experience even worse performance when exposed to a traffic burst due to its limited contention resources. Therefore, traffic prediction and identification of such a bursty nature in traffic would be highly beneficial in order to provide better QoS to high priority devices and in some cases it could eventually improve performance of all devices.

4.3 Existing Studies in Traffic Prediction

One of the major challenges in network planning and operation is how to deal with traffic uncertainty [42]. Proper traffic prediction minimizes over-provisioning in network planning [43] and also enables proactive resource management solutions to avoid performance degradation of a network [44]. Recently, deep learning based approaches have been popular alternatives for traffic prediction in mobile and wireless networks [43,45,46]. In [47], an

overview of artificial neural networks used in wireless systems was presented. More specifically, [45] employs artificial neural networks for traffic prediction in 4G gNBs wherein the predictions are utilized for efficient backhaul resource allocation. In [48], a traffic prediction mechanism based on convolutional long short-term memory (LSTM) was proposed. That model intended to use the spatial-temporal dependencies of traffic in predictions for network slicing applications in vehicular networks.

Machine learning models that jointly explore the spatio-temporal correlations were also proposed in [49]. Therein, LSTM is adopted for handling long term dependencies that are required for predictions. Considering the continuous movements of users within a given cellular network, the authors utilize the correlation among traffic flows across neighboring base stations for accurate learning and predictions. In [50], a multi-task learning architecture for mobile Internet traffic forecasting was presented. Therein, different deep learning methods were compared and the authors claim that a combination of convolution neural networks and recurrent neural networks approaches provides better overall performance. In [51], a novel deep learning architecture, named spatial-temporal cross-domain neural network, was proposed to capture the complex patterns hidden in cellular data. The proposed scheme adopts a convolutional LSTM network as a subcomponent and claims to have strong capabilities in modeling spatial-temporal dependencies of cellular data. A base station sleeping mechanism based on traffic prediction was proposed in [52]. Therein, a modified wavelet neural network is used to predict future traffic of base stations and based on that a base station sleeping mechanism is proposed for energy conservation.

In addition, statistical methods and hybrid approaches are also commonly used for traffic predictions. Auto-regressive integrated moving average models are popular for time series prediction and to capture the behavior of network traffic. In [53], an extension to auto-regressive integrated moving average schemes that supports direct modeling of the seasonal component of the traffic was proposed. A Gaussian process based model to predict traffic in a 4G network was proposed in [43]. Therein, the authors claim high prediction accuracy and better performance than what is achieved in [53]. In [43], the authors exploit information theory based entropy concept and analyze traffic predictability in cellular radio access networks. In a hybrid approach, [54] employs both LSTM and adaptive neuro-fuzzy inference time series models for cellular data prediction. A hybrid deep learning model for spatio-temporal cellular traffic prediction was presented in [55]. Therein, a novel auto-encoder based model was proposed for spatial modeling and an LSTM based approach was adopted for temporal modeling. In a more recent paper, [44] proposed traffic prediction tools that employ either statistical, rule based, or deep machine learning methods. An extensive evaluation of the designed tools was performed and the impact of different parameters in prediction accuracy was investigated.

While the existing approaches provide valuable contributions for traffic prediction at a base station, most studies have merely considered general traffic arrival scenarios with sporadic features. Therefore, prediction of a bursty traffic arrival specially considering a massive number of devices needs further investigation. We attempt to fill this research gap in Paper C of Part II as briefly outlined in the next subsection.

4.4 Outline of the Proposed Scheme

4.4.1 Traffic Prediction Model

In Paper C, we have adopted a machine learning based approach for traffic prediction and based on the predictions resource allocation for a GB procedure is presented. Specifically, a bursty traffic arrival process is considered. As illustrated in Fig. 4.1, the gNB is only aware of the number of successfully detected devices at each RA slot. Indeed, the number of initial arrivals and the total number of competing devices resulting from the addition of retransmissions to initial arrivals are unknown to the gNB. In Paper C, we propose a scheme to predict the initial arrivals based on the pattern of detected data variation at the gNB for a bursty traffic arrival.

The prediction utilizes supervised learning with a spanning tree based algorithm to predict arrivals. The model is trained using the arrival data and the resulting detected data obtained from an analytical modeling of the LTE-A RA process similar to [22] for a device population of 30000 devices. The model was then tested using the detection data at the gNB obtained from the simulation data of the LTE-RA process.

For the proposed prediction model, various supervised learning algorithms for prediction are tested and the one which provides the minimum root mean squared error between the observed and predicted data is selected as the machine learning algorithm for the prediction model. Based on the prediction, the gNB is able to identify a traffic burst by observing the predicted number of initial arrivals. When the predicted number of arrivals exceeds a given threshold, the gNB considers the arrival traffic pattern to have changed from normal to bursty.

4.4.2 Arrival Prediction based Preamble Allocation

Based on the arrival prediction, we develop a preamble allocation scheme, APPA which stands for arrival prediction based preamble allocation. For dynamic preamble allocation, we consider a scenario similar to the one which is mentioned in Paper A of Part II. Here, the grouped devices have L priority levels based on their URLLC requirements. Based on the predicted arrival levels, once the predicted traffic intensity exceeds a given threshold, dynamic preamble allocation for different GD levels will be enabled. Through such dynamic traffic aware preamble allocation, GDs will receive dedicated preambles based on their priority level and the detected bursty traffic level. In Paper C, simulations were performed considering three levels of groups with different bursty thresholds as the group enabling criteria. As a result, the bursty nature of arrivals at the gNB reduces and the performance of dynamically enabled grouped devices improves considerably.

4.4.3 Further Extensions

The study in Paper C has been further extended in [56] with a main focus of predicting the future number of detections from a given seed length that consists of the observed detections. An LSTM based recurrent neural network architecture was utilized for the prediction task. LSTMs are especially powerful for time series data prediction thanks to

its feedback loops and the ability to remember previous inputs. Different from Paper C, the work in [56] focuses on predicting the number of preambles successfully detected at the gNB. Another difference between these two papers is that the initial data set used for traffic prediction in [56] is merely the data collected up to a point after which congestion might happen (e.g., when the RA slot index is close to 400 in Fig. 4.1) and the prediction can be long term (e.g., for RA slot index up to 2000), whereas all data accumulated up to the prediction instant are needed for arrival estimation performed in Paper C.

In addition, [56] considers both homogeneous (30000 total device population similar to Paper C) and heterogeneous traffic arrivals consisting of different total number of device populations (i.e., 10000, 20000, 30000) when training and evaluating the prediction model. From the simulation results, the homogeneous traffic based prediction provides more accurate results whereas the heterogeneous case requires longer seed length for better accuracy.

4.5 Chapter Summary

In this chapter, we first introduce the types and modelling of traffic arrivals for mMTC networks. Thereafter, the motivation for traffic prediction followed by a summary of existing prediction approaches in the literature is presented. Afterwards, an outline of the proposed scheme in Paper C of Part II and a further enhancement is provided.

Chapter 5

User Reliability and Availability: A Dependability Theory Perspective

In this chapter, we first introduce existing availability and reliability measuring metrics that are prevailing in wireless networks. Second, we extend our discussions with a dependability theory based perception. Thereafter, a discussion on space and time domain availability is presented. Then, the outline of Paper D on Part II is provided.

5.1 Evaluating Availability and Reliability of a Wireless Network

The availability and reliability levels provided by a network to end users are two major key performance indicators of a wireless network. In general, the reliability of a wireless system from a conventional reliable communication point of view is measured by metrics such as bit error rate and packet delivery ratio. The number of bit errors is the number of received bits of a data stream over a communication channel that have been altered due to noise, interference, distortion, or bit synchronization errors. Bit error rate is defined as the number of bit errors occurred during a unit time. Packet delivery rate is the ratio between the number of packets delivered successfully at the intended receiver and the total number of packets sent from a transmitter to a receiver in the network during a given period of time.

According to the 3GPP [57], reliability can be evaluated by the success probability of transmitting a certain number of bytes within a certain delay. The 3GPP has set a general URLLC reliability requirement for one transmission of a packet as $1 - 10^{-5}$ for a packet size of 32 bytes with a user plane latency of 1 ms.

In 5G, URLLC services such as high-precision robot control, autonomous vehicles, factory automation over wireless links demand ultra-high reliability levels. The number $1-10^{-5}$ is based on packet delivery ratio and this metric inherently assumes the availability of the network and mainly considers the reliability and latency aspects. Concurrently, 5G is expected to provide anytime and anywhere communications to end users. For facilitating such communications, a service should be available in both time and space domains.

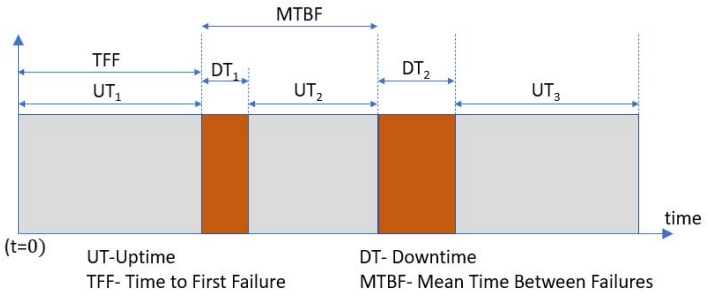


Figure 5.1: Illustration of operational state variation of a repairable system [59].

5.2 Dependability Theory Basics

Dependability is the ability of a system to operate as it is specified regardless of errors and to deliver intended level of services to its users [7]. Dependability theory deals with two types of systems, repairable and non-repairable. A repairable system is the one which can be restored to intended operation by taking necessary repair actions such as replacement of failed components or modification of configurations. In non-repairable systems, a system cannot be restored after the first failure. A failure is an event when a system transits from its specified operational state to a state where it cannot perform the intended level of services. Repair is the opposite transition of going back to the intended performance state from a failed state.

5.2.1 MUT, MDT, MTBF, and MTFF

Fig. 5.1 represents a repairable system with the system transferring from operational state to a failed state and vice versa. The mean time the system performs as intended is denoted as mean up time (MUT). The mean time that the system stays in the failure state is denoted as mean down time (MDT). For repairable systems, the mean time between failures (MTBF) is the mean time interval between two consecutive failure events.

On the other hand, the mean time to first failure (MTFF) is the time interval between the starting time of operation and the time instant when the first failure occurs. This is the main reliability measure for non-repairable systems.

5.2.2 Reliability and Availability

Moreover, the availability of a system can be analyzed mainly in two ways, i.e., instantaneous (point) availability and steady state availability. Instantaneous availability, denoted as $A(t)$, is the probability that the system is operating properly and is available to perform its functions at a specified time t, i.e.,

$$A(t) = P(Y(t) = 1) = E[Y(t)], \tag{5.1}$$

where $Y(t)$ has binary values either 1 or 0, as $Y(t) = 1$ if the system is operating at time t, and $Y(t) = 0$ otherwise [58]. The steady state availability, A_{SS}, can be obtained as the

average availability over a sufficiently longer period of time and it is expressed as

$$A_{SS} = \frac{MUT}{MUT + MDT}. \tag{5.2}$$

The reliability of a system is defined as the probability that a system will perform its intended functions without failure for a given interval of time under specified operating conditions [60].

5.3 System Level versus End User Availability in Time and Space Domains

From a network operator's perspective, it is essential to have a metric to measure *anytime and anywhere* operation of their networks. Such a measure would allow operators and their end users to benchmark different operators and to make comparisons among the services provided. For a credible claim of providing *anytime and anywhere* operation, a 5G network should be, regardless of its location and/or time, available to any end user for successful communication. The anytime component of such communications can be evaluated by using a time domain metric such as the instantaneous or steady state availability presented above. On the other hand, space domain availability is often overlooked and it is assumed to be present when reliability is evaluated.

As an effort from our research group, the authors in [61] first introduced a dependability theory based availability concept on space domain availability. The definitions on cell availability and system availability were presented and the availability as well as the probability of providing a guaranteed level of availability in a network were analyzed. That work was further extended in [62] and further system reliability analysis was performed therein by considering signal to interference ratio as a reliability measure. Additionally, several recent studies have contributed towards the development of this research area. In [63], system reliability was investigated from the viewpoint of space domain considering regular, irregular, and hybrid grid construction methods. In another work related to space domain analysis, [64] defined metrics for cell availability, system availability, individual user availability, and user-oriented system availability targeted at a cellular network.

While the above studies focused on space domain availability, we introduced a per-user availability metric combining both space and time domain parameters in our study [65]. That work is further consolidated with analysis considering user mobility and reliability impairments in Paper D of Part II. Therein, channel unavailability due to channel occupied or failed status is considered as a reliability impairment that decreases the network reliability and availability for a new user. [66] introduced another novel reliability metric that indicates the reliability level for downlink transmissions considering both time and space domain parameters. Similarly, in [67], the authors also presented availability and reliability metric definitions considering both spatial and temporal availability by taking into account the probability to achieve a targeted SIR threshold with a given confidence level to represent spatial reliability. Therein, the temporal availability was investigated

through a space availability driven channel access scheme and thereby, coupling the relation between spatial/temporal availability and reliability.

5.4 Outline of the Study in Paper D

Compared to the other parallel studies which mainly focused on system level availability [4], Paper D of Part II instead investigates in the per-user availability experienced by end users. We also analyze the impact of reliability impairments in both spatial and temporal domains and introduce a mobile user to reflect both the anytime and anywhere aspects. In the following, we discuss the main elements of the study in Paper D.

In Paper D, the proposed end user availability metric is evaluated under the presence of reliability impairments and user mobility. We consider error-prone channels together with channel occupancy status as examples of reliability impairments. The study consists of the following main tasks and will be elaborated further in separate subsections below.

- Representing node distribution of a multiple cell heterogeneous network using a Poisson point process (PPP).
- Reflecting user mobility via two different mobility models.
- Representing reliability impairments with channel impairment and modeling channel status behavior with a Markov chain.
- Proposing cell selection criteria at cell intersection areas.

5.4.1 Modeling Heterogeneous Multi-cell Network

To model a multi-cell cellular network, stochastic geometry which is a popular tool for modeling different aspects of wireless networks is applied. In particular, spatial point processes are adopted to model the distribution of locations of network nodes within a given geographic area. While spatial point processes are used for modelling both user and base station distributions, we do not focus on modelling user distributions in this study. The network nodes in our context refer to base stations which are assumed to be represented by points in the Euclidean space and they form a spatial point process.

Particularly, PPP is adopted to model the locations of network nodes in wireless networks due to its complete independence among points. A PPP represents a set of nodes distributed with a constant node density λ. Furthermore, there are several other point processes that also consider node repulsion. For example, if we consider the locations of base stations in reality, two base stations from the same network operator may not be placed very closer to each other. Therefore, other point processes like Ginibre point process and determinantal point processes take into account the repulsion between network nodes represented by the repulsion among the points [68].

Nevertheless, in Paper D, we employ a PPP to model the locations of base stations due to its tracebility as well as wide applicability. In general, cell coverage could change based on factors like different transmission power levels and antenna heights. Therefore,

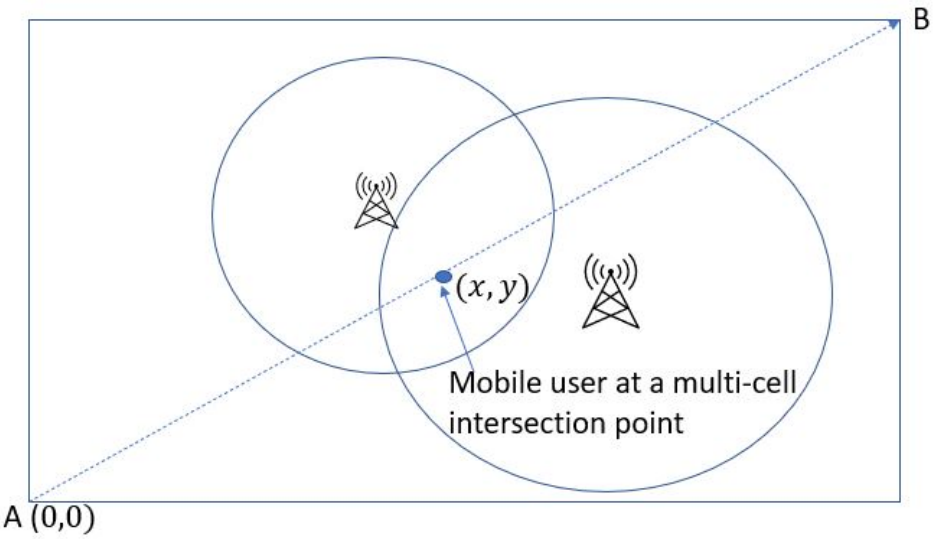


Figure 5.2: Mobile user at a point intersected by two cells.

we consider a heterogeneous multi-cell scenario where the cell coverage areas could differ cell wise.

5.4.2 Modeling User Mobility

Mobility models characterize the movements of mobile users with respect to their locations, velocities and directions over a period of time. There are several mobility models available to represent user mobility. Among them, random walk, random waypoint, and random direction are popular models for user mobility [69]. In our study, we employ two mobility models. The first one is a customarily designed model to represent the movement of a user from the origin $A(0, 0)$ depicted in Fig. 5.2 to the B point at the upper corner of the diagonal. The user may take different paths based on the angle with the x axis. In the second model, we adopt the random direction model which is a variant of the well-known random way point model within a rectangular area.

5.4.3 Reliability Impairments and Modeling of Channel Status

As mentioned in Subsection 5.2.2, the reliability and availability of a network could be impaired by several factors known as reliability impairments. These include factors like excessive interference, resource constrains, and system failures which need to be addressed when evaluating the URC level provided by a network operator.

In this study, channel unavailability is modeled as a reliability impairment. Ideally, a user would prefer a channel to be available whenever a service is required and the channel to successfully deliver the intended communication task. In reality, however, an incoming user may experience channel congestion due to resource limitations. On the other hand, although a channel is available, the communication process could still be affected by poor channel conditions and failures. Therefore, such channel conditions are also regarded as reliability impairments in our model.

In order to model channel status, a discrete time Markov model consisting of three states is constructed. In our model, states represent the status of being idle, busy, or failed channel. A channel is regarded as available to a new user only when it is in the idle state. Otherwise, it is considered to be an impaired channel in view of the new user.

5.4.4 Cell Selection Criteria for Cell Intersection Areas

In a multi-cell deployment within an area of interest, a mobile user may come across areas where it will be served by more than one cell. In such a situation, the end user needs to select one specific cell which it will be associated with. As depicted in Fig. 5.2, the user at the position (x, y) is currently covered by two cells and can be served by both of them. Based on the proposed availability metric, we propose three different strategies and criteria for the user to make this decision. While the first two strategies are mainly focusing on the availability for a new user, the third strategy emphasizes the reliability of ongoing sessions. Since end users have different availability and reliability requirements, it is reasonable to assume that end users can make a decision of cell association based on the input from base stations.

The first strategy is to select the cell with the highest steady state probability at the idle state. This strategy enables higher probability for a new user to access the cell. Therefore, a mobile user at an idle user state (without an ongoing voice or data session) would prefer to associate with this cell.

In the second strategy, the cell with the lowest state holding time at the failed state will be selected. Specifically, the state holding times are considered for such a decision making process. By doing so, the strategy aims to minimize the time spent on a failed state if any failure occurs. In other words, the cell status would be quickly transitioned from a failed state to an idle state thereby giving more opportunities for the user to communicate.

As the third strategy, the cell with the lowest occupied to failed transition probability will be selected. This strategy is designed by considering the reliability preference of ongoing communications. For an ongoing session, it is preferable to select a cell which is least likely to fail. As such, maximum reliability for ongoing traffic can be reached.

5.5 Chapter Summary

In this chapter, we first provide requirements and existing approaches for user reliability and availability measurement in wireless networks. Thereafter, a basic overview of dependability theory concepts is provided. Thereon, we present a summary of system level versus end user availability evaluation approaches from the literature followed by an outline of the work performed in Paper D.

Chapter 6

Conclusions and Future Work

This final chapter is organized into three sections. The first section presents the conclusions of the work presented in this dissertation. Afterwards, the main contributions of this thesis work are highlighted. To continue the research work relevant to the issues addressed in this thesis towards beyond 5G networks, a few potential research directions are pointed out.

6.1 Conclusions

5G NR facilitates various futuristic application scenarios with stringent requirements. The advent of 5G has initiated a surge of research interests in many areas, within and beyond wireless and mobile communications. In this dissertation, we endeavor to answer several research questions related to the URLLC and mMTC technological directions in 5G NR networks. Our major focus lies on initial access approaches for 5G medium access based on both GB and GF communications. In addition, we also make efforts on traffic arrival predictions and user availability evaluation. The following major conclusions can be reached from the research work performed in this dissertation.

Firstly, the proposed GB access schemes that employ device grouping or/and RA resource grouping techniques facilitated by the flexibility offered in the 5G NR frame structure exhibit superior performance in comparison with the legacy LTE-A RA procedure. Thanks to the benefits obtained through grouping, the proposed schemes enable reliable initial access with lower latency for critical mMTC devices with URLLC service requirements. Meanwhile, the performance of other non-grouped or non-URLLC devices is also improved as a consequence of reduced competing devices achieved through the grouping process.

Secondly, the proposed DSA-GF scheme enables service differentiation for heterogeneous traffic arrivals facilitating priority access to HPT devices without starving LPT access opportunities. The proposed 2D Markov chain accurately captures the dynamic behavior of the scheme. The performance evaluations and comparisons with traditional GF schemes reaffirm the advantages of the DSA-GF scheme with respect to throughput, delay, and packet loss.

Thirdly, the proposed supervised learning based arrival prediction model instigates a

novel approach based on the detected data patterns at a gNB in order to identify bursty traffic arrivals in an mMTC scenario. This capability enables the gNB to proactively and dynamically allocate preamble resources to various priority device groups. Consequently, the impact of bursty traffic arrivals on priority devices is diminished compared with that in traditional LTE-A based schemes.

Finally, the study proposes a metric to measure end user availability when both time and space domain reliability impairments are considered. Such a metric provides a valuable performance indicator when evaluating the anytime and anywhere performance of 5G networks.

6.2 Contributions

The main contributions of this dissertation are summarized as follows.

- Three GB initial access schemes are proposed. The DGDP, RAUG, and hybrid schemes enable URLLC or high priority devices to access the network with higher reliability and lower latency through either device or resource grouping by taking the advantage of the 5G NR frame structure. Through simulation and analysis, the performance of the proposed schemes in terms of reliability and latency is evaluated and compared with the legacy LTE-A RA and GF schemes.

- A GF dynamic slot allocation scheme is proposed to enable heterogeneous traffic access in an mMTC network. The proposed scheme facilitates the co-existence of high priority and low priority traffic by dynamically allocating the available GF slots for both traffic types. While the scheme provides better performance for high priority traffic, it also prevents starvation for low priority traffic. A 2D Markov chain model that combines the two types of traffic through a pseudo-aggregated process is developed. The performance of the proposed scheme is evaluated through both analysis and simulations and compared with two existing GF schemes.

- A machine learning based scheme for bursty traffic arrival prediction is proposed. The proposed model is capable of predicting bursty traffic arrivals at the base station. Based on the prediction, a resource allocation method is proposed.

- An end user availability evaluation metric is developed from a dependability theory perspective. Based on the metric, the availability of a mobile user is evaluated when reliability impairments are taken into account. Furthermore, three strategies for cell selection at a cell intersection area are proposed.

6.3 Future Directions

The work conducted during this thesis work can be extended towards multiple directions. In the following, we list four potential directions for future research.

- In Paper A, device grouping and group leader selection were assumed to be performed beforehand to facilitate the proposed DGDP scheme.Therefore, it is worthwhile to investigate and propose device grouping and group leader selection mechanisms as extensions to our work. Thereafter, different inter- and intra-group communication schemes and the resulting latency as well as complexities need to be explored.

- The analysis in Paper B was performed based on an assumption on ideal channel conditions. For future work, a promising direction is to consider channel impairments and the possibility of multi-user detection at the gNB.

- Artificial intelligence and machine learning are recognized as enabling technologies for the design and optimization of beyond 5G networks. Therefore, embodying intelligence to beyond 5G mMTC/mIoT networks by introducing network intelligence to edge nodes including both gNBs and IoT devices represents another promising research direction.

- Next generation wireless systems advocate extreme requirements of high reliability and lower latency for supporting intelligent services built upon the Internet of senses and real-time human-machine interactions powered by augmented and virtual reality advancements. Accordingly, joint consideration of reliability and latency improvements in such a context exhibits another appealing research direction.

References

[1] 3GPP RAN, "Overview of RAN aspects," Workshop on 3GPP submission towards IMT-2020, Oct. 2018, Brussels, [Online] Available: `https://www.3gpp.org/news-events/1976-imt_2020`, accessed on 18 Dec. 2020.

[2] 3GPP TS 22.104, "Service requirements for cyber-physical control applications in vertical domains," R17, v17.4.0, Sep. 2020.

[3] M. Bennis, M. Debbah, and H. V. Poor, "Ultra reliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[4] 3GPP TS 22.368, "Service requirements for machine-type communications (MTC); Stage 1," R16, v16.0.0, Jul. 2020.

[5] 3GPP TS 36.101-1, "User equipment (UE) radio transmission and reception; Part 1: Range 1 standalone," R16, v16.0.0, Sep. 2020.

[6] 3GPP TS 38.211, "NR; Physical channels and modulation," R16, v16.3.0, Sep. 2020.

[7] G. Rubino and B. Sericola, *Markov Chains and Dependability Theory*. New York, NY, USA: Cambridge University Press, 2014.

[8] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free random access solutions for URLLC services in 5G new radio," in *Proc. IEEE ISWCS*, Aug. 2019, pp. 607-612.

[9] I. Leyva-Mayorga, C. Stefanovic, P. Popovski, V. Pla, and J. Martinez-Bauset, "Random access for machine-type communications," *Wiley 5G Ref: The Essential 5G Reference Online*, 2019.

[10] S. R. Pokhrel, J. Ding, J. Park, O. S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796-131813, Jul. 2020.

[11] 3GPP TS 36.321, "Evolved universal terrestrial radio access (e-UTRA)," R16, v16.2.0, Sep. 2020.

[12] 3GPP TS 36.211, "Physical channels and modulation," R16, v16.1.0, Mar. 2020.

[13] 3GPP TS 36.331, "Radio resource control (RRC)," R16, v16.0.0, Mar. 2020.

[14] 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, 1st Quart., 2014.

[15] M. Cheng, G. Lin, H. Wei, and A. C. Hsu, "Overload control for machine-type-communications in LTE-advanced system," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 38-45, Jun. 2012.

[16] 3GPP TR 37.868, "Study on RAN improvements for machine type communications," R11, v11.0.0, Sep. 2011.

[17] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86-93, Jun. 2013.

[18] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, Jan. 2017.

[19] 3GPP TS 22.011, "Service accesssibility," R17, v17.2.0, Sep. 2020.

[20] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," in *Proc. IEEE ICC*, May 2016, pp. 1-6.

[21] 3GPP TS 22.261, "Service requirements for the 5G system: Stage 1," R18, v18.0.0, Sep. 2020.

[22] C. Wei, G. Bianchi, and R. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940-1953, Apr. 2015.

[23] 3GPP TS 36.881, "Study on latency reduction techniques for LTE," R14, v14.0.0, Jun. 2016.

[24] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System level analysis of uplink grant-free transmission for URLLC," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2017, pp. 1-6.

[25] R. B. Abreu, *Uplink grant-free access for ultra-reliable low-latency communications in 5G: Radio access and resource management solutions*, Aalborg Universitet, 2019. [Online] Avaliable: `https://vbn.aau.dk/ws/portalfiles/portal/315080379/PHD_Renato_Barbosa_Abreu_E_pdf.pdf`, accessed on 18 Dec. 2020.

[26] T. Le, U. Salim, and F. Kaltenberger, "Strategies to meet the configured repetitions in URLLC uplink grant-free transmission," in *Proc. IEEE ISWCS*, Aug. 2019, pp. 597-601.

[27] S. Ozaku, Y. Shimbo, H. Suganuma, and F. Maehara, "Adaptive repetition control using terminal mobility for uplink grant-free URLLC," in *Proc. IEEE VTC*, May 2020, pp. 1-5.

[28] S. Moon and J. W. Lee, "Integrated grant-free scheme for URLLC and mMTC," in *Proc. IEEE 3rd 5G World Forum (5GWF)*, Sep. 2020, pp. 98-102.

[29] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE J. Sel. Areas Commun.*, Early access article available in IEEE Xplore, Aug. 2020.

[30] M. Ghanbarinejad and C. Schlegel, "Irregular repetition slotted ALOHA with multi-user detection," in *Proc. IEEE WONS*, Mar. 2013, pp. 201-205.

[31] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," in *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[32] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179-189, Apr. 2020.

[33] N. Alsbou, D. Henry, and H. Refai, "R-ALOHA with priority (PR-ALOHA) in non ideal channel with capture effects," in *Proc. 17th Int. Conf. Telecommun.*, Apr. 2010, pp. 566-570.

[34] J. Sun, R. Liu, Y. Wang, and C. W. Chen, "Irregular repetition slotted ALOHA with priority (P-IRSA)," in *Proc. VTC*, May 2016, pp. 1-5.

[35] S. Lam, "Packet broadcast networks - A performance analysis of the R-ALOHA protocol," *IEEE Trans. Comput.*, vol. C-29, pp. 596-603, Jul. 1980.

[36] T. L. Sheu and H. Lin, "A priority-based channel allocation scheme for cellular networks," in *Proc. IEEE WCNC*, Mar. 2003, pp. 1066-1071.

[37] J. Frigon and V. Leung, "A pseudo-Bayesian ALOHA algorithm with mixed priorities for wireless ATM," in *Proc. IEEE PIMRC*, 1998, pp. 45-49.

[38] J. Huang, Z. Xiong, J. Li, Q. Chen, Q. Duan, and Y. Zhao, A priority-based access control model for device-to-device communications underlaying cellular network using network calculus, In: Z. Cai, C. Wang, S. Cheng, H. Wang, and H. Gao, (eds) *Wireless Algorithms, Systems, and Applications*. Springer, Cham, Lecture Notes in Computer Science, vol. 8491, 2014.

[39] K. Akkarajitsakul, E. Hossain, D. Niyato, and D. I. Kim, "Game theoretic approaches for multiple access in wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 372-395, 3rd Quart., 2011.

[40] K. Cohen and A. Leshem, "Distributed game-theoretic optimization and management of multichannel ALOHA networks," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1718-1731, Jun. 2015.

[41] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Supervised learning based arrival prediction and dynamic preamble allocation for bursty traffic," in *Proc. IEEE INFOCOM Workshops*, Apr. 2019, pp. 1-6.

[42] S. Yang and F. A. Kuipers, "Traffic uncertainty models in network planning," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 172-177, Feb. 2014.

[43] Y. Xu, W. Xu, F. Yin, J. Lin, and S. Cui, "High-accuracy wireless traffic prediction: A GP-based machine learning approach," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1-6.

[44] A. Azari, P. Papapetrou, S. Denic, and G. Peters, "User traffic prediction for proactive resource management: Learning-powered approaches," in *Proc. IEEE GLOBECOM*, Dec. 2019, pp. 1-6.

[45] I. Loumiotis, E. Adamopoulou, K. Demestichas, P. Kosmides, and M. Theologou, "Artificial neural networks for traffic prediction in 4G networks," in *Proc. 8th Int. Wireless Internet Conf.*, Nov. 2014, pp. 141-146.

[46] H. D. Trinh, L. Giupponi, and P. Dini, "Mobile traffic prediction from raw data using LSTM networks," in *Proc. IEEE PIMRC*, Sep. 2018, pp. 1827-1832.

[47] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 3rd Quart., 2019.

[48] Y. Cui, X. Huang, D. Wu, and H. Zheng, "Machine learning based resource allocation strategy for network slicing in vehicular networks," in *Proc. IEEE ICCC*, Aug. 2020, pp. 454-459.

[49] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, "Spatio-temporal wireless traffic prediction with recurrent neural network," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 554-557, Aug. 2018.

[50] C. Huang, C. Chiang, and Q. Li, "A study of deep learning networks on mobile traffic forecasting," in *Proc. IEEE PIMRC*, Oct. 2017, pp. 1-6.

[51] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1389–1401, Mar. 2019.

[52] J. Hu, W. Heng, G. Zhang, and C. Meng, "Base station sleeping mechanism based on traffic prediction in heterogeneous networks," in *Proc. International Telecommunication Networks and Applications Conference (ITNAC)*, Nov. 2015, pp. 83-87.

[53] Y. Shu, M. Yu, J. Liu, and O. W. W. Yang, "Wireless traffic modeling and prediction using seasonal ARIMA models," in *Proc. IEEE ICC*, May 2003, pp. 1675-1679.

[54] T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, "Intelligent hybrid model to enhance time series models for predicting network traffic," *IEEE Access*, vol. 8, pp. 130431–130451, Jul. 2020.

[55] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *Proc. IEEE INFOCOM*, May 2017, pp. 1-9.

[56] A. Søraa, T. N. Weerasinghe, I. A. M Balapuwaduge, and F. Y. Li, "Preamble transmission prediction for mMTC bursty Traffic: A machine learning based approach," in *Proc. IEEE GLOBECOM*, Dec. 2020, pp. 1-6.

[57] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies," R16, v16.0.0, Jul. 2020.

[58] E. De Souza, E. Silva, and H. R. Gail, "Calculating cumulative operational time distributions of repairable computer systems," *IEEE Trans. Comput.*, vol. C-35, no. 4 pp. 322-332, Apr. 1986.

[59] I. A. M. Balapuwaduge, *Channel access and reliability performance in cognitive radio networks: Modeling and performance analysis*. Ph.D. dissertation, University of Agder, Norway, 2016.

[60] H. Pham, *System Reliability Concepts*. London, UK: Springer London, Ch. System Software Reliability, pp. 9–75, 2006.

[61] H. V. K. Mendis and F. Y. Li, "Achieving ultra reliable communication in 5G networks: A dependability perspective availability analysis in the space domain," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2057–2060, Sep. 2017.

[62] H. V. K. Mendis, I. A. M. Balapuwaduge, and F. Y. Li, "Dependability-based reliability analysis in URC networks: Availability in the space domain," *IEEE/ACM Trans. Netw.*, vol. 27, no. 5, pp. 1915–1930, Sep. 2019.

[63] H. B. Nafea, E. M. Soultan, and F. W. Zaki, "Availability of cellular systems based on space domain," *Wireless Pers. Commun*, vol. 107, pp. 1881–1910, Apr. 2019.

[64] Y. Benchaabene, N. Boujnah, and F. Zarai,"Ultra reliable communication: Availability analysis in 5G cellular networks," in *Proc. IEEE International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, Dec. 2019, pp. 96-1029.

[65] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Per-user availability for ultra-reliable communication in 5G: Concept and analysis," in *Proc. IEEE WCNC*, Apr. 2018, pp. 1-6.

[66] I. A. M. Balapuwaduge and F. Y. Li, "A joint time-space domain analysis for ultra-reliable communication in 5G networks," in *Proc. IEEE ICC*, May 2018, pp. 1-6.

[67] M. Emara, M. C. Filippou, and I. Karls, "Availability and reliability of wireless links in 5G systems: A space-time approach," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2018.

[68] M. Haenggi, N. Jindal, J. G. Andrews, R. K. Ganti, and S. Weber,"A primer on spatial modeling and analysis in wireless networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 156–163, Nov. 2010.

[69] H. Tabassum, M. Salehi, and E. Hossain, "Fundamentals of mobility-aware performance characterization of cellular networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 2288-2308, 3rd Quart., 2019.

PART II

Paper A

Title:	Priority-based Initial Access for URLLC Traffic in Massive IoT Networks: Schemes and Performance Analysis
Authors:	Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, and Frank Y. Li
Affiliation:	Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway
Journal:	*Computer Networks*, vol. 178, Article 107360, September 2020
DOI:	10.1016/j.comnet.2020.107360
Copyright ©:	Elsevier

Priority-based Initial Access for URLLC Traffic in Massive IoT Networks: Schemes and Performance Analysis

Thilina N. Weerasinghe, Indika A. M. Balapuwaduge,
and Frank Y. Li

Abstract - At a density of one million devices per square kilometer, the 10's of billions of devices, objects, and machines that form a massive Internet of things (mIoT) require ubiquitous connectivity. Among a massive number of IoT devices, a portion of them require ultra-reliable low latency communication (URLLC) provided via fifth generation (5G) networks, bringing many new challenges due to the stringent service requirements. Albeit a surge of research efforts on URLLC and mIoT, access mechanisms which include both URLLC and massive machine type communications (mMTC) have not yet been investigated in-depth. In this paper, we propose three novel schemes to facilitate priority-based *initial access* for mIoT/mMTC devices that require URLLC services while also considering the requirements of other mIoT/mMTC devices. Based on a long term evolution-advanced (LTE-A) or 5G new radio frame structure, the proposed schemes enable device grouping based on device vicinity or/and their URLLC requirements and allocate dedicated preambles for grouped devices supported by flexible slot allocation for random access. These schemes are able not only to increase the reliability and minimize the delay of URLLC devices but also to improve the performance of all involved mIoT devices. Furthermore, we evaluate the performance of the proposed schemes through mathematical analysis as well as simulations and compare the results with the performance of both the legacy LTE-A based initial access scheme and a grant-free transmission scheme.

Keywords - mIoT and mMTC, URLLC, LTE-A and 5G NR, initial access.

A.1 Introduction

While the Internet of things (IoT) is revolutionizing our society at an unprecedented pace, more recent research and development focus on IoT is shifting towards the direction of massive IoT (mIoT). In parallel with this trend, *massive machine type communications (mMTC), which is an enabling technology for mIoT*, has been envisaged as one of the three major use cases for the fifth generation (5G) mobile and wireless networks. Indeed, the popularity of mIoT arises from the ever-increasing data traffic spurred by various applications ranging from smart cities to mission critical communications in cyber-physical systems and Industry 4.0 [1]. Consequently, the ever-growing network size, heterogeneity in applications, and energy constraints pose various new challenges for mIoT related research [2]- [4].

Together with mMTC, enhanced mobile broadband (eMBB) and ultra-reliable and low latency communication (URLLC) are the other two use cases for 5G applications. The current standardization activities led by the 3rd generation partnership project (3GPP) focus mainly on eMBB, which represents an evolutionary path from long term evolution-advanced (LTE-A) in order to provide ultra-high data rates to end users for applications like high resolution video streaming. Meanwhile, there is a surge of research interests in mIoT/mMTC and URLLC from both academia and industry [5]- [9]. For mIoT/mMTC applications including automated energy distribution in a large smart grid, control of large-scale industrial processes, and surveillance of critical infrastructure, how to provide medium access to a huge volume of devices appears as a challenging task. In contrast to eMBB, the URLLC use case focuses on achieving ultra-high levels of reliability and low latency for futuristic scenarios like remote surgery, remote monitoring and control, as well as augmented and virtual reality [9] [10]. For many applications, it is expected that the reliability level reaches 99.9999% or higher and the device to network latency becomes less than 1 ms [10]. However, achieving stringent URLLC in 5G is extremely challenging especially when considering that ultra-reliability and low latency represent two contradictory requirements. For instance, achieving high reliability requires parity check, coding or link redundancy, and packet retransmissions which in turn increase latency [9].

Addressing these mIoT and 5G challenges calls for novel approaches for system development and protocol design. Although a lot of work on eMBB has been done, URLLC and mIoT/mMTC are expecting more innovative contributions from the research community. Among others, one of the most paradoxical research questions to be answered is *how to satisfy service requirements when both mIoT/mMTC and URLLC are jointly taken into consideration*. This point is especially important for *initial access of IoT devices which occurs before actual data transmissions*. It is known that existing LTE/LTE-A based random access (RA) procedures are inefficient when there are a large number of device arrivals simultaneously, due to the constraint of a limited number of preambles or/and radio resource blocks for uplink or downlink traffic [2]. Although numerous initial access schemes have been proposed for fourth generation (4G) networks, the problem becomes more complex in 5G new radio (NR) since 5G NR Phase 1 is more advanced but still based on orthogonal frequency division multiple access (OFDMA). In an mIoT network, when traffic volume is high especially under bursty traffic conditions, the number of attempts for initial access could rise substantially, leading to high collision, low access success probability, and correspondingly increased latency. As such, it is imperative to develop customized solutions in 5G for devices that require URLLC access among mIoT devices.

In this paper, we propose three initial access schemes addressing the aforementioned research question. Considering a large number of mIoT/mMTC devices covered by a cell, we focus on providing ultra-reliable and low latency access for a portion of devices that require URLLC services. The proposed novel schemes utilize device grouping and resource grouping for low latency communications based on the LTE-A or NR frame structure. Furthermore, the performance of these schemes is analyzed mathematically based on an existing comprehensive model which was initially developed for LTE traffic but with our extension to fit the proposed schemes in our envisaged LTE-A and NR

scenarios. Extensive simulations are performed to validate the model and compare the performance of our schemes with that of three existing schemes.

In brief, the main contributions of this paper are summarized as follows.

- Three initial access schemes are proposed with the aim of providing services for mMTC[1] devices in two scenarios with location-bounded and location-spread URLLC devices, respectively. These schemes are specifically designed considering bursty traffic arrivals, posing a worst case scenario for devices sharing resources for initial access.
- Based on the advanced features of numerology and the frame structure in NR, a novel RA slot allocation method which enables flexible URLLC grouping is proposed. Accordingly, collisions among URLLC access contentions and latency are minimized.
- The performance of the proposed schemes is evaluated through analysis and simulations by taking into account a massive number of devices contending for network access and compared with the performance of both the existing LTE-A RA scheme which serves as a baseline scheme and with a grant-free (GF) transmission scheme.

The rest of the paper is organized as follows. Sec. A.2 summarizes the related work. Then, Sec. A.3 provides preliminaries to help readers better comprehend the work presented in the paper. In Sec. A.4, the network scenarios and assumptions are presented. In Sec. A.5, the proposed schemes are explained in details, followed by performance analysis in Sec. A.6. Thereafter, Sec. A.7 illustrates the numerical results. Finally, the paper is concluded in Sec. A.8.

A.2 Related Work

As an enabling technology for mIoT operation in licensed bands, mMTC follows the procedures defined by 3GPP. Since these procedures are highly relevant to the work presented in this paper, we first outline existing solutions for RA channel (RACH) congestion avoidance for initial access that occurs prior to data transmissions in LTE-A and 5G NR and then introduce a few mathematical models for LTE-A RA process performance evaluation.

A.2.1 RACH Congestion in LTE-A: Initial Access and Solutions

A main constraint of the LTE-A RA process is the limited number of preambles available in a cell, e.g., 64 preambles within one RA slot (to be clarified in Sec. A.4). Out of these 64 preambles, a certain amount, typically 10, is reserved for contention-free transmissions while the rest is shared by other devices. RA collision occurs when multiple devices select the same preamble to transmit in the same RA slot (to be clarified in the next section), causing unsuccessful detection of transmitted preambles at the evolved nodeB (eNB) [13]. This in turn results in an increased number of retransmissions, further escalating the problem.

[1] In the rest of this paper, the terminologies, IoT and MTC, or mIoT and mMTC, are interchangeably used.

In [2] [41], 3GPP recommended several solutions to resolve this problem. Two of the most popular approaches are access class barring (ACB) and extended access baring (EAB) [41] [14]. Initially, ACB provides an effective access control mechanism in order to prevent potential overload of a network. In ACB, devices are classified into multiple classes with different priority levels. An eNB broadcasts the configuration information periodically through the master information block (MIB) and system information block (SIB) messages. Via SIB Type 2 (SIB2), the eNB broadcasts the current ACB configurations including a barring rate and a barring timer to guide various classes of devices to run a random access procedure in case of possible network overload. When a device intends to access the channel, it will pursue a random access procedure if its selected random number is lower than the barring rate. Although ACB provides higher priority devices with higher access probabilities, it does not guarantee their access privilege [15]. This is because ACB schemes still follow contention based access and collisions could still happen for example when there are too many high priority devices.

The performance of ACB schemes may vary with different parameter configurations. In [32], an ACB scheme for dealing with physical RACH (PRACH) overload was studied and the impact of its configuration parameters on network performance was analyzed. In [33], an optimal ACB control and resource allocation scheme to acquire system capacity under a limited total number of resource blocks was proposed.

Furthermore, in order to prevent overload of the network, EAB introduces another more restrictive method to control access attempts from devices that can tolerate more access restrictions for instance MTC devices which can tolerate longer delays. EAB provides a *deterministic* access control mechanism, preventing devices belonging to certain types access classes from obtaining access [1]. If congestion occurs, the network could restrict the access of these classes of EAB devices while still allowing access from other EAB devices specified through the advertised SIB messages and ACB devices according to the barring rate [2].

On the other hand, in both ACB and EAB, the detection of traffic conditions by an eNB is performed in a reactive manner and devices also behave passively based on the received SIB messages. Although these schemes improve the access success of higher priority devices, such behavior will cause additional delays which are detrimental for achieving low latency communications, especially upon the arrival of a traffic burst.

In addition to ACB and EAB, [2] has also proposed several other schemes. For instance, an MTC specific backoff approach introduces separate backoff times for MTC and human type communication (HTC) traffic by assuming that HTC traffic always has higher priority. However, when it comes to URLLC, we cannot prioritize HTC traffic over MTC traffic as both types will have similar importance levels. Other approaches include slot based and pull based access or eNB initiated access. For uplink URLLC access, however, these approaches may not be efficient since URLLC devices cannot wait until the eNB has initiated a communication process. In [45], the coexistence of scheduled and non-scheduled URLLC services and the difficulties for achieving stringent latency requirements under such a scenario were discussed. Furthermore, grouping based methods have also been studied for collision avoidance in LTE-A RA. In [16], a grouping based method was proposed to diminish collisions at the eNB. Using this method, all group devices

send their data to a group coordinator based on device-to-device (D2D) communications and group coordinators transmit uplink data following the standard 4-step RA procedure. This scheme was further analyzed in [17]. Recently, a compressed sensing based RACH protocol was proposed in [18].

Furthermore, cluster based access schemes were proposed in [35] [36] to mitigate potentially severe collisions of MTC devices that access to an eNB concurrently. In another study performed in [37], spatial group based reusable preamble allocation was proposed. According to clustering-reuse preamble allocation proposed in [36], complementary preamble sets are allocated to clusters with similar distances and the same preamble set is allocated to clusters that are far away. In [38], a cluster based group paging scheme for congestion and overload control was proposed. This method is based on IEEE 802.11ah by collecting the sensed data from MTC devices and upload data to the LTE/LTE-A cellular network. However, 802.11ah limits the number of devices.

In a nutshell, although many schemes have contributed to a large extent RACH congestion avoidance, most of them are targeted at LTE-A networks without considering the stringent low latency requirements for URLLC services. Despite much progress, the performance gap for RA in terms of providing ultra-high reliability and low latency simultaneously in mMTC networks remains largely unresolved and calls for more research efforts.

A.2.2 Initial Access for 5G NR

For medium access in NR Phase 1, an OFDMA based RA scheme similar to the LTE-A RA scheme was recommended [3] [20]. Its main difference in comparison with LTE-A is the introduction of beam steering techniques for synchronization in higher frequency operations, as further discussed in Sec. A.3 below. Additionally, the NR frame structure with shorter transmission time intervals (TTIs) ensures faster RA process and allows more flexible numerology [22] [20]. In general, with proper parameter tunning, the ACB and EAB mechanisms presented above which are initially designed for LTE/LTE-A are also applicable to NR Phase 1 initial access.

Additionally, there have been numerous access schemes proposed for 5G NR. Among them, [23] proposed a contention based access scheme by allowing multiple transmissions of the same packet in consecutive TTIs. By deducing the optimal number of consecutive transmissions, the low latency and high reliability requirement can be satisfied. Another type of popular approaches is grant-free access, also known as configured grant [7] [25], in which devices are allowed to transmit their data messages without following the standard grant based (GB) process [24] [8]. In [24], a GF radio access scheme was proposed for low complexity IoT devices where highly reliable access with bounded delay was achieved with long battery lifetime. Accordingly, devices directly transmit their data packets in pre-configured grant-free slots defined by the next generation NodeB (gNB). Rather than waiting for an acknowledgment (ACK) or negative ACK (NACK) message which takes additional time, a device may transmit replicas of its message up to k times in randomly selected k GF slots within a subframe for achieving high reliability and low latency. When multiple devices transmit at the same time, different techniques like successive

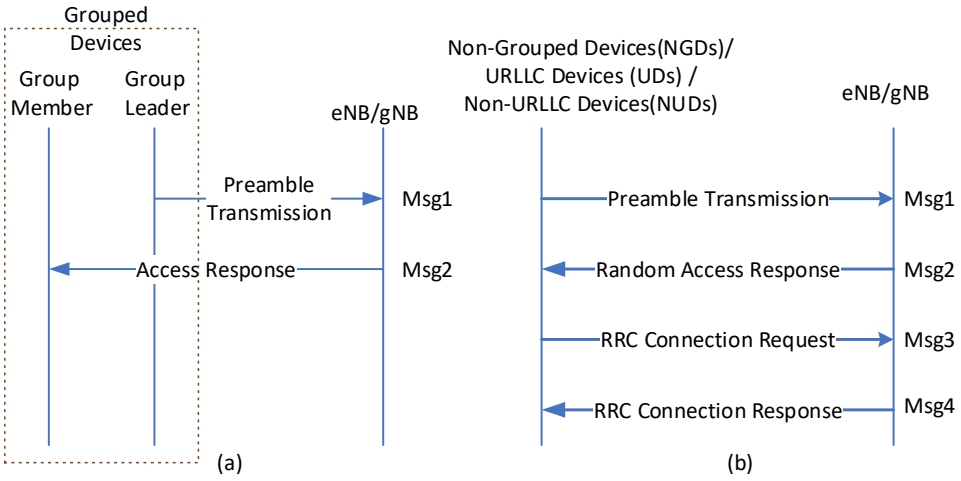


Figure A.1: Illustration of (a) the 2-step access procedure for UDs and (b) the 4-step access procedure for LTE-A, NGDs, UDs, and NUDs.

interference cancellation (SIC) can be employed to cancel out interference and detect data associated with a specific user. However, GF transmissions are targeted at small size data packets with sporadic arrival patterns [1]. When a large number of devices transmit at the same time, grant-free access could result in high collision probability and increased delay considering the additional time required for resolving collisions [7]. As such, how to ensure URLLC in 5G NR based mIoT networks remains as an open research question.

A.2.3 Modelling LTE-A RA Process

Modeling precisely an LTE-A RA procedure is not an easy task. As mentioned in Sec. I of [28], the performance evaluation of RA schemes is oftentimes conducted by means of simulations due to the fact that the RA procedure of LTE-A is difficult to model analytically. Among the research efforts reported in the literature, [29] provided a model with a focus on the first preamble transmission. Although few other analytical models that consider the complete RA process exist, the accuracy of these models needs to be improved when comparing with simulation results. In [43], a general model to analyze the performance of the RACH procedure was proposed and validated via simulations, focusing on the case of highly synchronized MTC traffic. Furthermore, an in-depth review on the accuracy of existing models was presented in [28]. However, most of these models have ignored access delay which is a key performance indicator. This aspect is especially important in the case of URLLC since the latency performance needs to be properly analyzed. [8] presented a comprehensive analytical model for performance evaluation of the LTE based RA process which also serves as the basis for our performance analysis presented later in Sec. A.6. Therein, the authors adopted Stirling numbers of the second kind to derive an exact expression for the probability distribution of the number of successful preamble transmission attempts over multiple RACH slots. Moreover, the drift approximation was used to model a complete and detailed LTE RA procedure based on a 3GPP standard [7].

Furthermore, it is worth mentioning that the schemes proposed in this paper differ from existing work in several ways. Firstly, a salient feature of this work is the consideration of

both mMTC and URLLC requirements that is largely overlooked in most other studies. Secondly, the proposed schemes are built on top of the LTE-A or NR based RA procedure and we advance the state-of-the-art techniques by introducing priority based grouping approaches for initial access of URLLC traffic. Thirdly, unlike other existing priority based approaches for instance ACB and EAB, which do not provide guaranteed access with low latency, our schemes ensure access privilege based on device grouping or RA slot grouping, providing URLLC devices with guaranteed or highly probable access. Lastly, while most other schemes like ACB and EAB follow a *reactive* principle as mentioned above, our schemes behave in a *proactive* manner which is beneficial for achieving low latency and the parameters are reconfigurable. By proactive, it is meant that device grouping is performed in an intended manner and a dedicated preamble is assigned to each group leader. The parameters involved in this procedure, e.g., number of devices in each group, are configurable, however, over a comparatively long period much larger than a MIB or SIB cycle.

A.3 Preliminaries

This section provides preliminaries that form the bases for the schemes to be presented in the rest of this paper.

A.3.1 RA Process in LTE/LTE-A and 5G NR

An RA process occurs when devices require initial access, e.g., upon network deployment or update, or transition from an idle mode to a connected mode. Such an RA process needs to be performed for initial access, after a signaled disconnection from the gNB, or a device has just woken up from the power saving or sleep mode. The LTE/LTE-A RA process recommended by 3GPP consists of the exchange of four handshake messages between a device and its associated eNB, as illustrated in Fig. A.1(b).

- *Step 1 (Msg1): Preamble transmission.* Whenever a device needs to communicate with an eNB, it first selects an RA preamble from a set of available preambles and transmits it in the next available RA slot. *An RA slot is a subframe within which devices are allowed to send their selected preambles.* It is defined by eNB and broadcast periodically over paging cycles via the SIB2 messages.

- *Step 2 (Msg2): Random access response (RAR).* When the eNB receives preamble transmissions without collision, it transmits Msg2 in the handshake process. Through RAR, the eNB schedules uplink resources for the transmission of the next message. Additionally, RAR contains also information about the detected RA preamble sequence, for which the response is valid, timing advance details, and a cell radio-network temporary identifier (C-RNTI) for further communication of a particular device.

- *Step 3 (Msg3): Radio resource control (RRC) connection request.* Using the received C-RNTI and uplink resources, the device transmits its RRC request to the eNB

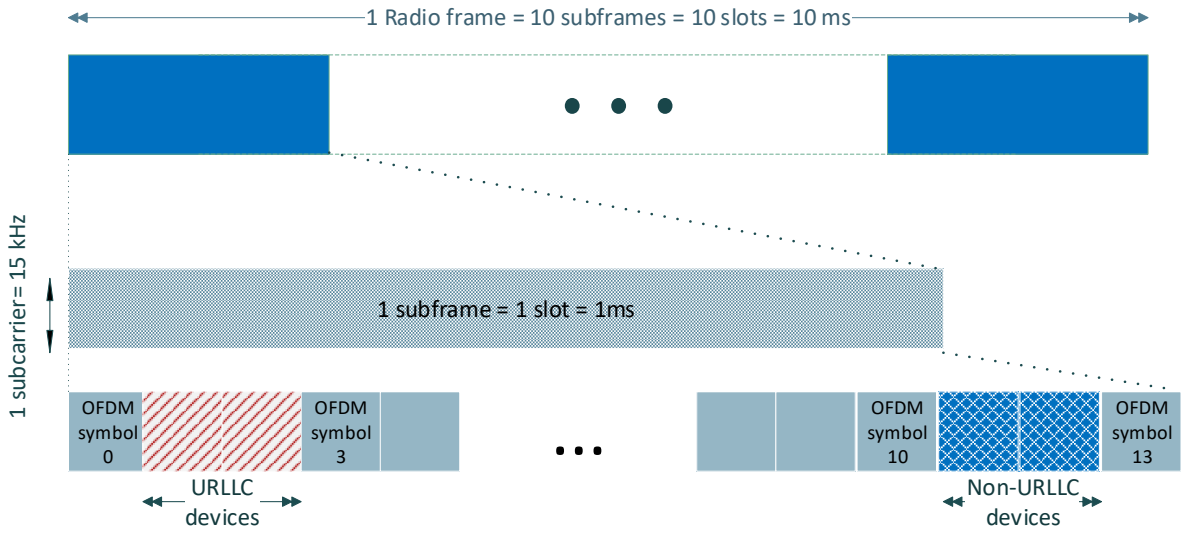


Figure A.2: Illustration of the NR frame structure for $\mu = 0$ and OFDM symbol allocation in the second proposed initial access scheme.

based on the uplink radio resources assigned by the RAR message. Msg3 includes the device temporary C-RNTI which is used for contention resolution in the fourth step.

- *Step 4 (Msg4): RRC connection response.* Devices receive the RRC setup message from the eNB. Only the devices which have their transmitted and received identities matched in Msg3 and Msg4 declare their RA procedure to be successful. After this step, the four-step handshake procedure for initial access is complete. Then devices and eNB perform data transmissions based on the C-RNTI of each device.

In case that there is more than one device transmitting the same preamble, a collision occurs and the competing devices may not receive the corresponding RAR message. If any step in one of the four handshake steps fails[2] the involved device will wait for a random backoff period from a window of size w_{BO} and repeat the RA process by retransmitting an RA preamble. The maximum number of transmissions allowed is limited by a given number, n_{PT}.

In 5G NR, the initial access procedure between a device and its associated gNB is similar to the one employed in LTE-A when operating in the sub-6 GHz frequency range, often referred to as frequency range 1 (FR1). For frequency range 2 (FR2), which includes frequency bands from 24.25 GHz to 52.6 GHz, the initial access involves procedures for cell search and synchronization using beam sweeping [20] [20]. However, to study these physical layer details is beyond the scope of this paper.

A.3.2 5G NR Frame Structure and Numerologies

NR introduces novel scalable numerology and frame structure with the aim of facilitating the expected capacity and latency requirements in 5G. In contrast to the 15 kHz only

[2]An unsuccessful message transmission may also occur due to channel impairments for uplink and/or downlink. This effect is partially reflected in the message error probability expression presented later in Sec.A.6.

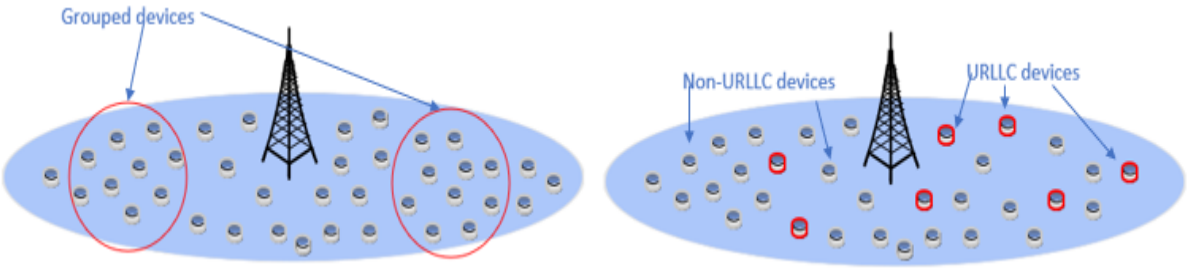


Figure A.3: (a) Scenario 1: Location-bounded URLLC devices versus (b) Scenario 2: Location-spread URLLC devices.

option in LTE/LTE-A, NR supports multiple subcarrier spacing. NR defines 15 kHz as a baseline and introduces 5 numerologies based on subcarrier spacing $\Delta f = 2^{\mu} * 15$ kHz where $\mu = 0, 1, ..., 4$ is the numerology index [20]. The radio frame duration in NR is the same as in LTE/LTE-A, i.e., 10 ms, and one *frame* consists of 10 *subframes* each with 1 ms duration, as shown in Fig. A.2. Moreover, one NR subframe may have one or more *slots* based on the numerology index. For $\mu = 3$ and $\mu = 4$ which are used in our study, the number of slots per subframe would be 8 and 16, respectively. With the increased subcarrier spacing and a larger value of μ, the slot duration reduces according to $1/2^{\mu}$ ms. When $\mu = 3$ and $\mu = 4$, the slot duration would be 125 μs and 62.5 μs respectively. Furthermore, each slot contains 14 (or 12 for extended cyclic prefix (CP)) OFDM *symbols*. However, not all numerologies are applicable to any type of physical channels. Instead, a specific numerology is used only for a given type of physical channels. For more details about NR numerology, refer to [20] [31].

A.3.3 A 3GPP Model for Bursty Traffic

A bursty traffic arrival process occurs when a large number of IoT devices attempt to access the same network simultaneously during a short period of time. This is especially observable under mMTC scenarios where the number of devices could be huge. In [2], 3GPP recommends applying a Beta distribution based arrival process to model the arrival intensity during bursty traffic arrivals, shown as follows.

$$A(i) = L \int_{t_i}^{t_{i+1}} p(t)dt, \tag{A.1}$$

where $A(i)$ represents the access intensity for a total number of L devices contending in an RA slot i between time t_i and t_{i+1}. In (A.1), $p(t) = (t^{\alpha-1}(T-t)^{\beta-1})/\ (T^{\alpha+\beta-1}Beta(\alpha,\beta))$ where $Beta(\alpha,\beta)$ is the Beta function with $\alpha = 3$ and $\beta = 4$. T is the total observation time for traffic arrivals [2].

As an example, we illustrate in Fig. A.4 the numbers of initial arrivals, initial arrivals plus retransmissions, and successful detections within an RA slot under a traffic burst of 10 sec based on 30k devices and 54 preambles [44]. It is clear that the actual number of arrivals consisting of both initial arrivals and retransmissions is much higher than the initial arrivals itself. With such bursty traffic arrivals, the number of devices competing for access in an RA slot is unusually high and providing URLLC services in such a scenario

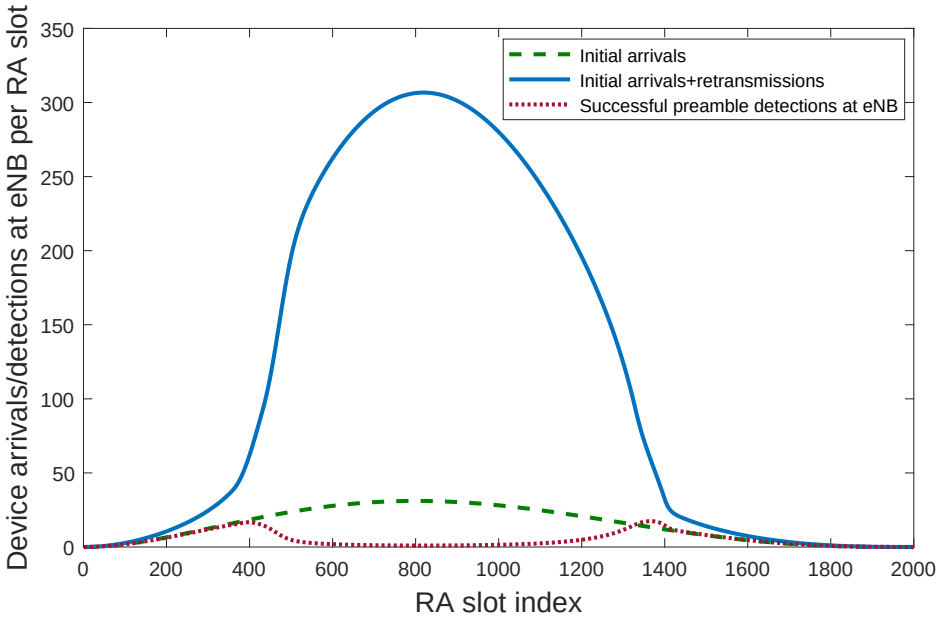


Figure A.4: Number of initial arrivals, retransmissions, and detections in LTE-A random access for 30k devices with 54 preambles following a bursty arrival process [44].

is a challenging task since GF based access schemes which were discussed in Subsec. A.2.2 above would result in high collisions. For this reason, the proposed schemes in this paper focus on grant-based initial access instead of GF transmissions and radio resource allocation. Later on in Subsec. A.7.6, we provide a brief comparison of our schemes versus a GF scheme.

A.4 Network Scenarios and Assumptions

The envisaged network scenarios in this work are inspired by the futuristic cyber-physical mIoT applications recently presented by 3GPP in [10]. In many such applications, devices are battery powered with power saving mode enabled. Upon the occurrence of a mission critical event, for instance, it is likely that many devices will require initial access at almost the same time leading to a traffic burst as presented above.

In this study, we consider that all devices are covered by one cell although some of them may lie comparatively far away from the eNB and that proper preamble formats are allocated to all RA slots [40]. For each scenario, there are L number of IoT devices within the coverage area of an eNB or a gNB[3] and ϕ number of preambles that can be allocated to this cell in a given RA slot. The number of orthogonal preambles that can be allocated in a given RA slot depends on the cell coverage [39]. According to [7] [40], there are 64 preambles that can be allocated in a cell with a coverage radius of 7.4 km and a delay spread of 6 μs and these preambles are designed to be orthogonal to each other.

Scenario 1: Location-bounded URLLC Devices. Although a large number of mMTC devices are deployed across a cell, a set of devices in the immediate vicinity of a point of interest are monitoring the same natural or physical phenomenon, e.g., for process automation within a service area of 100 m $\times$ 100 m as given in [10].

[3]For the rest of this paper, the abbreviations eNB and gNB are interchangeably used as both LTE/LTE-A and 5G NR Phase 1 follow the same procedure for initial access.

In this scenario, we categorize the total population of L IoT devices into γL grouped devices (GDs) and $(1-\gamma)L$ non-grouped devices (NGDs) where γ is a scalar with $0 < \gamma < 1$. The traffic generated by IoT devices could be deterministic periodic, deterministic aperiodic, or non-deterministic [10]. In this work, we focus on a case where devices abruptly require uplink access after sensing an event triggered in a *non-deterministic* and bursty manner, thus representing a worst case scenario among the aforementioned traffic types. Accordingly, the GDs require URLLC access while NGDs still generate traffic but without demanding URLLC services. Although semi-persistent scheduling for URLLC access is another option, it may not guarantee the required performance due to the stringent delay requirements especially when the number of URLLC devices is huge. Furthermore, maintaining semi-persistent scheduling for a massive number of devices is rather difficult and costly as mMTC traffic is often sporadic. For this reason, we propose to reserve merely a small amount of resources (preambles) for grouped devices and obtain the necessary amount of uplink resources for all other group devices through group leader's communications with the gNB (to be clarified in the next section).

Furthermore, we assume that device grouping including group leader selection is performed beforehand based on a specific criterion, e.g., the functionality or geographic proximity of the IoT devices. Device grouping is reconfigurable, however, over a comparatively long period much longer than a SIB2 cycle. A triggering event would be detected by all IoT devices in the same group including the group leader. All the GDs that sensed the triggering event need their measurements to be transmitted to the gNB as each device may report a different facet of the same event. Once a preamble is received, the gNB is assumed to have enough radio resources to allocate to all these grouped devices.

The rationale of the above assumption is as follows. Although the amount of available physical downlink (PDCCH) resources is always limited in reality, the flexibility provided by NR enables the use of more PDCCH resources compared with that of LTE-A. Based on the NR numerology and frame structure presented above and the flexibility provided for PDCCH scheduling [20], more downlink control information (in terms of both information volume and broadcast interval) can be transmitted via PDCCH within a given 5G NR subframe compared with what is possible in LTE-A. Moreover when considering the privilege of URLLC traffic, it is common in the literature to employ techniques such as preemptive scheduling which provides immediate downlink resources to URLLC traffic by overriding parts of already assigned resources for eMBB or another type of lower priority traffic. Such a mechanism is justifiable considering the stringent latency requirements of URLLC devices. Accordingly, we may introduce a potential solution which combines preemptive scheduling with the NR frame structure to accommodate extra PDCCH resources to URLLC traffic. In this way, resource constraint which might appear as a bottleneck to complete the initial access procedure could be abbreviated.

Scenario 2: Location-spread URLLC Devices. Consider another scenario where the IoT devices that require URLLC services are not confined to certain areas within the coverage but could be spread anywhere across the cell. The devices in this scenario could be process monitoring devices which are *static* or mobile robots which are *non-static* [10]. Among these L devices, a certain portion, i.e., ηL where η is a scalar with $0 < \eta < 1$, of devices are considered to require URLLC services whereas the remaining $(1-\eta)L$ devices

do not have such a requirement. Hereafter, these two categories of IoT devices are denoted as URLLC device (UD) and non-URLLC device (NUD), respectively.

Further Clarification: Different from GDs in Scenario 1 which are restricted to certain small areas, UDs in Scenario 2 could be distributed geographically throughout the cell. During the bursty traffic arrival duration, all these devices are considered to be active, i.e., having at least one packet to transmit. The portions of devices which belong to GDs or UDs, i.e., γ and η, are determined by the eNB as a compromise of performance (collision probability, delay, etc.) and configurable parameters. Since these values are configured periodically and the gNB needs to inform all devices about any update, extra signaling overhead is expected. However, to study such extra overhead is beyond the scope of this paper.

Furthermore, in both scenarios, a single frequency band is considered. For NR frame structure based initial access scheme design, the parameter configurations and assumptions including numerologies, PRACH selection, and slot scheduling will be explained in the next section.

A.5 Proposed Initial Access Schemes

Based on the scenarios presented above, we propose three schemes for initial access of mMTC devices. While the first two schemes are tailored to the two scenarios (device grouping with dedicated preambles (DGDP) for scenario 1 and RA-slot based URLLC grouping (RAUG) for scenario 2), respectively, the third one combines the merits of the first two schemes and applies to both scenarios.

A.5.1 Device Grouping with Dedicated Preambles

The main feature of the DGDP scheme is that GDs obtain access privilege to the network through *a contention-free 2-step scheme* [8], as illustrated in Fig. A.1(a) and explained below. Meanwhile, NGDs follow the legacy LTE-A 4-step contention based RA procedure, as shown in Fig. A.1(b). It is expected that a 2-step RACH scheme will bring benefits to channel access in terms of both reduced latency and lower overhead. Although 2-step RACH approaches are presently under discussion within 3GPP, the current draft [46] does not state which type(s) of traffic should apply the 2-step scheme.

A.5.1.1 Access scheme for grouped devices

Consider a single group as an example. At the initial network deployment phase, devices communicate and register themselves with their associated eNB. During the registration process, the eNB collects information about all IoT devices inside the group and their location information to infer the required timing advance details. A *unique and permanent* address, which is different from the C-RNTI mentioned in Sec. A.3, is allocated to each device and the group also receives *a dedicated preamble for uplink communication to be used by the group leader*. The eNB stores these details in a database for further references.

Furthermore, a group leader is selected by the eNB based on a given criterion, e.g., device battery level, device location, or uplink channel quality among group members. All group members will periodically communicate with the eNB and the updated information will be used for group leader selection in the next period of time. In other words, the group leader could be dynamically changed based on the adopted criterion by the eNB and newly collected information from group members. To tackle a rare case where the group leader's preamble transmission fails, e.g., due to uplink channel impairment, the eNB also assigns a backup group leader. A backup leader may also initiate a preamble transmission if necessary. The coordination between a serving group leader and the backup group leader can be performed by various methods with or without the involvement of the gNB. For instance, we can set a timer which expires after a pre-defined period from an event and triggers the backup leader to act as the serving group leader. Alternatively, we can assume an out-of-band D2D communication protocol between the serving leader and the backup leader. However, to design a protocol or procedure for group leader and backup group leader selection is beyond the scope of this paper. In what follows, we explain the 2-step scheme illustrated in Fig. A.1(a).

Step 1 (Msg1): Event triggered dedicated preamble transmission. Once the deployment phase is finished, IoT devices enter into the operational stage. In an event where the observed measurements of IoT devices exceed a pre-defined threshold, a triggering event will be initiated. We assume that the group leader can sense this triggering event and correspondingly it immediately transmits its allocated preamble in the next available RA slot. Other GDs in the same group will not transmit any preamble but they overhear this transmission and wait for the access response from the eNB. In a rare case if the group leader does not sense the triggering event, or the group leader's uplink channel quality is below the required level, the backup group leader will transmit the preamble *after the timeout duration of the access response has elapsed.*

Step 2 (Msg2): Access response from the eNB: When the eNB receives a preamble that is reserved for a specific group, it identifies the group from the preamble. Since each group leader in different groups has its own dedicated preamble, this access process is collision-free. Once the eNB identifies the corresponding group which the received preamble belongs to, it retrieves the information about the registered group members. The eNB is aware of the immediate access requirement of these GDs. *It then allocates resource blocks to individual group members* based on the addresses assigned during the registration process. The eNB transmits the relevant timing advance information for each group member based on the calculations from the registration process so that each member can adjust their transmission time accordingly for radio frame synchronization. Since devices are static, the timing advance values would remain the same unless an update is performed.

A.5.1.2 Access for non-grouped devices

The NGDs inside the same cell follow the legacy LTE-A RA scheme [7] with a 4-step procedure for initial access as explained in Sec. A.3.1. Since n_G preambles are reserved for n_G group leaders, the number of available preambles for NGDs is reduced by n_G (where

$n_G < \phi$), i.e., it becomes $\phi - n_G$. Concurrently, the number of NGDs competing for the $\phi - n_G$ preambles also shrinks to $(1-\gamma)L$. If a collision happens, the collided devices will retransmit their preambles after waiting for a backoff interval based on a random number selected from a uniformly distributed range $[0 \sim w_{BO} - 1]$. For successfully transmitted preambles, Msg3 and Msg4 will be transmitted subsequently to complete the RA process as shown in Fig. A.1(b). In this paper, we do not consider explicitly how a message transmission could be affected by channel impairment for any specific type of channels between the gNB and devices. However, the transmissions of Msg3 and Msg4 are subject to failures as presented in the next section.

As mentioned earlier, the group formation of IoT devices in the DGDP scheme is predefined and the parameters are reconfigurable. While having a higher n_G would enable access for a larger number of grouped devices, the selection of n_G and γ needs to be performed carefully to avoid performance degradation of NGDs. Generally, the number of devices per preamble gives an indication about the possibility of different devices selecting the same preamble and thereby causing collisions. In LTE-A without grouping, this ratio is L/ϕ. In DGDP with n_G number of groups and γL grouped devices, this ratio is given by $(1-\gamma)L/(\phi - n_G)$ for NGDs. In order to improve the performance level that will be achieved by NGDs without grouping, the following condition must hold

$$\frac{(1-\gamma)L}{(\phi - n_G)} < \frac{L}{\phi}. \tag{A.2}$$

Reformulating the above inequality into $(1-\gamma)L\phi < L(\phi - n_G)$, (A.2) can be expressed in a simplified form, as $n_G < \gamma\phi$. This relationship can be utilized when deciding n_G and γ so that the performance of NGDs is not compromised.

A.5.2 RA-slot based URLLC Grouping

Consider now an mIoT cell as presented earlier in Scenario 2 where the number of IoT devices that require URLLC services could be potentially large and their locations may spread across the cell. In this case, it is prohibitive to assign many dedicated preambles to these UDs as we did in DGDP since the total number of preambles in cell, i.e., ϕ, is very small. In what follows, we propose another scheme, RAUG, which grants access privilege to certain devices *without assigning dedicated preambles*. This scheme is designed largely based on the NR frame structure and numerology outlined in Subsec. A.3.2.

A.5.2.1 The principle of RAUG

In RAUG, all devices follow the 4-step RA initial access procedure but separate RA slot resources are assigned to URLLC and non-URLLC preamble transmissions respectively. As depicted in Fig. A.2, each subframe provides RA opportunities and dedicated RA slots are reserved for UDs in order to provide them with URLLC access. As mentioned in Sec. A.4, only a portion of IoT devices, i.e., ηL of them, will have URLLC requirements during a given period of time. Note that although it is possible to form groups with very small URLLC device population, very little benefit would be observed if the group size is too small considering the scarcity of the number of preambles. Accordingly, each

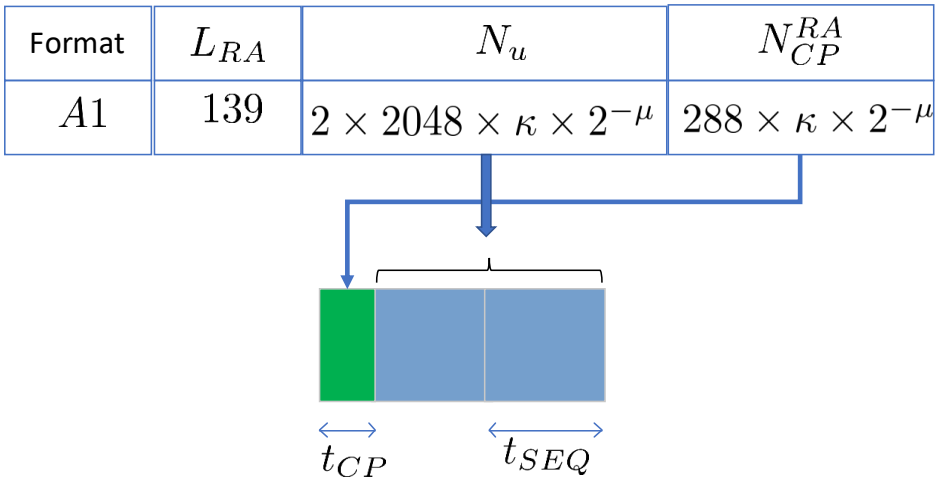


Figure A.5: Illustration of the format of preamble type A1.

particular device will transmit its preamble only in the assigned RA slot for UDs that is broadcast by the gNB beforehand and periodically, e.g., via the SIB2 message.

Different from ACB [2], RAUG does not assign any probabilities for any type of devices to transmit their preambles. In other words, *both UDs and NUDs have equal opportunity when competing for network access, however, through dedicated RA slots assigned inside a 5G NR subframe.* Although having a dedicated RA slot for URLLC devices significantly increases access probability, the time interval between two consecutive RA slots for UDs needs to be minimized in order to reduce latency. Distinct from the slotted access schemes presented in [2] where low latency is not a priority concern, the RAUG scheme utilizes the 5G NR frame structure and numerology concept for the purpose of latency reduction.

A.5.2.2 Frame format in RAUG

To demonstrate the concept of RAUG, we use numerology $\mu = 0$ as an example. It corresponds to the 15 kHz subcarrier spacing and each subframe in the radio frame structure consists of a single slot. Among the 13 preamble formats available in NR, a short sequence can be used for numerology $\mu = 0$ [20]. Fig. A.5 illustrates the preamble format adopted in this study, known as A1, and the values mentioned therein will be used to calculate the preamble duration. The value of L_{RA}, which is the preamble sequence length, is related to the short sequence while N_u and N_{CP}^{RA} provide the total sequence length and the CP length of the preamble in samples respectively. To convert them into seconds, we need to multiply the given values by $T_c = 0.509 \times 10^{-6}$ ms where T_c denotes the basic time unit in NR.

Denote by term κ the ratio between the basic time unit of LTE/LTE-A (T_s) and T_c. According to 3GPP [20], it ends up with $\kappa = 64$. Based on Fig. A.5, the total duration of preamble format A1 is equal to $t_{cp} + 2 \times t_{seq}$ where $t_{cp} = 288\kappa \times 2^{-\mu} = (288 \times 64 \times 0.509 \times 10^{-6}) = 0.0094$ ms and $t_{seq} = 2048\kappa \times 2^{-\mu} = 0.0667$ ms. Hence, the total duration can be calculated as $t_{cp} + 2 \times t_{seq} = 0.0094 + 2 \times 0.0667 = 142.8$ μs. Note that this duration is similar to the time duration of two OFDM slots in $\mu = 0$ and hence, the preamble can be transmitted using two OFDM slots including CP. Similarly we adopt index 106 mentioned in Table 6.3.3.2-2 in [20] for the PRACH configuration in this study.

As such, it is possible to transmit a PRACH preamble in every subframe.

In order to provide priority to devices with low latency requirements, we introduce an option to allocate two RA slots inside a given subframe for initial access. This is possible for specific types of preamble formats available under a given numerology that satisfies the preamble length and OFDM symbol duration requirements mentioned above. Further details regarding these formats can be found in Table 6.3.3.1-2 of [20]. Accordingly, their duration can be calculated similar to the aforementioned calculation. Hence, considering the above configuration by having two RA slots inside a slot (one slot equals to one subframe for $\mu = 0$), both UDs and NUDs obtain an opportunity for an initial access attempt in every slot. Table 11.1.1-1 in [3] defines which symbols could be allocated for uplink and downlink transmissions. However, different from the legacy initial access procedure, the RA slot OFDM symbols in RAUG are different for UDs and NUDs. Correspondingly, *both types of IoT devices can share the same set of preambles in the same subframe, however, in different RA slots*. Furthermore, once the gNB receives a set of preambles in the URLLC RA slot, it treats these requests with higher priority. Hence, the required timing for the transmission of remaining messages is reduced.

Further discussions on the distinctions between RAUG and DGDP: Firstly, no prior grouping based on service types or device location is involved when deciding UDs in RAUG. UDs could be deployed in any location inside a cell and do not have to share any common application with their neighboring devices. Furthermore, each UD could perform its individual task supporting a specific application. Secondly, unlike GDs, UDs need to transmit the preambles themselves and compete with other UDs for initial access. Thirdly, UDs do not necessarily need to be static in deployment whereas GDs are considered to be static for timing advance synchronization purposes needed in the 2-step initial access procedure. However, different from the legacy RA scheme, UDs do not need to compete with NUDs since they have their separate RA slots to transmit the selected preambles. This would ensure better access opportunities for UDs in comparison with devices in the legacy scheme. Furthermore, unlike NGDs in DGDP which compete for $\phi - n_G$ preambles, NUDs in RAUG have all ϕ available preambles for access competition in their allocated RA slot.

A.5.3 Hybrid Scheme (HS)

While DGDP is designed for providing URLLC services for a specific set of GDs, it cannot be applied to a large number of IoT devices with such requirements. RAUG releases this constraint by providing *high reliability and low latency access* for a potentially much larger number of UDs inside a cell regardless of their locations. However, since RAUG follows a 4-step contention based RA procedure, the achieved reliability and latency could be lower than what is obtained in DGDP. In this subsection, we propose a hybrid scheme which combines the merits of the other two schemes proposed above.

More specifically, HS is a combined access scheme in which both device based grouping and slot based allocation apply. In this scheme, we still have GDs and NGDs but NGDs are further categorized into UDs and NUDs. UDs will use the first RA slot to transmit its preambles but still follow a contention based procedure. GDs and NUDs will use the

Table A.1: Main features of the three proposed schemes.

	DGDP	RAUG	HS
Type of devices	GDs, NGDs	UDs, NUDs	GDs, UDs, NUDs
Pre-grouping of devices	Yes	No	Yes
RA slot based grouping	No	Yes	Yes
URLLC enabled for	GDs only	UD only	GDs, UDs
Guaranteed reliability	for GDs	No	for GDs

second RA slot inside the same subframe, however, GDs still have dedicated preambles. In this way, *GDs and UDs can share the same preambles but in different slots*. Hence, a larger number of IoT devices with URLLC requirements can be accommodated via GDs and UDs while utilizing the benefits of having multiple RA slots inside a subframe.

Accordingly, there will be γL GDs. Among the remaining $(1-\gamma)L$ NGDs, $\eta(1-\gamma)L$ will be UDs and $(1-\eta)(1-\gamma)L$ devices will be NUDs. As a result, $\eta(1-\gamma)L$ UDs will compete for ϕ preambles inside the first RA slot in a subframe whereas $(1-\eta)(1-\gamma)L$ NUDs will compete for $\phi - n_G$ preambles in the second RA slot inside the same subframe.

Moreover, it is worth reiterating that the proposed schemes for IoT device initial access in this paper are targeted at both 4G and 5G NR Phase 1, i.e., OFDMA based networks, and the operation of RAUG and HS relies on the support of NR numerologies. Enabled by the flexibility supported through different numerologies in 5G NR, allocating two RA slots inside one subframe becomes configurable. Meanwhile, reservation of radio resources is also feasible in both 4G and 5G NR. Therefore, to apply the proposed scheme(s) to a specific type of IoT technology, e.g., narrowband IoT (NB-IoT), proper parameter tuning based on the corresponding physical layer specifications is required. In Table A.1, we summarize the main features of the three proposed initial access schemes.

A.6 Performance Analysis

In this section, the performance of the proposed schemes is analyzed. Recall that a contention-free 2-step procedure applies to GDs whereas the other types of IoT devices, i.e., NGDs, UDs, and NUDs follow a contention based 4-step procedure however *with different number of preambles and different number of device arrivals for each type of devices*. Therefore, the same analytical model applies to these three types of devices. In Table A.2, we summarize the number of IoT devices and the number of available preambles per RA slot in each type, denoted as $\hat{L}$ and $\hat{\phi}$ respectively, for our performance evaluation. The main notations, their meanings, and the respective numerical values[4] used in this study are listed in Table A.3.

In the rest of this section, the performance evaluation of GDs is presented first. Then, an analytical model used to evaluate the performance of NGDs, UDs, and NUDs is developed. For performance evaluation, three metrics which are recommended by 3GPP [2], i.e., preamble collision probability, access success probability, and average delay for successful transmissions, are selected as our performance metrics.

[4]In Table A.3, the numbers inside () corresponded to values used by UDs.

Table A.2: $\hat{L}$ and $\hat{\phi}$ values for different type of devices in the three proposed schemes.

	Initial Access Scheme						
	DGDP		RAUG		HS		
	GDs	NGDs	UDs	NUDs	GDs	UDs	NUDs
$\hat{L}$	γL	$(1-\gamma)L$	ηL	$(1-\eta)L$	γL	$\eta(1-\gamma)L$	$(1-\eta)(1-\gamma)L$
$\hat{\phi}$	n_G	$\phi - n_G$	ϕ	ϕ	n_G	ϕ	$\phi - n_G$

Table A.3: Notations, explanations, and values [2] [8].

Notation	Explanation	Value
t_{AP}	Duration of an arrival period (in terms of subframes).	10000
L	Total number of devices in a cell which request service during t_{AP}	10000-300000
w_{BO}	Backoff window size (in terms of subframes)	21, (1)
t_{RAS}	Interval between two successive RA slots (in terms of subframes). The t_{RAS} value in RAUG is 8 OFDM symbols (Refer to Fig. 2)	5, 1
ϕ	Total number of preambles in an RA slot available for access competition	54
n_{PT}	Maximum number of preamble transmissions	10
w_{RAR}	Length of the RA response window (in terms of subframes)	5, (2)
p_j	Preamble detection probability of the j^{th} preamble transmission	$p_j = 1 - \frac{1}{e^j}$
p_f	HARQ retransmission probability for Msg3 and Msg4	0.1
n_{HARQ}	Maximum number of HARQ transmissions for Msg3 and Msg4	5
t_{HARQ}	Time interval required for receiving HARQ ACK (in terms of subframes)	4, (1)
t_{RQ}	Gap of Msg 3 retransmission	4, (1)
t_{RAR}	Processing time required by the eNB to detect transmitted preambles (in terms of subframes)	2, (1)
n_G	Number of groups	5, 10, 15
γ	Portion of devices from L that are grouped	0.1, 0.2, 0.3
η	Portion of devices from L that require URLLC services	0.1, 0.2, 0.3, 0.5
n_{UL}	Maximum number of devices acknowledged within an RA response window	15
t_D	Delay from a preamble transmission to the reception of the RAR response	w_{RAR}+t_{RAR}
μ	5G NR subcarrier spacing configuration numerology	0 - 4

A.6.1 Performance of GDs

Since each group has its dedicated preamble reserved for GDs, the access process for GDs is contention-free. Hence, the probability of occurring a preamble collision at the eNB is 0. However, although there is no preamble collision, there is no guarantee that the preamble will be successfully received considering the effect of channel impairments. This

is represented by the preamble detection probability P_j at the eNB for the j^{th} preamble transmission of the group leader. The value of P_j is calculated based on $P_j = (1-e^{-j})$, as recommended by 3GPP [2], and it monotonically increases as more transmission attempts are conducted. Although the detection probability is not high enough after the first few attempts, it reaches the value of $P_j > 0.9999$ when $j = n_{PT} = 10$. Accordingly, we claim that the access success probability for GDs will be 1 even in the worst case given that up to $n_{PT} - 1$ retransmissions can be performed.

For detecting a preamble successfully, at least one transmission attempt is required from the group leader. Whether a retransmission is needed or not depends on the detection status of the previous transmission, up to $n_{PT} - 1$ times. Let $s(j)$ be the probability of success after the j^{th} preamble transmission and it is given by $s(j) = (1 - P_1)(1 - P_2) \cdots (1 - P_{j-1})P_j$. This expression is equivalent to the probability mass function of success at the j^{th} preamble transmission. Therefore, the expected value of the number of preamble transmissions required for a successful detection can be obtained by $\sum_{j=1}^{n_{PT}} js(j)$. After a t_D duration from a successful preamble transmission, the group members receive Msg2 from the eNB with the granted access and allocated radio resources, as shown in Fig. A.1(a). Correspondingly, the group leader will wait for a duration of t_D before initiating a retransmission attempt. Therefore, considering the number of required retransmissions, the average delay for successfully transmitting a preamble and receiving the corresponding Msg2, denoted as D_a, can be calculated as follows

$$D_a = t_D \sum_{j=1}^{n_{PT}} js(j) = t_D \sum_{j=1}^{n_{PT}} \left(jP_j \prod_{k=1}^{j-1} (1 - P_k) \right). \tag{A.3}$$

To be more precise, the access delay for grouped devices would be slightly different from the access delay of their group leader if other factors such as the location of devices and extra cost for intra-group communications are included in this calculation. For analysis simplicity, we do not consider additional delay occurred for intra-group communications. Instead, the delay obtained in (A.3) is considered as an representative value since grouped devices are normally deployed in relatively close proximity to their group leader.

A.6.2 Performance of NGDs, UDs, and NUDs

A.6.2.1 Modeling the initial access procedure

Consider a burst of initial traffic arrivals for the duration of t_{AP}. Fig. A.6 illustrates the timing diagram with RA slots and arrivals. As explained earlier, the initial access procedure for NGDs, UDs, and NUDs follows the legacy RA process. Hence, a common analytical model is adopted as the baseline for analyzing these three types of devices. Based on a comprehensive analytical model proposed in [8] which provides sufficiently high accuracy for LTE-A RA processes, we present below our analysis tailored for performance evaluation of mMTC networks consisting of four types of devices according to the envisaged scenarios and the proposed schemes.

Initial arrivals: The average number of device arrivals at the i^{th} RA slot is calculated

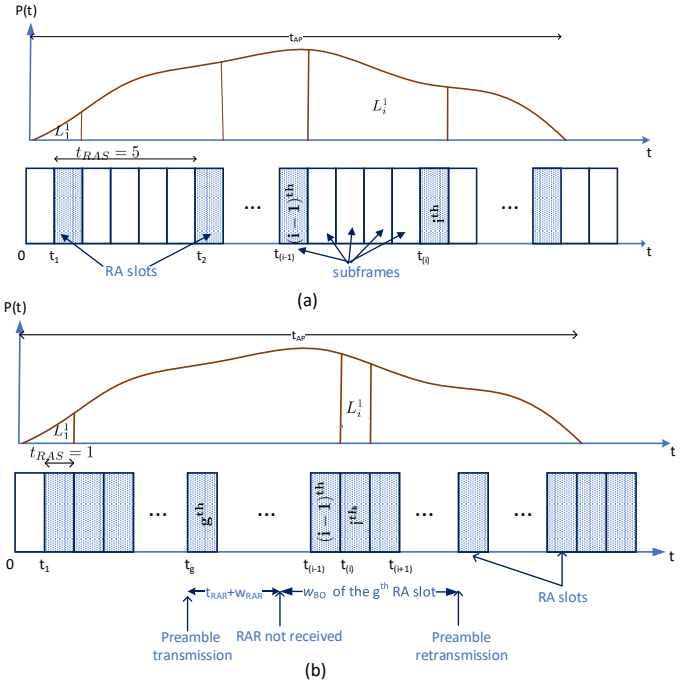


Figure A.6: Timing diagram denoting RA slots, initial bursty arrivals per slot and the related timing parameters (a) $t_{RAS} = 5$ (b) $t_{RAS} = 1$.

by the following equation

$$L_i^1 = \hat{L} \int_{t_{i-1}+1}^{t_i+1} p(t)\, dt, \tag{A.4}$$

where $p(t)$ is based on Beta distribution and t_i is the starting time of the i^{th} RA slot as explained in Sec. A.3. The superscript of L_i^1 represents the initial arrival, i.e., $j = 1$. Term $\hat{L}$ in (A.4) denotes the total number of IoT devices based on each device type and the adopted access scheme, as illustrated in Table A.2. Accordingly, the initial access device intensity at a given RA slot, L_i^1, is the integration of the number of new device arrivals between the end points of the previous and current RA slots.

Retransmissions: For a given RA slot i, in addition to the initial arrivals, there would be IoT devices attempting their j^{th} preamble transmissions ($1 < j \leq n_{PT}$) due to previously failed $(j-1)^{th}$ preamble transmissions at the g^{th} RA slot. The positions of the g^{th} and i^{th} RA slots are demonstrated in Fig. A.6. The number of IoT devices performing their j^{th} preamble transmission on the i^{th} RA slot, denoted by L_i^j, is calculated as follows

$$L_i^j = \sum_{g=G_{\min}}^{G_{\max}} \alpha_{g,i} L_{g,F}^{j-1}, \tag{A.5}$$

where G_{min} and G_{max} denote respectively the lower and upper limit of the window of the RA slot values that g could take. That is, in order to transmit the j^{th} transmission on the i^{th} RA slot, the $(j-1)^{th}$ transmission failure should occur between G_{min} and G_{max} time before t_i. $\alpha_{g,i}$ denotes the percentage of the backoff interval of the g^{th} RA slot that overlaps with the transmission interval of the i^{th} RA slot. The G_{min}, G_{max}, and $\alpha_{g,i}$ values are calculated as follows [8], $G_{\min} = (i-1) - \frac{t_D + w_{BO} - 1}{t_{RAS}}$, $G_{\max} = i - \frac{t_D + 1}{t_{RAS}}$.

$$\alpha_{g,i} = \begin{cases} \frac{t_g+t_D+w_{BO}-t_{i-1}}{w_{BO}}, & \text{if} \quad G_{min} \leq g \leq i - \frac{t_D+w_{BO}}{t_{RAS}}; \\ \frac{t_{RAS}}{w_{BO}}, & \text{if } i - \frac{t_D+w_{BO}}{t_{RAS}} < g < (i-1) - \frac{t_D}{t_{RAS}}; \\ \frac{t_i-(t_g+t_D)}{w_{BO}}, & \text{if } (i-1) - \frac{t_D}{t_{RAS}} \leq g \leq G_{max}; \\ 0, \text{ otherwise.} \end{cases}$$

Furthermore, the number of IoT devices that failed their j^{th} preamble transmission at the i^{th} RA slot, $L^j_{i,F}$, can be calculated from the relationship $L^j_i = L^j_{i,S} + L^j_{i,F}$, where

$$L^j_{i,S} = \begin{cases} L^j_i e^{-\frac{L_i}{\phi - n_G}} p_n, & \text{if } \sum_{j=1}^{n_{PT}} L^j_i e^{-\frac{L_i}{\phi - n_G}} p_n \leq n_{UL}; \\ \frac{L^j_i e^{-\frac{L_i}{\phi - n_G}} p_n n_{UL}}{\sum_{j=1}^{n_{PT}} L^j_i e^{-\frac{L_i}{\phi - n_G}} p_j}, & \text{otherwise.} \end{cases} \tag{A.6}$$

Here, $L_i = \sum_{j=1}^{n_{PT}} L^j_i$. Note that, even if the preamble transmission is performed without collision, there is no guarantee on the successful reception of the RA response due to channel impairments as discussed above and the constraint on the maximum number of IoT devices that would be acknowledged within an RA response window, denoted by n_{UL}. Hereafter, term $L^{j-1}_{g,F}$ in (A.5) can be calculated accordingly.

As mentioned earlier, the transmissions of Msg3 and Msg4 may not always be successful due to channel impairments. A message transmission is considered to be failed if the transmission of Msg3 or MSg4 exceeds n_{HARQ} times. Accordingly, we calculate the error probability of message transmission, $P_{e,MSG}$, including the hybrid automatic repeat request (HARQ) process as follows,

$$P_{e,MSG} = p_f^{n_{HARQ}} + \sum_{k=0}^{n_{HARQ}-1} p_f^k (1-p_f) p_f^{n_{HARQ}}. \tag{A.7}$$

A.6.2.2 performance metrics

Using the outcome from the above modeling, we are able to obtain the number of initial arrivals and retransmissions at each RA slot as well as the number of successful and failed devices at each RA slot. Based on this information, closed-form expressions for the three performance parameters of interest are obtained as follows.

Collision probability, denoted as P_c, is the ratio between the number of collided preambles and the total number of preambles transmitted. As the number of collided preambles equals to the total number of preambles minus the number of successful and idle preambles, P_c is obtained as follows

$$P_c = \frac{\sum_{i=1}^{I_R} \left(\hat{\phi} - L_i e^{-\frac{L_i}{\hat{\phi}}} - \hat{\phi} e^{-\frac{L_i}{\hat{\phi}}} \right)}{I_R \hat{\phi}} = \frac{\sum_{i=1}^{I_R} \left(\hat{\phi} - e^{-\frac{L_i}{\hat{\phi}}} \left(L_i + \hat{\phi} \right) \right)}{I_R \hat{\phi}}. \tag{A.8}$$

In (A.8), term I_R denotes the number of RA slots inside the observation time duration. Term $\hat{\phi}$ denotes the total number of preambles available for each type of IoT devices under a specific asccess scheme, as explained in Table A.2.

Access success probability, denoted by P_s, is the probability that an IoT device successfully completes the RA procedure within n_{PT} transmission attempts. That is, P_s = (total number of successfully accessed devices) / (total number of devices arrived in t_{AP}), as given in (A.9). Note that an access success means not only a successful preamble transmission but also the completion of all four steps in the RA procedure. Therefore, the number of successfully accessed devices that transmit the j^{th} preamble within the i^{th} RA slot is equal to $L^j_{i,S}(1 - P_{e,MSG})$. Considering that the values for $P_{e,MSG}$ are negligibly low in reality, P_s can be expressed and estimated as follows,

$$P_s = \frac{\sum_{i=1}^{I_R}\sum_{j=1}^{n_{PT}} L^j_{i,S}(1 - P_{e,MSG})}{\hat{L}} \approx \frac{\sum_{i=1}^{I_R}\sum_{j=1}^{n_{PT}} L^j_{i,S}}{\hat{L}}. \tag{A.9}$$

Average delay for successful devices, denoted by D'_a, equals to the accumulated access delay experienced by those devices which experience successful access divided by the total number of successfully accessed devices. It is given by

$$D'_a = \frac{\sum_{i=1}^{I_R}\sum_{j=1}^{n_{PT}} L^j_{i,S} T_n}{\sum_{i=1}^{I_R}\sum_{j=1}^{n_{PT}} L^j_{i,S}}, \tag{A.10}$$

where T_n is the average access delay of a successfully accessed device that performs exactly n preamble transmissions.

Moreover, it is well understood that backoff mechanisms may lead to long delays and induce heavy-tailed delay distributions, especially when the number of competing devices is large. In our schemes, however, the number of preamble transmissions is strictly bounded by a parameter, n_{PT}. Therefore, the time an RA request can wait for access is also bounded by this constraint.

A.7 Numerical Results and Discussions

This section presents the numerical results obtained from both the analytical model and simulations for an mMTC cell with its parameters configured as listed in Table A.3. The analytical results are obtained following the model presented in Sec. A.6. For simulations, we develop a program in MATLAB which mimics the behavior of the proposed schemes as well as the baseline scheme for LTE-A based initial access and the GF transmission scheme. The results reported in this section are the average values obtained from a large number of simulation runs for all considered schemes. For traffic arrivals, the Beta distribution based arrival intensity function expressed in (A.1) is adopted. The performance of the studied schemes is evaluated by configuring ϕ, γ, η, and n_G to certain values according to Table A.3 while varying the number of IoT devices, i.e., L, in each case. More specific configuration details will be elaborated when presenting the performance under each scenario. Consequently, each configuration will in turn affect the $\hat{L}$ and $\hat{\phi}$ values in each scheme, as explained in Table A.2. In order to reflect bursty traffic in *a massive*

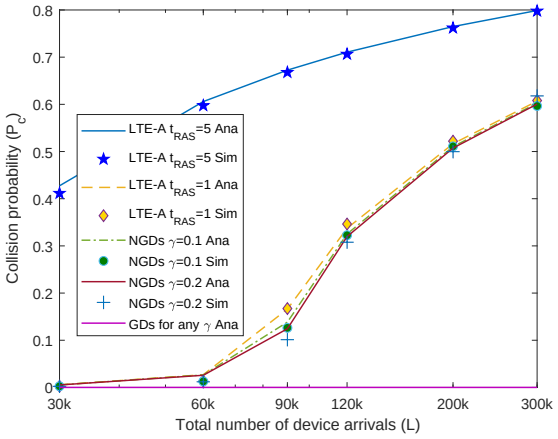


Figure A.7: Collision probability in DGDP: GDs versus NGDs.

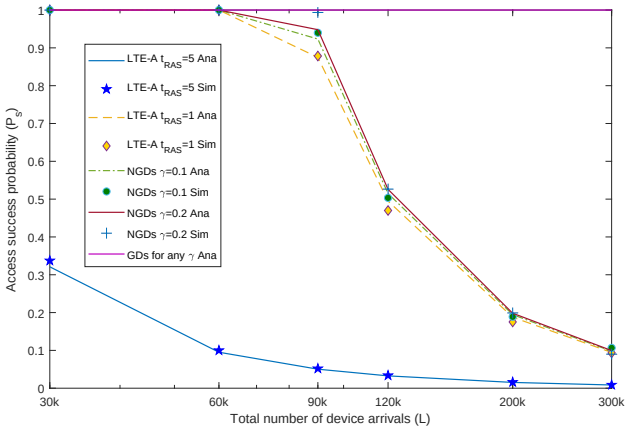


Figure A.8: Access success probability in DGDP: GDs versus NGDs.

MTC network, we let L vary from 30k up to 300k which is 10 times as large as what was typically considered in early studies, e.g., in [8] which considered merely an MTC network with a moderate size.

The performance of the proposed schemes is first evaluated and compared with that of the legacy LTE-A RA scheme. Then the access success probability is compared with that of GF transmission. To perform the comparison, we enable two PRACH configurations by selecting the t_{RAS} value alternatively between 5 and 1. When $t_{RAS} = 5$, the access schemes behave as what is commonly used in LTE-A PRACH [2] [8], i.e, an IoT device gets an initial access opportunity in every fifth subframe. By configuring $t_{RAS} = 1$, which is a feature supported by multiple PRACH configurations in NR and also supported in LTE-A, IoT devices are entitled to transmit their preambles in every subframe. These two initial access options are illustrated in Fig. A.6(a) and Fig. A.6(b), respectively.

A.7.1 DGDP Performance

The performance of the DGDP scheme is evaluated based on the n_G and γ values configured as $\gamma = 0.1, 0.2$ with corresponding $n_G = 5, 10$, respectively. In order to further reduce latency in the 2-step handshake procedure, GDs need faster responses from eNB. Accordingly, $w_{RAR} = 2$ and $t_{RAR} = 1$ are configured for the initial access of UDs.

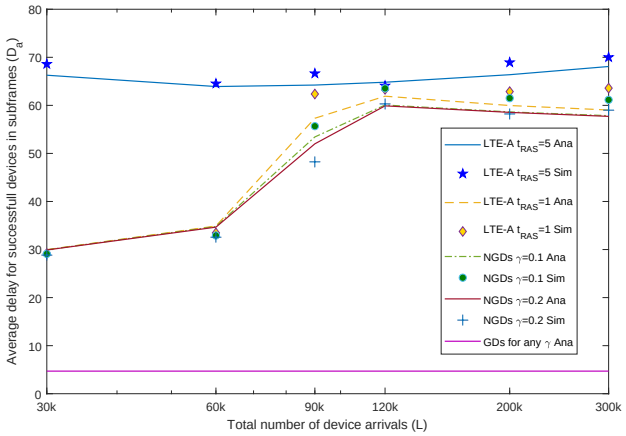


Figure A.9: Average delay of the successfully accessed devices in DGDP.

A.7.1.1 Collision probability and access success probability

As discussed in Sec. A.6, $P_c = 0$ for GDs since the initial access of GDs is contention-free. Furthermore, by allowing up to $n_{PT} - 1$ retransmissions, GDs have guaranteed access even when channel impairment is taken into account, leading to $P_s = 1$. In Fig. A.7 and Fig. A.8, we depict respectively the collision probability and access success probability achieved by DGDP, for both GDs and NGDs, and compare them with the performance of the legacy LTE-A scheme. It is evident that, in addition to the guaranteed performance of GDs, NGDs have also achieved better or much better performance over the legacy scheme for both γ values. The same observation applies to the other figures illustrated later in this section, even though not explicitly highlighted in result discussions.

For NGDs, P_c monotonically increases as the number of IoT devices, L, grows. With a large device population, a higher number of devices will select the same preamble and transmit it in the same RA slot, resulting in collisions. The collided transmissions prompt more retransmissions, leading to further collisions per RA slot. As a result, P_s for NGDs decreases with a larger L. With $\gamma = 0.2$, which means that more IoT devices are grouped in comparison with $\gamma = 0.1$, the performance of both metrics is marginally better. This is due to the fact that, although the number of competing NGDs is reduced with a larger γ, the number of available preamble, $\hat{\phi}$, has also shrunk, leading to limited performance gain. In Subsection A.7.4 below, we will further elaborate this relationship.

A.7.1.2 Average delay for successfully accessed devices

The average delay for the successfully accessed GDs obtained based on (A.3) equals approximately to 5 subframes according to our parameter configuration. This is significantly lower in comparison with the delay that a successful IoT device would experience without grouping, i.e., via LTE-A based access, as presented in Fig. A.9.

Note that the delay behavior of the GDs is governed by (A.3) and it is independent of the number of IoT devices in the group. For NGDs, in all configurations except when $t_{RAS} = 5$ for LTE-A, the average delay of the successfully accessed devices increases up to when there are $L = 120$k devices. Beyond this point, the average delay exhibits a slightly descending trend. This behavior can be explained by referring back to Fig. A.8 which

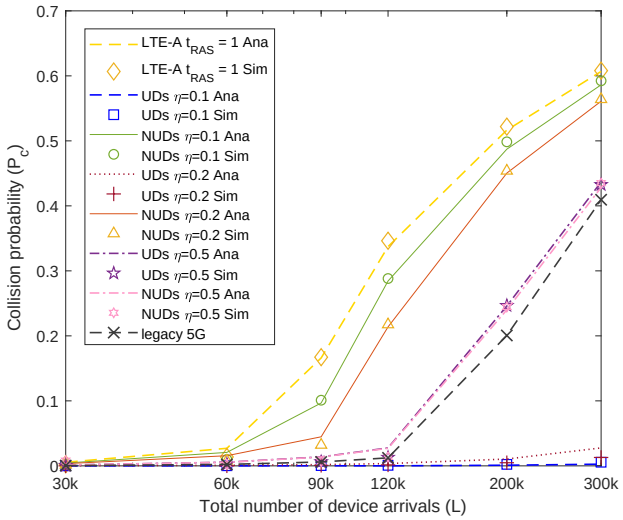


Figure A.10: Collision probability in RAUG: GDs, UDs, versus NUDs.

shows that the P_s values obtained at 200k is approximately 1/3 of the value at 120k. That is, the total number of successful devices is much lower at 200k in comparison with when there are 120k IoT devices. Among these successful ones, transmission successes occur at the initial or final phase of an arrival burst since heavy losses happened during the peak of the burst. In other words, the successful devices have experienced relatively low access delays, leading to a slightly lower average delay.

A.7.2 RAUG Performance

The performance of RAUG needs to be evaluated with respect to UDs and NUDs. As the number of UDs and NUDs depends on the value of η, we evaluate the impact of η on the performance of each type of IoT devices.

As introduced in Sec. A.5, UDs and NUDs transmit their preambles in separate RA slots of the same subframe. This enables eNB to recognize UDs from the arriving RA slot in a subframe and to perform the remaining handshake steps faster. For this purpose, we adopt two different timing values for UDs and NUDs in our network configuration. This is a reasonable approach since UDs require minimum latency. The flexible frame structure in NR with shorter TTI values also enables such a privilege for UDs. Accordingly, the backoff window size w_{BO} is reduced to 1 in order to speed up the retransmission process in case of a transmission failure due to collisions or channel impairment. Furthermore, we configure the w_{RAR} value as $w_{RAR} = 2$ [3]. In addition to LTE-A with $t_{RAS} = 1$, we have considered another scheme that follows the legacy LTE-A access procedure but allows two RA slots within a subframe for the purpose of further comparison. Hereafter this scheme is referred to as *legacy 5G* as this configuration is possible considering the flexibility provided by the 5G NR frame structure. Note however that although RAUG also provides two RA slots per subframe, each type of devices (UD or NUD) has only one RA slot available within one subframe.

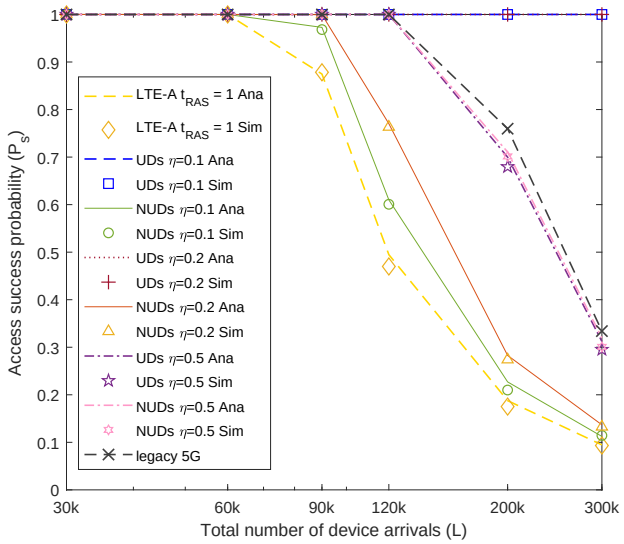


Figure A.11: Access success probability in RAUG: GDs, UDs, vs. NUDs.

A.7.2.1 Collision probability and access success probability

As expected, UDs achieve lower P_c and higher P_s for all ranges of L when $\eta = 0.1, 0.2$, as illustrated in Figs. A.10 and A.11. Having all ϕ preambles available for access competition of a small fraction of L enables such significant improvements. On the other hand, the performance of NUDs deteriorates with larger L values. However, NUDs still exhibit better performance when compared with the baseline scheme and also with NGDs when γ is configured with the same value as η. This comparison will be further discussed in Subsec. A.7.5. For an extreme case with $\eta = 0.5$, the performance of UDs also degrades when $L >$ 120k. However, this configuration will significantly improve the performance of NUDs as the number of NUDs would reduce substantially. In Subsec. A.7.4, the performance tradeoff between UDs and NUDs with respect to the value of η will be further elaborated.

As shown in Figs. A.10 - A.11, the performance of the legacy 5G scheme is similar to that of the UDs and NUDs given that $\eta = 0.5$. Since legacy 5G does not employ device grouping, the number of devices competing for RA slots is twice as many as for UDs and NUDs with $\eta = 0.5$. At the same time, the total amount of available resources for legacy 5G is also doubled for UDs and NUDs with the same η value due to the fact that there are two RA slots within each subframe. Accordingly, the amount of resources used by each device type is half of what is available for legacy 5G. Therefore, the performance of these three schemes is similar based on the given configuration. From these figures, it is clear that the UDs still exhibit better performance when the $\eta = 0.1$ or 0.2 thanks to the concept of having separate resources for URLLC traffic.

A.7.2.2 Average delay for successfully accessed devices

As shown in Fig. A.12, when $\eta = 0.1, 0.2$, the achieved average delay for UDs is approximately 10 subframes and this value keeps comparatively stable regardless of the IoT device population. With a low collision probability as presented above, devices can transmit their preambles successfully with a low number of transmission attempts, resulting in

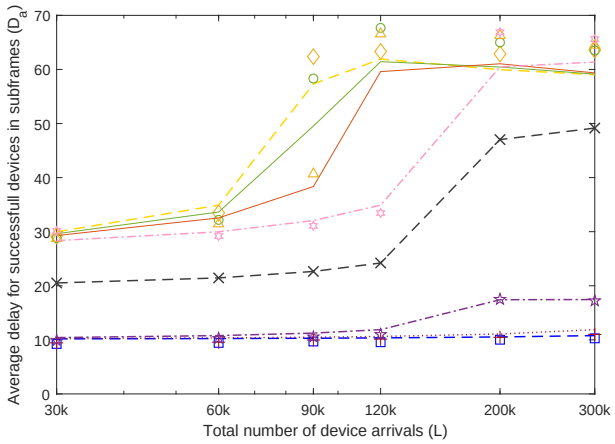


Figure A.12: Average delay of the successfully accessed devices in RAUG (The legend is identical to the ones shown in Fig.A.11).

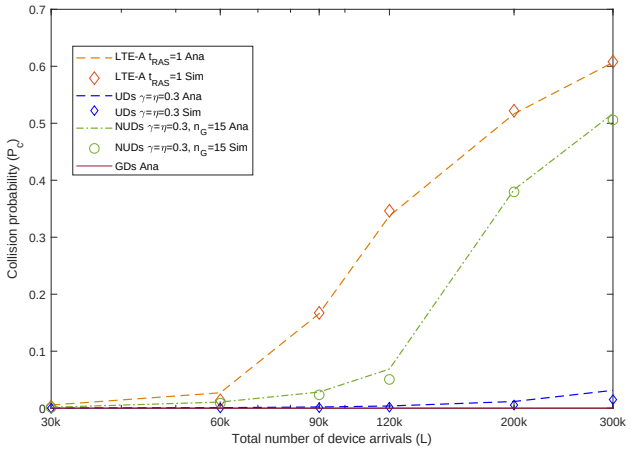


Figure A.13: Collision probability in HS: GDs, UDs, versus NUDs.

reduced overall delay. Additionally, the shortened response time configured for UDs further contributes to latency reduction. Compared with UDs, and legacy 5G, NUDs have a significantly higher delay and the corresponding value generally increases with a higher L. However, in comparison with the baseline scheme, NUDs still attain lower latency. When $\eta = 0.5$, which indicates lesser competition among NUDs, shorter latency for NUDs is achieved at a cost of slightly increased latency for UDs.

A.7.3 HS Performance

The performance of the HS scheme is illustrated in Figs. A.13 - A.15. It is clear that the performance of GDs in HS is similar to what is observed in DGDP. Furthermore, UDs, which are a subset of NGDs, exhibit also similar performance as what is observed in the RAUG scheme. Recall, however, that the number of competing IoT devices in each device type will be different when NGDs are categorized into UDs and NUDs.

As a result, NUDs in HS achieve much better performance compared with NUDs in RAUG and NGDs in DGDP even though their available number of preamble, $\hat{\phi}$, is lower than in RAUG or DGDP. Furthermore, since both GDs and UDs coexist in HS, a much

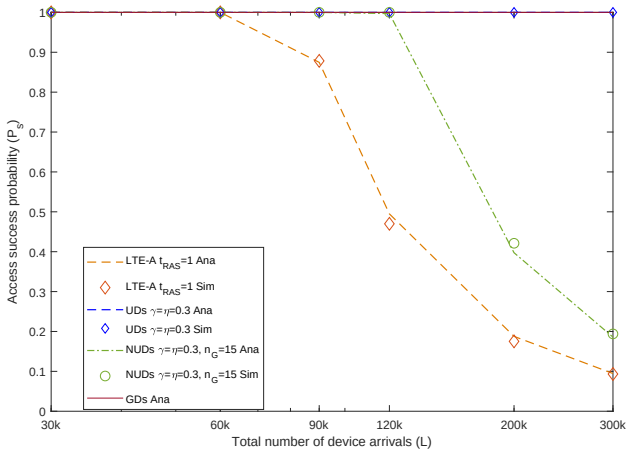


Figure A.14: Access success probability in HS: GDs, UDs, versus NUDs.

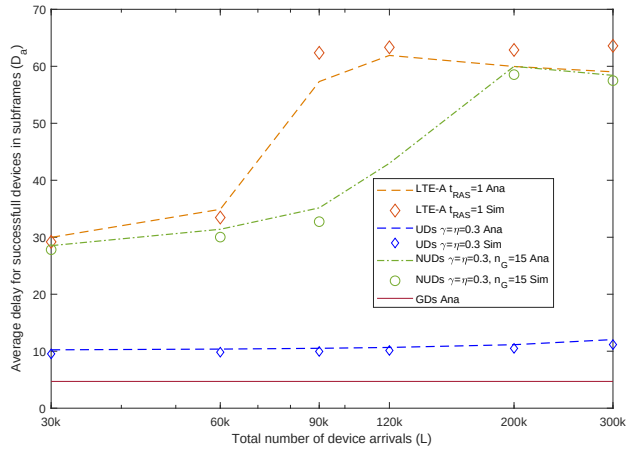


Figure A.15: Average delay of the successfully accessed devices for different types of devices in HS.

larger number of devices with URLLC requirements can be accommodated when HS is employed. Observe Fig. A.14 and take $L = 200$k, $\gamma = \eta = 0.3$, and $n_G = 15$ as an example. The total number of IoT devices that achieve $P_s = 1$ would be as many as 102k including $\gamma L = 60$k GDs grouped in 15 groups plus $\eta(1-\gamma)L = 42$k UDs.

A.7.4 The Impact of γ, η, and n_G

As mentioned earlier, the values of γ, η, and n_G are reconfigurable. In a cell with other parameters like L and ϕ fixed, the adopted values of these three variables play a significant role in determining the performance of the proposed schemes. A higher γ value means a larger number of GDs and accordingly n_G also needs to be enlarged. The performance of NGDs in DGDP depends on the *joint configuration* of γ and n_G values. Similarly, increasing η would lead to a higher number of UDs in RAUG indicating more competition among UDs and better access opportunities for NUDs, respectively.

To achieve optimal performance from the proposed initial access schemes, it is vital to configure network parameters appropriately so that, while GDs and UDs enjoy URLLC service, NGDs and NUDs could also achieve better or at least similar performance in

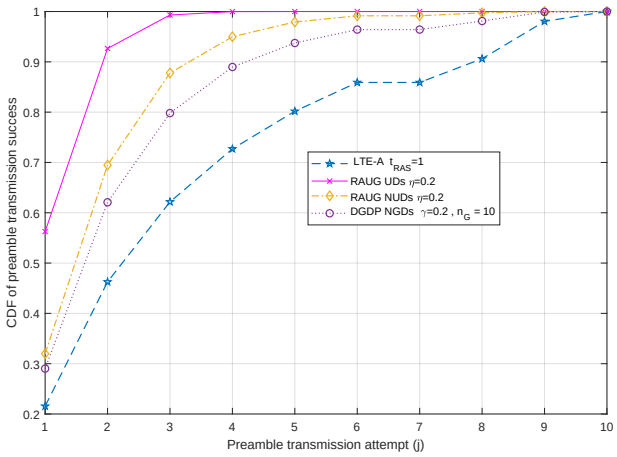


Figure A.16: CDF of successful preamble transmissions for different types of devices under LTE-A, DGDP, and RAUG respectively.

comparison with the baseline scheme. Observing the presented numerical results for DGDP, it is evident that the selected γ values satisfy the criterion given in (A.2). Any violation of this criterion would deteriorate the performance of NGDs as further discussed in [8]. Furthermore, the impact of η values on the performance of NUDs has a simpler proportional relationship. Whenever η is increased, NUDs will obtain better performance owing to reduced competition, as observed in the numerical results for RAUG. However, *η should only be enlarged to a level up to which the required performance for UDs is still guaranteed.*

A.7.5 Performance Comparison among Our Schemes and versus LTE-A

As demonstrated above, the proposed schemes outperform the baseline scheme under all studied configurations. To elaborate the performance distinctions, we further differentiate the results obtained from the baseline scheme with two configuration options, i.e., when the interval between two successive RA slots is configured as $t_{RAS} = 5$ and $t_{RAS} = 1$, respectively.

The baseline scheme with $t_{RAS} = 5$ performs worst among all the studied schemes. Although this configuration is commonly adopted in LTE-A, our results reveal that this is not an effective option when the number of IoT devices could increase promptly, e.g., under mMTC bursty traffic scenarios. When $t_{RAS} = 1$, the performance of the baseline scheme improves significantly, thanks to a much higher number of RA slots (10000 for $t_{RAS} = 1$ versus 2000 for $t_{RAS} = 5$) available for preamble transmissions of arriving devices. However, when the number of IoT devices is very large, i.e., $L > 90$k, the performance of this configuration also degrades more seriously than what is achieved in our proposed schemes.

Among the proposed schemes, DGDP provides the best URLLC performance for GDs since GDs always enjoy guaranteed access privilege with null collision based on their contention-free access. The proposed 2-step handshake procedure combined with lower response times further reduces the latency for GDs. The performance of UDs in the HS

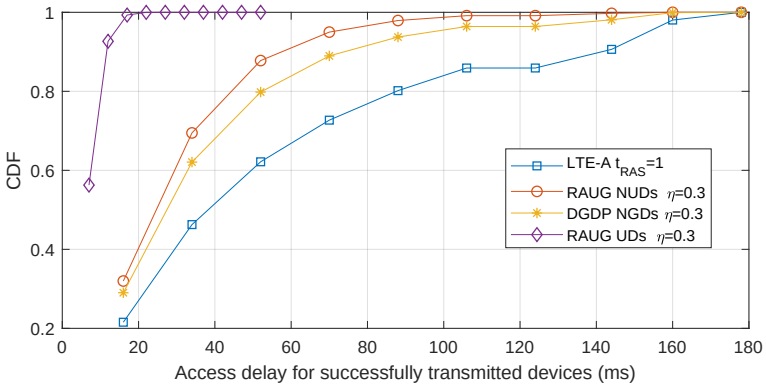


Figure A.17: CDF of access delay for successful UDs, NGDs, and NUDs: Comparison of RAUG, DGDP, and LTE-A.

and RAUG schemes is better than any NUDs or NGDs in all cases. UDs benefit from the proposed dedicated RA slots with reduced latency obtained by allowing multiple slots inside one subframe for preamble transmission and also from the shortened response times. However, the performance of UDs in RAUG is not as superb as GDs in DGDP since UDs in RAUG still need to follow the 4-step RA procedure and to compete with other UDs. Nevertheless, unlike GDs, UDs have more flexibility in terms of device implementation and the support of various IoT applications (since no pre-grouping is required and no requirement on static deployments). Moreover, with the same γ and η configuration, NUDs in RAUG achieve generally better performance in comparison with NGDs. Since two dedicated RACH slots are enabled inside a subframe and no preambles are pre-allocated to GDs, the access opportunities for NUDs are based on all ϕ preambles. In contrast, NGDs in DGDP have only $(\phi - n_G)$ preambles, leading to slightly degraded performance in comparison with NUDs in RAUG.

Furthermore, for the purpose of performance comparison under a medium size device population, we reconfigure the network as $L = 90$k, $\gamma = \eta = 0.2$, and $n_G = 10$. In Fig. A.16, we illustrate the cumulative distributed function (CDF) of successful preamble transmissions for different types of IoT devices under LTE-A, DGDP, and RAUG, respectively. As can be observed, almost all UDs in RAUG obtain network access within three preamble transmissions. Moreover, NGDs and NUDs have also achieved higher CDF values compared with the baseline scheme. With a cross-reference of the respective P_s values in Fig. A.16, we ascertain that DGDP and RAUG provide faster access to the network than the baseline scheme does. For instance, to achieve $P_s = 95\%$, NUDs in RAUG need on average merely 4 preamble transmissions whereas about 6 and $7 \sim 8$ transmissions are required for NGDs in DGDP and devices in LTE-A respectively. In Fig. A.17, we further illustrate the CDF of the access latency experienced by successfully transmitted devices in milliseconds based on the four studied schemes. It is evident that all devices in our schemes including UD, NUDs, and NGDs have achieved better performance in comparison with that of LTE-A and among them UDs obtain the best performance.

Moreover, HS offers best opportunities to all types of IoT devices owing to its hybrid nature. When HS is employed, both GDs and UDs could coexist without compromising each other's performance, thus supporting a higher number of IoT devices with URLLC

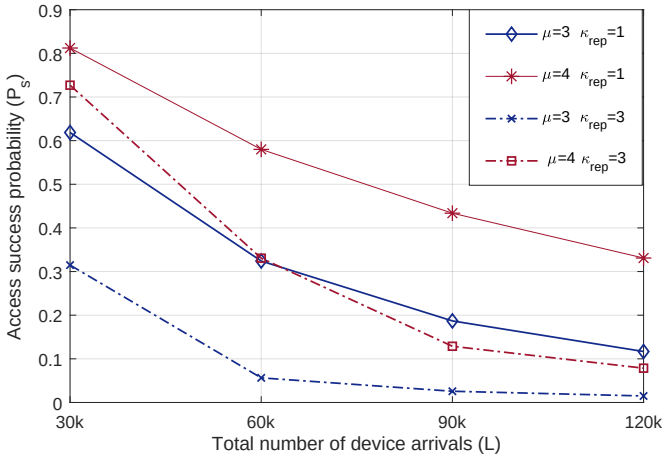


Figure A.18: Access success probability for GF transmissions under bursty traffic.

requirements. Although NUDs in HS possess a smaller set of preambles, i.e., $(\phi - n_G)$, the same as NGDs in DGDP, the number of NUDs is meanwhile significantly reduced to $(1-\eta)(1-\gamma)L$ which is lower than that of NGDs in DGDP, i.e., $(1-\gamma)L$. In this way, the performance of NUDs in HS is also improved.

A.7.6 Access Success Probability Comparison with Grant-free Transmission

As mentioned in Subsec. A.2.2, GF transmission appears as an attractive mechanism for data reporting in various mMTC and URLLC scenarios, especially for small data and sporadic traffic. In this subsection, we compare through simulations the performance of the proposed schemes with GF in terms of access success probability by considering two numerologies $\mu = 3$ and $\mu = 4$ which have 8 and 16 slots respectively. To maximize resource allocation for GF transmissions, we assume that all these available slots can be utilized by GF traffic. For GF transmissions, we adopt a popular transmission scheme known as *k repetitions* [9]. Accordingly, a number of k_{rep} replicas of the same packet will be transmitted within a subframe. A packet transmission is regarded as successful if at least one of these k_{rep} transmissions is successful.

For GF transmissions, all devices that arrived during a given subframe will compete for transmission in the next subframe. Each device will randomly select one or more (if $k_{rep} > 1$) slots based on the configuration and transmit k_{rep} replicas of its packet in the selected slot(s) *within the same subframe*. A collision happens if two or more devices have selected the same slot for transmitting any replica of their packets.

Fig. A.18 illustrates the obtained access success probability of GF devices according to a bursty traffic arrival pattern which was presented in Subsec. A.3.3. As expected, the success probability monotonically decreases with a higher number of device arrivals. When comparing the results for $\mu = 3$ and $\mu = 4$, it is clear that providing a higher number of slots for GF data transmission would result in a higher access success probability. On the other hand, it is counter-intuitive that having a higher number of repetitions does not help to increase access success. This is because with $k_{rep} = 3$, the number of competing

transmissions per slot increases, leading to an even lower success probability.

Finally, let us compare the access success probability achieved by GF devices with what is achieved in the proposed DGDP and RAUG schemes which belong to GB schemes (with the results shown in Fig. A.8 and Fig. A.11 respectively). By comparing the curves in these figures, it is evident that the GB schemes perform better. This is because during a traffic burst, a very higher number of arrivals within a short interval have occurred, causing a higher number of collisions for both GB access and GF transmissions. Initially, the number of arrivals for each subframe is the same for both GB and GF. Although a GB scheme has to deal with retransmissions, it has the advantage of transmitting up to n_{PT} transmissions across multiple subframes. On the other hand, a *k repetitions* GF scheme has to finish all k_{rep} transmissions within one subframe without the possibility of retransmissions. As a consequence, GF transmissions experience higher collisions than GB transmissions, resulting in a lower access success probability. Based on this observation, we ascertain that, although GF communication reduces extra overhead by skipping the initial access phase and it provides lower latency when traffic load is light, it is not better suited for providing URLLC services *in the presence of bursty traffic with a high number of arrivals.*

A.7.7 Further Discussions

The proposed schemes are developed based on multiple assumptions as presented above. For instance, the procedures for intra-group communications between group members and their group leader are not included in our scheme design. Nor is the coordination between a serving group leader and its backup group leader considered. In spite of having a very lower probability, it is not impossible that neither of the group leaders sensed an event or the transmissions of both leaders failed. If such an extreme case occurs, extra intra-group communication is needed. Although intra-group communications could be performed with or without the involvement of downlink message coordinations through a gNB, extra protocol overhead and longer delay are unavoidable. As such, the reported results in this section represent an upper bound of the performance of our schemes.

A.8 Conclusions and Future Work

Targeting at two massive IoT traffic scenarios, we have proposed in this paper three LTE-A or 5G NR based initial access schemes which provide URLLC access to a selected portion of mMTC devices. The schemes were developed by considering various mission critical and cyber-physical IoT applications envisaged by 3GPP. The first scheme, DGDP, provides contention-free access with low latency to grouped IoT devices based on dedicated preamble reservation. The second scheme, RAUG, is still contention based but facilitates reserved random access slots allowing multiple occurrences inside each subframe and hence produces lower latency and very high access success probabilities to those IoT devices with URLLC requirements. The third scheme, HS, combines the merits of these two schemes and provides more flexibility to a larger number of URLLC devices as well as non-grouped

and non-URLLC devices. Furthermore, the performance of all four types of IoT devices under these three schemes has been evaluated based on both analysis and simulations, in comparison with the legacy LTE-A initial access as well as grant-free transmission. Through performance comparison, we demonstrate that, by fine-tuning a few configurable network parameters, the proposed schemes are able to provide ultra-high reliability and low latency to grouped devices and URLLC devices while still improving the performance of non-grouped and non-URLLC devices. As future work, we will further study both inter- and intra-group communications in a two-tier architecture for mMTC networks, intra-group communications among devices and group leaders, and initial access for beyond 5G networks together with data transmission and radio resource allocation after the initial access phase. For protocol design, we will also consider more realistic channel conditions, the constraint of radio resource blocks, as well as minimized extra protocol overhead.

References

[1] H. Habibzadeha, T. Soyataa, B. Kantarci, A. Boukerche, C. Kaptan, Sensing, communication and security planes: A new challenge for a smart city system design, Comput. Netw. 144 (2018) 163–200.

[2] O. Galinina, S. Andreev, M. Komarov, S. Maltseva, Leveraging heterogeneous device connectivity in a converged 5G-IoT ecosystem, Comput. Netw. 128, (2017) 123-132.

[3] 3GPP TS 22.368, Service requirements for machine-type communications (MTC); Stage 1, R15, v15.0.0, 2019.

[4] M.S. Ali, E. Hossain, D.I. Kim, LTE/LTE-A random access for massive machine-type communications in smart cities, IEEE Commun. Mag. 55 (1) (2017) 76–83.

[5] G. Hampel, C. Li, J. Li, 5G ultra-reliable low-latency communications in factory automation leveraging licensed and unlicensed bands, IEEE Commun. Mag. 57 (5) (2019), 117–123.

[6] S. Zhang, Y. Wang, W. Zhou, Towards secure 5G networks: A Survey, Comput. Netw. 162 (2019) 1–22.

[7] G.J. Sutton, J. Zeng, R.P. Liu, W. Ni, D.N. Nguyen, B.A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, T. Lv, Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives, IEEE Commun. Surv. Tut. 21 (3) (2019) 2488–2524

[8] T.N. Weerasinghe, I.A.M. Balapuwaduge, F.Y. Li, Preamble reservation based access for grouped mMMTC devices with URLLC requirements, in: Proc. IEEE ICC, 2019, pp.1-6.

[9] M. Bennis, M. Debbah, H.V. Poor, Ultra reliable and low-latency wireless communication: Tail, risk, and scale, Proc. IEEE, 106 (10) (2018) 1834–1853.

[10] 3GPP TS 22.104, Service requirements for cyber-physical control applications in vertical domains, R17, v17.0.0, 2019.

[11] 3GPP TR 37.868, Study on RAN improvements for machine type communications, R11, v11.0.0, 2011.

[12] 3GPP TS 36.321, Evolved universal terrestrial radio access (e-UTRA), R15, v15.6.0, 2019.

[13] P. Castagno, V. Mancuso, M. Sereno, M.A. Marsan, Limitations and sidelink-based extensions of 3GPP cellular access protocols for very crowded environments, Comput. Netw. 168, (2020) 1–15.

[14] M. Tavana, A. Rahmati, V. Shah-Mansouri, Congestion control with adaptive access class barring for LTE M2M overload using Kalman filters, Comput. Netw. 141 (2018) 222–233.

[15] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks, in: Proc. IEEE ICC, 2016, pp. 1–6.

[16] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, C. Zhou, J. Eichinger, Ö. Bulakci, On the way to massive access in 5G: Challenges and solutions for massive machine communications, in: Proc. International Conference on Cognitive Radio Oriented Wireless Networks (CROWNCOM), 2015, pp. 708–717.

[17] B. Han, H.D. Schotten, Grouping-based random access collision control for massive machine-type communication, in: Proc. IEEE GLOBECOM, 2017, pp. 1–7.

[18] H. Seo, J. Hong, W. Choi, Low latency random access for sporadic MTC devices in Internet of things, IEEE Internet Things J. 6 (3) (2019) 5108-5118.

[19] 3GPP TS 38.213, NR; Physical layer procedures for control, R15, v15.6.0, 2019.

[20] G. Sanfilippo, O. Galinina, S. Andreev, S. Pizzi, G. Araniti, A concise review of 5G new radio capabilities for directional access at mmWave frequencies, in: Proc. International Conference on Next Generation Wired/Wireless Advanced Networks and Systems (NEW2AN), 2018, pp. 340–354.

[21] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, H. Wei, 5G new radio: Waveform, frame structure, multiple access, and initial access, IEEE Commun. Mag. 55 (6) (2017) 64–71.

[22] 3GPP TS 38.211, NR; Physical channels and modulation, R16, v16.1.0, 2020.

[23] B. Singh, O. Tirkkonen, Z. Li, M.A. Uusitalo, Contention-based access for ultra-reliable low latency uplink transmissions, IEEE Wireless Commun. Lett. 7 (2) (2018) 182–185.

[24] 3GPP TR 38.824, Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC), R16, v16.0.0, 2019.

[25] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H.V. Poor, B. Vucetic, High-reliability and low-latency wireless communication for Internet of things: Challenges, fundamentals and enabling technologies, IEEE Internet Things J. 6 (5) (2019) 7946–7970.

[26] A. Azari, P. Popovski, G. Miao, C. Stefanovic, Grant-free radio access for short-packet communications over 5G networks, in: Proc. IEEE GLOBECOM, 2017, pp. 1-7.

[27] A.T. Abebe, C.G. Kang, Comprehensive grant-free random access for massive and low latency communication, in: Proc. IEEE ICC, 2017, pp. 1-6.

[28] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme, IEEE Trans. Wireless Commun. 16 (12) (2017) 7785–7799.

[29] P. Zhou, H. Hu, H. Wang, H.H. Chen, An efficient random access scheme for OFDMA systems with implicit message transmission, IEEE Trans. Wireless Commun. 7 (7) (2008) 2790-2797.

[30] C. Wei, G. Bianchi, R. Cheng, Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks, IEEE Trans. Wireless Commun. 14 (4) (2015) 1940–1953.

[31] 3GPP TS 38.331, NR; Radio resource control (RRC) protocol specification, R15, v15.6.0, 2019.

[32] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. Vidal, V. Casares-Giner, L. Guijarro, Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic, IEEE Trans. Veh. Technol. 67(4) (2018) 3505-3520.

[33] J. Li, Q. Du, L. Sun, P. Ren, Queue-aware joint ACB control and resource allocation for mMTC networks, in Proc. IEEE GLOBECOM Workshops, 2018, pp. 1-6.

[34] J. Zhang, X. Tao, H. Wu, N. Zhang, X. Zhang, Deep reinforcement learning for throughput improvement of uplink grant-free NOMA system, IEEE Internet Things J., Early Access, DOI:10.1109/JIOT.2020.2972274, (2020).

[35] L. Liang, L. X. B. Cao, Y. Jia, A cluster-based congestion-mitigating access scheme for massive M2M communications in Internet of things, IEEE Internet Things J. 5(3) (2018) 2200–2211.

[36] F. Wu, B. Zhang, W. Fan, X. Tian, S. Huang, C. Yu, Y. Liu, An enhanced random access algorithm based on the clustering-reuse preamble allocation in NB-IoT system, IEEE Access 7 (2019) 183847–183859.

[37] T. Kim, H. S. Jang, D. K. Sung, An enhanced random access scheme with spatial group based reusable preamble allocation in cellular M2M networks, IEEE Commun. Lett., 19(10) (2015) 1714–1717.

[38] Q. Pan, X. Wen, Z. Lu, W. Jing, L. Li, Cluster-based group paging for massive machine type communications under 5G networks, IEEE Access 6 (2018) 64981–64904.

[39] S. Sesia, I. Toufik, M. Baker, *LTE - the UMTS long term evolution: From theory to practice*, 2nd Edition, John Wiley & Sons Ltd., 2011.

[40] M. Rahnema, M. Dryjanski, *From LTE to LTE-Advanced Pro and 5G*, Artech House, 2017.

[41] 3GPP TS 36.331, Radio resource control (RRC); Protocol specification, R15, v15.8.0, Dec. 2019.

[42] I. Leyva-Mayorga, C. Stefanovic, P. Popovski, V. Pla, J. Martinez-Bauset, Random access for machine-type communications, Wiley 5G Ref: The Essential 5G reference Online, 2019.

[43] O. Arouk, A. Ksentini, General model for RACH procedure performance analysis, IEEE Commun. Lett. 20 (2) (2016) 372–75.

[44] T. Weerasinghe, I. A. M. Balapuwaduge, F. Y. Li, Supervised learning based arrival prediction and dynamic preamble allocation for bursty traffic, in Proc. IEEE INFOCOM Workshops, 2019, pp.1-6.

[45] A. Azari, M. Ozger, C. Cavdar, Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning, IEEE Commun. Mag. 57(3) (2019) 42–48.

[46] 3GPP RP-191677, Revised work item proposal: 2-step RACH for NR, 3GPP TSG RAN Meeting #85, 2019.

[47] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, T. H. Jacobsen, Uplink grant-free random access solutions for URLLC services in 5G new radio, in Proc. IEEE ISWCS, 2019, pp. 607-612.

Paper B

Title:	Priority-based Initial Access for URLLC Traffic in Massive IoT Networks: Schemes and Performance Analysis
Authors:	Thilina N. Weerasinghe[†], Vicente Casares-Giner[‡], Indika A. M. Balapuwaduge[†], and Frank Y. Li[†]
Affiliation:	[†]Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway [‡]Dept. of Communications, Universitat Politècnica de València (UPV), 46022 València, Spain
Journal:	*IEEE Transactions on Communications*, early access article, Jan. 2021

Priority Enabled Grant-Free Access with Dynamic Slot Allocation for Heterogeneous mMTC Traffic in 5G NR Networks

Thilina N. Weerasinghe, Vicente Casares-Giner, Indika A. M. Balapuwaduge, and Frank Y. Li

Abstract - Although grant-based mechanisms have been a predominant approach for wireless access for years, the additional latency required for initial handshake message exchange and the extra control overhead for packet transmissions have stimulated the emergence of grant-free (GF) transmission. GF access provides a promising mechanism for carrying low and moderate traffic with small data and fits especially well for massive machine type communications (mMTC) applications. Despite a surge of interest in GF access, how to handle heterogeneous mMTC traffic based on GF mechanisms has not been investigated in depth. In this paper, we propose a priority enabled GF access scheme which performs dynamic slot allocation in each 5G new radio subframe to devices with different priority levels on a subframe-by-subframe basis. While high priority traffic has access privilege for slot occupancy, the remaining slots in the same subframe will be allocated to low priority traffic. To evaluate the performance of the proposed scheme, we develop a two-dimensional Markov chain model which integrates these two types of traffic via a pseudo-aggregated process. Furthermore, the model is validated through simulations and the performance of the scheme is evaluated both analytically and by simulations and compared with two other GF access schemes.

Keywords - Grant-free access, NR numerology, mMTC traffic, dynamic slot allocation, two-dimensional Markov chain, pseudo-aggregated process.

B.1 Introduction

Simultaneous packet transmissions over the same radio resource cause performance deterioration for wireless access due to potential collisions among transmissions from competing devices. In fourth generation (4G) cellular networks, i.e., long term evolution-advanced (LTE-A), this problem was primarily addressed using grant-based (GB) communications. For GB channel access, a device follows a four-step handshake procedure for initial access with an evolved nodeB (eNB) by first transmitting a preamble before it obtains a grant for its data packet transmission. Once access is granted by the eNB, a data packet can be successfully transmitted without collision under ideal channel conditions. The initial preamble transmission, however, is still subject to collision(s) and could require multiple transmissions depending on traffic load and the availability of preamble resources at the eNB.

In LTE-A, the time required for initial four-step handshaking, which occurs prior to a data transmission, is in the order of 15 ms [1]. This is not a major concern since many 4G

applications do not have stringent low latency requirements. In emerging fifth generation (5G) networks specified by the 3rd generation partnership project (3GPP), however, a variety of applications necessitate novel approaches for ultra-reliable low latency communications (URLLC) and massive machine type communications (mMTC). For small data transmissions which are common for mMTC traffic, the amount of control overhead required before an actual data transmission in GB schemes is too high with respect to the actual data to be transmitted and the handshake procedure lasts too long [2].

Although GB initial access is still kept as a legacy mechanism in 5G new radio (NR) networks, to perform such a four-step initial access procedure requires extra delay and protocol overhead [3] [4]. As an alternative to reduce overall latency, another category of mechanisms for data transmission, known as grant-free (GF), configured grant, or without grant, has emerged [4] [5]. Different from the GB principle, devices in GF communications transmit their data packets together with (or without using specific) control messages directly to a 5G NR nodeB (gNB) in available GF slots *without requiring the initial access procedure*. In other words, no dedicated preamble transmission for granting access and allocating radio resources is required for GF communications before starting a data packet transmission [3]. The benefits brought by this principle in terms of shortened delay and reduced protocol overhead make GF mechanisms attractive for various applications with URLLC/mMTC requirements and small data packets [1].

For periodic or deterministic traffic, a gNB can allocate dedicated slots to devices for their data transmissions. However, such a mechanism will lead to resource underutilization and long delay when traffic load is low or sporadic which is the case for many mMTC applications. Due to the unpredictability of sporadic traffic arrival patterns, it is beneficial to apply a *random access* protocol for GF data transmissions based on the principles of ALOHA or slotted ALOHA [6]. Furthermore, GF transmissions are generally recommended for small data transmission with a low or moderate level of traffic arrivals [7] [8].

B.1.1 Related Work

B.1.1.1 GF communications

While GF is a more popular terminology favored by the research community, similar mechanisms are commonly referred to as configured grant or without grant in 3GPP specifications [4] [5] [9] [10]. In brief, existing GF based transmission schemes can be classified into four major categories, as summarized below. (i) *GF reactive*: A device needs to send its GF transmission and wait for an acknowledgment (ACK) or a negative ACK (NACK) from the gNB. If no ACK is received within the ACK timeout, or a NACK is received, the same packet will be retransmitted up to a retry limit; (ii) *GF reactive with power boost*: In order to increase successful reception probability, the transmit power of each retransmission could be higher than that of the previous unsuccessful transmission; (iii) *K repetitions without feedback*: A device transmits proactively $K > 1$ replicas of the same data packet across different GF slots in the same subframe [9]; and (iv) *K repetitions with feedback*: Similar to (iii), but it requires feedback from a gNB regarding its transmission status. Accordingly, a device will stop its transmission attempt once an ACK

is received. Furthermore, the 3GPP states clearly that *at least an uplink transmission scheme without grant is supported for URLLC* and *an uplink transmission scheme without grant is targeted to be supported for mMTC [5].*

On the other hand, recent academic efforts foresee the feasibility of facilitating multi-packet reception by applying more advanced technologies for instance non-orthogonal multiple access (NOMA) and multiple-input multiple-output (MIMO) to GF transmissions. By treating collisions as interference through successive joint decoding or successive interference cancellation (SIC), [11] derived expressions for outage probability and throughput for GF-NOMA transmissions. In [12], a semi-GF scheme which provides dedicated GB access for one user while facilitating the other users with GF opportunistic access was proposed. Another recent work investigated the suitability of applying non-orthogonal sequences for abbreviating preamble collisions for GF transmissions and concluded that such sequences did not necessarily lead to better performance than the orthogonal ones [13]. In general, GF access exhibits the characteristic of slotted ALOHA-alike access mechanisms as presented below.

B.1.1.2 Slotted, framed slotted, and SIC-enabled slotted ALOHA for MTC access

Depending on multi-packet reception is enabled or not, numerous variants of ALOHA-alike protocols, including framed slotted ALOHA (FSA), multi-channel slotted ALOHA, and SIC-enabled slotted ALOHA play an important role for medium access in mMTC [2] [14].

Based on the requirements for mMTC applications and design principles, FSA can be operated with either fixed or flexible frame length [15]. On the other hand, channels in multi-channel slotted ALOHA regard to different kinds of orthogonal resources such as codes or preambles which are used in the same, for instance time slot, during the initial access procedure. Using different orthogonal resources, multiple devices can access to a common channel simultaneously [1]. However, the amount of resources is still limited. For random access of mMTC traffic without multi-packet reception capability, a collision happens if two or more devices select the same preamble for their initial access or transmit their packets simultaneously in the same slot. More recent work intends to resolve collision following the principle of SIC through coded slotted ALOHA, e.g., in the form of frameless ALOHA [16].

Furthermore, priority oriented schemes in FSA have been studied previously. In [17], a pseudo-Bayesian ALOHA algorithm with mixed priorities was proposed. Similar to the pseudo-Bayesian ALOHA scheme presented in [18], that algorithm allows multiple independent Poisson traffic streams compete for a slot or a batch of slots in a frame each with an assigned transmission probability. Following the idea on resource sharing, an adaptive framed pseudo-Bayesian ALOHA algorithm was proposed in [19].

Considering that the subframe length in NR is constant as 1 ms regardless of the adopted NR numerology [20], we adopt a fixed subframe length for our scheme design. Furthermore, since no dedicated preamble for initiating access and resource allocation is needed for GF transmissions, the access scheme proposed below in Sec. B.3 allows devices transmit their packets directly to the associated gNB in the allocated GF slots.

The scheme is designed upon the FSA principle but is based on the NR frame structure to be presented in Subsec. B.2.1.

B.1.2 Contributions

So far, little work has been done considering GF access for heterogeneous mMTC traffic. In this paper, we consider heterogeneous GF traffic arrivals with different reliability and/or latency requirements and propose a novel GF based access and data transmission scheme with dynamic slot allocation (DSA) in each NR subframe. Hereafter, the scheme is referred to as DSA-GF which stands for DSA for GF based access for heterogeneous traffic. Targeting at providing better performance to high priority traffic (HPT), the scheme accommodates the remaining slots in the same subframe to low priority traffic (LPT) so that higher total slot utilization is achieved.

In contrast to most existing work which generally neglected *slot based* GF transmissions and slot utilization, this paper targets at 5G NR numerology as the basis for our scheme design and intends to maximize slot utilization for heterogeneous traffic integrated with priority enabled access. Through dynamic slot allocation, the dependence of two types of GF traffic is handled and modeled through a pseudo-aggregated process where both traffic types share available slots in each subframe and slot allocation to HPT is independent of that of LPT.

In brief, the main contributions of this paper are summarized as follows.

- Based on the NR frame structure, a novel GF based data transmission scheme, DSA-GF, which considers arrivals of heterogeneous GF traffic is proposed. The scheme performs traffic estimation, access control, and dynamic slot allocation on a subframe-by-subframe basis. Based on our scheme, both HPT access privilege and LPT resource preservation are achieved and they are bound together smoothly in each subframe.

- To evaluate the performance of the proposed scheme, a two-dimensional (2D) Markov chain model, in which a pseudo-aggregated process is defined to link two types of GF traffic by considering their coherence for slot allocation in a common subframe, has been developed. For a network with the same configuration, the number of states in our model is much less than what is needed in conventional Markov models.

- Extensive discrete-event based simulations have been performed to validate the preciseness of the developed model and assess the performance of the DSA-GF scheme. Through performance assessment under various HPT/LPT traffic variations and comparison with two other GF schemes, the effectiveness of the proposed scheme is further demonstrated.

In a nutshell, the uniqueness and novelty of our paper are reflected by the fact that this work is anchored at a niche with an intersection among 5G NR numerology, traffic estimation based dynamic slot allocation, proper handling of heterogeneous traffic considering the performance of both HPT and LPT, and pseudo-aggregated 2D Markov chain

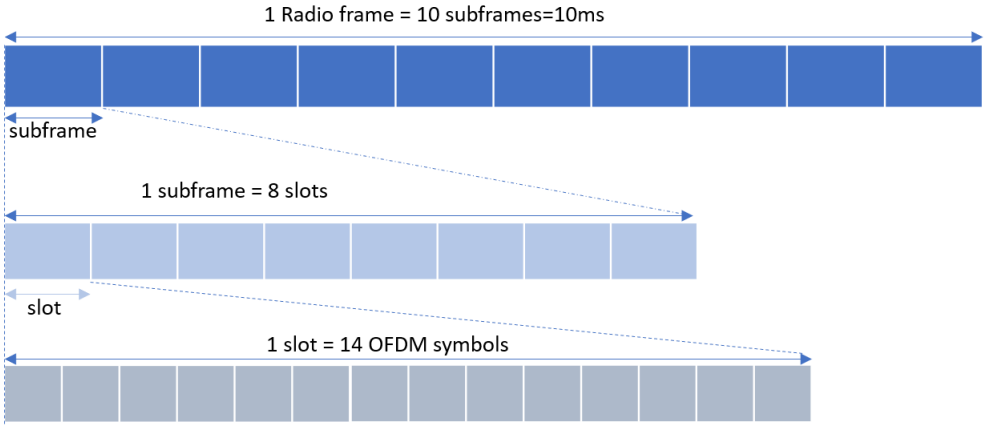


Figure B.1: 5G NR frame structure for numerology $\beta = 3$.

modeling for heterogeneous traffic. To the best of our knowledge, this is the first attempt which is dedicated for 5G NR numerology based GF transmission with dynamic slot allocation at the subframe level for heterogeneous traffic, combined with a Markov model with a significantly reduced state space bridging both types of traffic together for performance evaluation.

Furthermore, it is worth mentioning that the 3GPP has newly decided to discontinue NOMA as a work-item for 5G NR but leave it as a study-item for beyond 5G [21]. Under such a circumstance, the importance of investigating viable GF schemes based on the existing NR frame structure remains significant and it becomes even an imperative task as such schemes may serve as the basis or at least references for NOMA based GF scheme design.

The remainder of the paper is organized as follows. Sec. B.2 provides preliminaries on NR numerology and presents the network scenario. In Sec. B.3, the proposed scheme is explained in details. Then we develop a 2D Markov model in Sec. B.4 to analyze its performance. Thereafter, Sec. B.5 illustrates the numerical results. Finally, the paper is concluded in Sec. B.6.

B.2 Preliminaries, Scenario and Assumptions

This section presents the NR frame structure which forms the basis for our scheme design and outlines the scenario.

B.2.1 5G NR Frame Structure and Numerologies

With 15 kHz as a baseline for subcarriers as used in 4G, 5G NR defines five numerologies based on subcarrier spacing $\Delta f = 2^{\beta} * 15$ kHz, where $\beta = 0, 1, ..., 4$ is the numerology index, with different slot duration lengths downwards from 1 ms to 62.5 μs [20] [22]. As depicted in Fig. B.1, the per frame duration in NR is still 10 ms, and the same as in LTE-A, one *frame* consists of 10 *subframes* each with 1 ms duration. Moreover, one NR subframe may have one (for $\beta = 0$) or multiple (up to 16) *slots* depending on the value of the numerology index β.

Depending on the size of a packet, one or multiple orthogonal frequency division multiplexing (OFDM) symbols out of the available 14 symbols within a slot can be utilized by GF traffic [9] [10]. Considering that GF transmissions are targeted at small data packets in mMTC networks [24], we assume in this study that a packet with the size of less than 14 OFDM symbols is sufficient for one GF packet transmission. The remaining symbol(s) within the same slot can be allocated to other data traffic (for instance GB transmissions) and control information exchange as NR allows flexible uplink and downlink scheduling *at a symbol level* within one NR slot [20]. As such, *all slots in a subframe* can be utilized for GF data transmissions.

B.2.2 Scenario and Traffic Arrivals

Consider a scenario where an NR cell covers a large number of mMTC devices. Although both GF and GB devices may coexist, this study focuses only on GF data transmissions. More specially, GF data transmissions considered in this study are performed in each subframe following the DSA-GF scheme presented in the next section. A device is regarded as *active* if it has one packet ready to transmit. The transmission of a device is regarded as successful if no other device transmits in the same slot and it is confirmed through an ACK message provided at the end of each subframe. If two or more devices transmit in the same slot, a collision occurs and all involved transmissions are considered to be failed. If a device does not obtain a transmission opportunity due to the constraint of the permission probability in the current subframe or its transmission in the current subframe collided, it will try again in the next subframe based on a new permission probability broadcast by the gNB right before the next subframe begins.

Although the total number of mMTC devices covered by a cell could be huge [25], they generate typically sporadic traffic with small packet sizes. Therefore, the number of arrivals per subframe, i.e., within 1 ms, is rather limited. To reflect this, we adopt a combination of number of devices and activation probability as an indicator to represent offered traffic.

Without loss of generality, we consider numerology $\beta = 3$ as an example in most figures and descriptions in this paper. Later on in Subsec. B.5.6, we further demonstrate the applicability of the scheme to two other numerologies, i.e., $\beta = 2$ and $\beta = 4$ which have 4 and 16 slots per subframe respectively.

Two categories of traffic arrivals are considered, known as HPT and LPT respectively. While HPT requires superior performance, LPT can tolerate longer access delay and higher packet loss. For slot allocation in each subframe, HPT has access privilege over its counterpart, i.e., LPT. In the considered cell covered by one gNB, there are a finite number of HPT and LPT devices, denoted by M_x with $x = 1$ for HPT and $x = 2$ for LPT, respectively. The arrival process for both categories follows a Bernoulli process. That is, each device generates one data packet per subframe with activation probability a_x. This assumption means that each device has at most one packet ready to transmit at each subframe. Furthermore, we assume that the ACK message transmission from the gNB is always successful. No channel impairment is considered in this study and propagation delay is regarded to be negligible compared with access delay.

B.3 Proposed Transmission Scheme for GF Traffic

The DSA-GF scheme focuses on the NR frame structure and features the flavors of both 4G and 5G access mechanisms such as access class barring and unified access class [1] [23], imposing different permission probabilities to heterogeneous types of traffic. It is operated on a subframe-by-subframe basis. First of all, an observation-based slot allocation algorithm assigns an optimal number of slots to serve HPT transmissions in order to achieve maximum throughput, low access delay and reduced packet loss probability. In the meantime, the algorithm takes into account the performance of LPT through *slot preservation to LPT in order to avoid starvation of LPT*. To do so, the maximal number of slots to be allocated to HPT is restricted to the total number of slots per subframe *minus* one (for $\beta = 2$ and 3) or two (for $\beta = 4$). Then the remaining slots in the same subframe will be allocated to LPT. For a given subframe, the more slots allocated to HPT, the less slots assigned to LPT.

For a given numerology, the total number of slots per subframe, denoted by U, is a constant and it is decided by the NR frame structure presented above. Inspired by the pseudo-Bayesian broadcast algorithm for slotted ALOHA proposed in [18], we develop a novel random access scheme for NR based GF transmissions as presented below. While the algorithm in [18] targeted at slotted ALOHA with a single slot, the protocol designed in this paper is tailored to operations where multiple slots together form one subframe, taking into account the NR frame structure.

B.3.1 Transmission Principles of DSA-GF

At the beginning of each subframe, the gNB provides to all devices through a broadcast message with the permission probability for each type of traffic, denoted as p_x where $x = 1$ for HPT and $x = 2$ for LPT respectively. With probability p_x, each *active* device randomly selects one of the allocated slots to type x within the current subframe to transmit its packet. With probability $1 - p_x$, the device postpones its transmission to the next subframe. The permission probability is updated for each subframe based on two ingredients.

For each type, the gNB first observes each slot of the current subframe and counts the number of holes h (a slot that is not occupied by any transmission(s) is referred to as a hole), successes s (a slot with a single packet transmission), and collisions c (a slot with more than one packet transmissions). Then, it proceeds to estimate the number of packets involved in the transmissions of the current subframe.

Second, the gNB estimates the new arrivals of type x during the current subframe, which together with the backlogged devices will attempt to transmit their packets with an updated permission probability in the next subframe. Backlogged devices are those devices that postpone their transmission in the current subframe due to the imposed permission probability *plus* those devices that were involved in collisions. Furthermore, active devices comprise both backlogged devices and new arrivals. In the next subframe, all active devices will attempt to transmit following the permission probability for each traffic type (details are given in the next subsection).

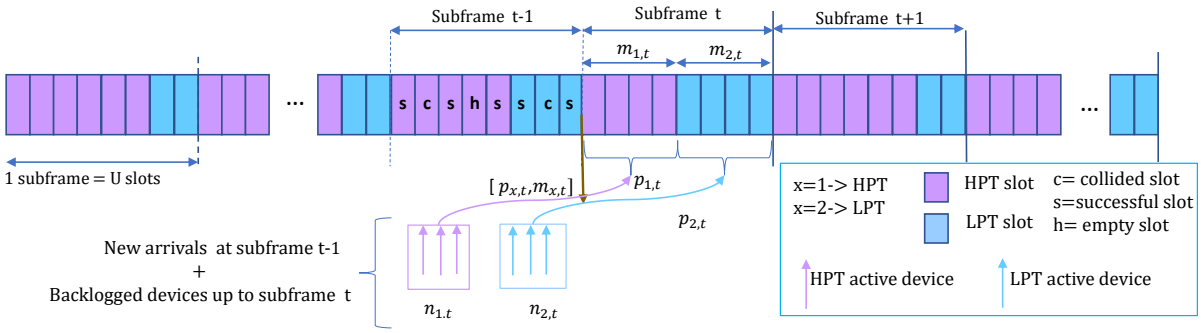


Figure B.2: Illustration of DSA-GF: With priority, the slots are dynamically divided into two groups, one for HPT and the other for LPT.

In DSA-GF, new arrivals follow the immediate first transmission (IFT) principle. By IFT, it is meant that any just-arrived packet in the current subframe will be potentially transmitted in the next immediately available subframe according to the updated permission probability provided by the gNB. Upon the successful reception of a packet transmission, immediate feedback is performed. The operation of the DSA-GF scheme is illustrated in Fig. B.2.

B.3.2 Detailed Access Procedure for Heterogeneous Traffic

Within each subframe, there are a total number of U slots shared by both streams, one from each type of devices. Let m_x denote the number of slots allocated to the HPT ($x = 1$) and LPT ($x = 2$) flows respectively. We have $m_1 + m_2 = U$. The same notations for subscripts apply to other expressions throughout the context. Denote by $u_{1,min}$ ($u_{1,max}$) the minimum (maximum) number of slots that can be allocated to HPT at any subframe, in such a way that $1 \leq u_{1,min} \leq m_1 \leq u_{1,max} < U$. By keeping $u_{1,max} < U$, our scheme reserves *at least one slot per subframe* for LPT so that no starvation happens to LPT regardless of HPT traffic intensity. In what follows, we present how m_x and p_x are updated from subframe to subframe.

During each subframe, the gNB observes what happened in each slot. Let $(h, s, c)_{x,t}$ denote the number of holes, successes and collided slots, respectively, observed for type x during subframe t. Obviously, we have $h_{x,t} + s_{x,t} + c_{x,t} = m_{x,t}$ with $m_{1,t} + m_{2,t} = U$. Furthermore, let $\hat{\lambda}_{x,t}$ be the estimation of new arrivals assessed by the gNB, i.e., the estimated number of devices that have generated a packet during subframe t. Then, the $(m, p)_{x,t} \rightarrow (m, p)_{x,t+1}$ update is performed according to the three steps presented below.
Step 1: Update the estimated number of active devices for HPT and LPT, at the end of subframe t.

- First, for $x = 1, 2$, based on the observations $(h, s, c)_{x,t}$ and the estimated number of active devices at the beginning of subframe t, $w_{x,t}$, the gNB estimates the number of backlogged devices at the end of this subframe t, $\hat{w}_{x,t+1}$. For that purpose, we extend Rivest's pseudo-Bayesian broadcast control algorithm [18] to data transmissions with multiple slots in each subframe so that $\hat{w}_{x,t+1} = w_{x,t} + \frac{c_{x,t}}{e-2} - (h_{x,t} + s_{x,t}) \approx w_{x,t} + 1.3922 \times c_{x,t} - (h_{x,t} + s_{x,t})$. In this expression, $1.3922 \times c_{x,t}$ represents the

estimated number of collided packets which will attempt to transmit again following the rule given in **Step 3**, and $h_{x,t} + s_{x,t}$ represent the idle slots plus successful transmissions that will not retransmit in the next subframe.

- Second, for $x = 1, 2$, the gNB estimates the number of *new arrivals* during subframe t, $\hat{\lambda}_{x,t}$. Considering a network in the steady state where the offered traffic and the carried traffic reach an equilibrium, we set $\hat{\lambda}_{x,t}$ equal to $s_{x,t}$, the number of successes at subframe t. A more elaborated estimator of the arrival process is however out of the focus of this paper.

- Third, the total number of active devices at the end of subframe t ready for transmission at subframe $t+1$ is the sum of $\hat{w}_{x,t+1}$ and $\hat{\lambda}_{x,t}$. Since $w_{x,t+1}$ cannot be negative, we set $w_{x,t+1} = \max(\hat{w}_{x,t+1}, 0) + \hat{\lambda}_{x,t}$. Note that $w_{x,t+1}$ can be any non-negative real number.

Step 2: Update the number of slots to be allocated in the next subframe $t + 1$.

- To give higher priority to HPT, we first allocate $m_{1,t+1} = \max\big(u_{1,min}, \min(\lceil w_{1,t+1} \rceil, u_{1,max})\big)$ and then configure $m_{2,t+1} = U - m_{1,t+1}$. That is, the capacity that is not assigned to HPT will be allocated to LPT. In the above expression, a ceiling function is introduced considering that $m_{x,t+1}$ is an integer number such as $u_{x,min} \leq m_{x,t+1} \leq u_{x,max}$.

Step 3: Update the permission probabilities for subframe $t + 1$ for each type of traffic.

- Set $p_{x,t+1} = \dfrac{m_{x,t+1}}{\max(m_{x,t+1}, w_{x,t+1})} = \min\left(\dfrac{m_{x,t+1}}{w_{x,t+1}}, 1\right)$. That is, when the estimated number of active devices, $w_{x,t+1}$, is greater than the number of allocated slots, $m_{x,t+1}$, the assigned permission probability becomes less than 1. Otherwise, it is 1. Note that for each type of traffic *the same permission probability applies to all active devices.*

B.4 Discrete-time Markov Model for DSA-GF

To evaluate the performance of the proposed DSA-GF scheme for heterogeneous GF traffic, we develop a 2D Markov model which integrates HPT and LPT through a pseudo-aggregated process. During each subframe, every device generates one data packet with probability a_x according to a Bernoulli process. For packet buffering, a packet rejection mechanism is adopted meaning that a packet is rejected when it arrives at a device and finds the buffer full [26].

B.4.1 Building a Discrete-Time Markov Model

Thanks to the memoryless property of the arrival processes, we can build a discrete-time Markov chain for the presented DSA-GF scheme. For this purpose, let us observe the system at the border of two consecutive subframes, e.g., at the time instant when subframe $t-1$ ends (or subframe t begins), where $t \in \mathbb{Z}$ ($\mathbb{Z}$ is the set of integer numbers).

Subframe by subframe, these time instants are regarded as the *transition instants* in the developed Markov model defined by a set of three random variables for each type of traffic. For traffic type x where $x = 1$ (HPT) or $x = 2$ (LPT), let $W_{x,t}$ be the random variable (r.v.) representing the number of active devices estimated by the gNB at transition instant t, $U_{x,t}$ be the r.v representing the number of slots in subframe t allocated to trafic type x, and $N_{x,t}$ be the r.v. representing the actual number of active devices (new arrivals plus backlogged devices) ready for transmission in subframe t.
The transition probabilities of the Markov chain, in a compact format, are as follow.

$$\Pr\big((\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t+1}|(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_t,(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t-1},(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t-2},\dots\big) = \Pr\big((\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_{t+1}|(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_t\big) \quad \text{(B.11)}$$

where $(\boldsymbol{W},\boldsymbol{U},\boldsymbol{N})_t$ denotes $\boldsymbol{W}_t = [W_{1,t}, W_{2,t}]$, $\boldsymbol{U}_t = [U_{1,t}, U_{2,t}]$, and $\boldsymbol{N}_t = [N_{1,t}, N_{2,t}]$. Note that $U_{1,t} + U_{2,t} = U_t = U$. In the expressions presented hereafter, we have introduced a compact expression based on (B.11) with simplified notations $\Pr(W_{x,t} = w_{x,t})$, $\Pr(U_{x,t} = m_{x,t})$, and $\Pr(N_{x,t} = n_{x,t})$ respectively.

It is worth mentioning that the Markov chain defined in (B.11) entails high complexity. In what follows, we opt a lightweight and consistent procedure which consists of the following three phases. 1) Subsec. B.4.2 performs the analysis of HPT since its behavior is independent of that of LPT; 2) Subsec. B.4.3 builds a pseudo-aggregated process which takes into account the correlation or dependence between these two types of traffic for slot allocation in the same subframe; and 3) Subsec. B.4.4 presents the performance of LPT.

B.4.2 The Analysis of High Priority Traffic

B.4.2.1 Modeling the HPT process

Consider that a total number of M_1 devices generate data packets according to a Bernoulli process with probability a_1. Clearly, a Markov chain can be built at the transition instants as defined above. Using (B.11) and omitting the random variables related to the notations of LPT, we have the corresponding transition probabilities, i.e.,

$$\Pr\big((W_1,U_1,N_1)_{t+1} = (\nu,v,j)|(W_1,U_1,N_1)_t = (\mu,u,i)\big) = P_{\mu,i;\nu,j}. \quad \text{(B.12)}$$

In (B.12), the following short notations are used: $(w,m,n)_{1,t} \equiv (\mu,u,i)$ and $(w,m,n)_{1,t+1} \equiv (\nu,v,j)$. For convenience, we restrict the values of $w_{1,t}$ $t \in \mathbb{N}$ to natural numbers (notice that, according to **Step 1** of the DSA-GF scheme, $w_{1,t}$ can be any real number). Such a restriction makes it possible to enumerate or list the states of the Markov chain. Since there exists a deterministic relationship between $u = m_{1,t}$ and $\mu = w_{1,t}$, i.e., $u = \max\big(u_{1,min}, \min(\mu, u_{1,max})\big)$, only two random variables, $W_{1,t}$ and $N_{1,t}$, are sufficient to fully describe this Markov chain. In other words, the Markov chain with three sets of r.v. defined in (B.11) shrinks to a 2D model. Accordingly, the short notation of $P_{\mu,i;\nu,j}$ in (B.12) represents the set of corresponding transition probabilities from subframe t to

subframe $t+1$. For expression clarity in the rest of this subsection, subscript "1" which is meant for HPT is intentionally omitted unless it is explicitly necessary.

Let us first derive the explicit expressions for $P_{\mu,i;\nu,j}$ starting with the transition $\mu \to \nu$. Based on the observations of slots for traffic type 1 during subframe t, i.e., $(h,s,c)_t$ where $h_t + s_t + c_t = u$, the gNB uses a function $f((h,s,c)_t) = c_t/(e-2) - (h_t + s_t)$ to estimate the number of backlogged devices, i.e., $\hat{w}_{t+1} = \mu + f((h,s,c)_t)$ (see **Step 1**: First presented in Subsec. B.3.2). After that and following **Step 1**: Second and **Step 1**: Third, the estimated number of new arrivals during subframe t is taken into account, such that the estimated number of devices active at the beginning of subframe $t+1$ is $\nu = \max(\hat{w}_{t+1}, 0) + \hat{\lambda}_t$.

Although $\mu = w_t$ is set to an integer number, in general, neither $\hat{w}_{t+1}$ nor $\hat{\lambda}_t$ is an integer number. As $\nu = w_{t+1}$ is also set to be an integer number, we introduce the 'ceil' operation such that,

$$\nu = \lceil \max(\mu + f((h,s,c)_t), 0) + \hat{\lambda}_t \rceil; \;\; v = \max\big(u_{1,min}, \min(\nu, u_{1,max})\big); \;\; p_{t+1} = \min\left(\frac{v}{\nu}, 1\right) \quad \text{(B.13)}$$

Note that the updated probability p_{t+1} applies to all active devices at subframe $t+1$ and it is restricted to be a fraction of two integer numbers.

Second, we evaluate the transition probability $i \to j$ referred to in (B.12). For that purpose, we consider in the first step the departure process, i.e., for packets that successfully finished their transmissions during the actual subframe t. At the beginning of subframe t, each of the i active devices chooses to transmit with permission probability p_t or postpone its transmission with probability $1 - p_t$, respectively. Then, the probability that z out of i active devices ($0 \leq z \leq i$) transmit in subframe t follows a binomial distribution, $B_z^i(p_t) = \binom{i}{z} p_t^z (1-p_t)^{i-z}$. With $u = m_{1,t}$, let $R_{s_t,c_t}^{z,u}$ denote the probability that z packets (active devices) intend to access over u slots of subframe t resulting in s_t successful transmissions and c_t collided slots. For any packet transmission, each of the z active devices chooses, with equal probability, one of the u slots of subframe t. Jointly considering these two sequential and independent actions, we obtain the probability that within subframe t, s_t out of i active devices succeed in the transmission of its own packet whereas the other $i - s_t$ devices were involved in collisions or deferred their transmissions. Analytically, it is expressed as,

$$D_{s_t,c_t}^{i,u}(p_t) = \sum_{z=s_t}^{i} B_z^i(p_t) R_{s_t,c_t}^{z,u}. \quad \text{(B.14)}$$

In (B.14), $R_{s_t,c_t}^{z,u}$ can be evaluated using, for instance, the recursions given at [28].

In the second step, we take into account the number of devices that will be active at the transition instant at the end of subframe t. Since the arrival of packets comes from M_1 sources each one with activation probability a_1, the arrival process follows a binomial distribution. Jointly considering the departure and arrival processes, which are independent of each other, we have,

$$P_{\mu,i;\nu,j} = \sum_{(h,s,c)_t \in \Omega_1} D^{i,u}_{s_t,c_t}(p_t) A^{M_1-i+s_t}_{j-i+s_t}(a_1), \tag{B.15}$$

where $A^{M_1-i+s}_{j-i+s}(a_1)$ follows the binomial distribution, as $A^k_l(a_1) = B^k_l(a_1) = \binom{k}{l} a_1^l (1-a_1)^{k-l}$. The set Ω_1 defined in (B.15) represents the set of $(h,s,c)_t$ values observed in subframe t that satisfy the following two conditions,

$$\Omega_1 \stackrel{\text{def}}{\equiv} \begin{cases} u = h_t + s_t + c_t & = \max\big(u_{1,min}, \min(\mu, u_{1,max})\big); \\ \nu & = \lceil \max(\mu + f((h,s,c)_t), 0) + \hat{\lambda}_t \rceil. \end{cases} \tag{B.16}$$

Then, the solution in the steady state regime is given by the stochastic row vector $\boldsymbol{\pi}$ ($\boldsymbol{\pi}\mathbf{e} = \mathbf{1}$) which can be obtained from the linear equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ with $\boldsymbol{\pi} = \{\pi_{\mu,i}\}, \mathbf{P} = \{P_{\mu,i;\nu,j}\}$. Here, $\mathbf{e}$ is a column vector of all 1's, $\pi_{\mu,i}$ is the probability that at the start of an arbitrary subframe the number of active devices estimated by the gNB is μ and the actual number of active devices is i.

B.4.2.2 Throughput, access delay, and packet loss probability for HPT

Based on $\boldsymbol{\pi}$, we derive below expressions for the performance of HPT in terms of four parameters as defined below.

Firstly, the mean value of the number of successfully transmitted packets *within one subframe*, defined as throughput per subframe, is obtained according to,

$$\gamma_1^{sf} = \sum_{s_t=u_{1,min}}^{u_{1,max}} s_t \sum_{(\mu,i)\in\Delta} \pi_{\mu,i} \sum_{c_t \in \mathcal{C}} D^{i,u}_{s_t,c_t}(p_t) = \sum_{(\mu,i)\in\Delta} i p_t \left(1 - \frac{p_t}{u}\right)^{i-1} \pi_{\mu,i}. \tag{B.17}$$

In (B.17), the set Δ shown as $(\mu,i) \in \Delta$ contains all possible values in μ and i and the set $\mathcal{C}$ shown as $c_t \in \mathcal{C}$ contains all possible collided slots such that $h_t + s_t + c_t = u$. Observe that the relationship between $\mu = w_t$ and $u = m_t$ is given in (B.16). The second equality is obtained after some algebraic operations and the details are omitted for the sake of brevity. Instead, a short clue is outlined as follows. Using DSA-GF, the expected number of successful transmissions when i active devices access to a set of u slots with permission probability $p_t = \min(u/\mu, 1)$ is given by,

$$E\big(\text{success} | (N_t = i, U_t = u, p_t = p = \min(u/\mu, 1)\big) = ip\left(1 - \frac{p}{u}\right)^{i-1}. \tag{B.18}$$

Then, the last equality in (B.17) is a weighted sum of (B.18) with probabilities $\pi_{\mu,i}$.

To give further insights on HPT performance in terms of resource utilization, how long a packet has to stay in a buffer, and how likely a packet may get lost, we define three other parameters. The mean value of the number of successfully transmitted packets *within one slot*, i.e., throughput per slot, which represents resource utilization is obtained based on (B.17),

$$\gamma_1^{slot} = \gamma_1^{sf} / \sum_{(\mu,i)\in\Delta} u\pi_{\mu,i}. \tag{B.19}$$

Thirdly, access delay in this study, d_1^{sf}, is defined as the mean sojourn time a packet stays in a buffer until it is successfully transmitted. Using Little's formula, the average number of customers in our steady state system (which is the mean number of active devices at the beginning of an arbitrary subframe, obtained by $\sum_{(\mu,i)\in\Delta} i\pi_{\mu,i}$) equals to d_1^{sf} multiplied by the average arrival rate (which is the average number of successful transmissions per subframe, γ_1^{sf}). Therefore, we have

$$d_1^{sf} = \sum_{(\mu,i)\in\Delta} i\pi_{\mu,i}/\gamma_1^{sf}. \tag{B.20}$$

The fourth performance parameter, packet loss probability, is defined as the ratio of the rejected, i.e., offered minus carried traffic, to the offered traffic. For HPT, it is expressed as

$$\theta_1 = (M_1 a_1 - \gamma_1^{sf})/M_1 a_1. \tag{B.21}$$

B.4.3 Linking HPT and LPT with a Pseudo-Aggregated Process

Based on the 2D Markov chain that models the HPT behavior, denoted as X, we construct a tailored pseudo-aggregated process that links the HPT process with the LPT process.

Inspired by the procedure presented in [27], we make a partition of the states of the Markov chain X. Let $E = \{(\mu, i)\}$ be the set of states of our initial Markov chain X, where ($1 \le \mu \le E_{1,max}$, $0 \le i \le M_1$). It is understood that $E_{1,max}$ represents the maximum number of active HPT devices that the gNB can estimate. Let us sort the set of states into the following order,

$$\begin{aligned}\mathcal{E}_\mu = \{(\mu,0),(\mu,1),\ldots,(\mu,M_1)\}; \qquad & \mu = 1,2,\ldots,u_{1,min}-1,u_{1,min},\\ & u_{1,min}+1,\ldots,u_{1,max}-1,u_{1,max},u_{1,max}+1,\ldots,E_{1,max}-1,E_{1,max}.\end{aligned}$$

Now, let $\mathfrak{F} = \{\mathcal{F}(u_{1,min}),\ldots,\mathcal{F}(u_{1,max})\}$ be a partition of E such that

$$\begin{aligned}&\mathcal{F}(u_{1,min}) = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_{u_{1,min}}; \qquad \mathcal{F}(\mu) = \mathcal{E}_\mu, \quad u_{1,min} < \mu < u_{1,max};\\ &\mathcal{F}(u_{1,max}) = \mathcal{E}_{u_{1,max}} \cup \mathcal{E}_{u_{1,max}+1} \cup \cdots \cup \mathcal{E}_{E_{1,max}}.\end{aligned}$$

Let F be the set of integers $\{u_{1,min}, ..., u_{1,max}\}$. Based on the initial Markov chain X with known values on E, we associate the pseudo-aggregated Markov chain Y with potential values on F, defined by: $Y_t = m \iff X_t \in \mathcal{F}(m)$ for all values of $t \in \mathbb{Z}$. Observe that, due to the mapping procedure, the pseudo-aggregated process includes the statistics of the number of slots allocated to HPT devices in the same subframe. Then, the transition probabilities of the pseudo-aggregated Markov chain Y are given as follows,

$$\hat{P}_{u,v} \stackrel{\text{def}}{\equiv} \sum_{i\in\mathcal{F}(\mu)} \left(\pi_{\mu,i} / \sum_{h\in\mathcal{F}(\mu)} \pi_{\mu,h}\right) \sum_{j\in\mathcal{F}(\nu)} P_{\mu,i;\nu,j}; \tag{B.22}$$

where $u = \max\big(u_{1,min}, \min(\mu, u_{1,max})\big)$ and $v = \max\big(u_{1,min}, \min(\nu, u_{1,max})\big)$.

Clearly, the probabilities $\hat{P}_{u,v}$ for $u_{1,min} \leq u, v \leq u_{1,max}$ constitute the Markov chain that counts the number of slots per subframe allocated to HPT devices. The Markov chain defined by (B.22) preserves the mean values (sojourn times in each set of state) of the original process, but in general higher statistical moments of these two processes are different from each other.

By solving the linear equation $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}\hat{\mathbf{P}}$ with $\hat{\boldsymbol{\pi}} = \{\hat{\pi}_u\}$ and $\hat{\mathbf{P}} = \{\hat{P}_{u,v}\}$, the stochastic vector $\hat{\boldsymbol{\pi}}$ ($\hat{\boldsymbol{\pi}}\mathbf{e} = \mathbf{1}$) is obtained. Accordingly, the statistics of the r.v. number of slots allocated per frame for HPT can be easily obtained. This pseudo-aggregated Markov chain provides a link between HPT and LPT. This link will be used to analyze the performance of LPT as presented next.

B.4.4 The Analysis of Low Priority Traffic

B.4.4.1 Modeling the LPT process

The analysis of LPT can be derived in a similar and parallel way as its HPT counterpart. The main difference is that the number of slots per subframe allocated to LPT is dictated by the dynamic behavior of the HPT occurring in the same subframe. A link between both types of traffic is established based on the pseudo-aggregated process defined above, hence simplifying the analysis of LPT. Intuitively, this approach could loose the "synchronization" or the existing coupling between HPT and LPT. However, the rationale behind our analysis lies on the fact that this approach largely captures the behavior of LPT, which utilizes the remaining capacity, i.e., a number of slots in the same subframe that are not allocated to the HPT transmissions.

Correspondingly, in a parallel way to (B.12) and omitting the r.v. of HPT, we have

$$\Pr\big((W_2, U_2, N_2)_{t+1} = (\nu, v, j) | (W_2, U_2, N_2)_t = (\mu, u, i)\big) = P_{\mu,u,j;\nu,v,j}. \tag{B.23}$$

In (B.23), the same notations as in (B.12) have been introduced, *but now it is referred to LPT*, i.e., $(w, m, n)_{2,t} \equiv (\mu, u, i)$ and $(w, m, n)_{2,t+1} \equiv (\nu, v, j)$. However, the difference between (B.12) and (B.23) is that in the LPT case the transitions $w_{2,t} \equiv \mu \to w_{2,t+1} \equiv \nu$ and $m_{2,t} \equiv u \to m_{2,t+1} \equiv v$ evolve independently of each other, whereas the second transition is dictated by the behavior of the HPT process. Accordingly, in contrast to the HPT process which is represented by 2 random variables, 3 random variables are needed to identify the Markov chain of the LPT process.

The evaluation of (B.23) is similar to the counterpart model of HPT. First, for the transition $(\mu, i) \to (\nu, j)$, we consider the packets that have been successfully transmitted, i.e., the departure process (see (B.14)).

$$D^{i,u}_{s_t,c_t}(p_t) = \sum_{z=s_t}^{i} B^i_z(p_t) R^{z,u}_{s_t,c_t}, \tag{B.24}$$

where $R^{z,u}_{s_t,c_t}$ has the same meaning as in (B.14). Furthermore, in parallel to (B.15), we have the following expression for LPT.

$$P_{\mu,i;\nu,j} = \sum_{(h,s,c)_t \in \Omega_2} D^{i,u}_{s_t,c_t}(p_t) A^{M_2-i+s_t}_{j-i+s_t}(a_2), \tag{B.25}$$

where $A^{M_2-i+s}_{j-i+s}(a_2)$ follows the binomial distribution similar to the one presented in (B.15) but for LPT. In (B.25), the set Ω_2 defined as $(h,s,c)_t \in \Omega_2$ is the set of values that satisfy the following two conditions,

$$\Omega_2 \overset{\text{def}}{\equiv} \begin{cases} u = h_t + s_t + c_t & \neq \max\big(u_{2,min}, \min(\mu, u_{2,max})\big); \\ \nu & = \lceil \max(\mu + f((h,s,c)_t), 0) + \hat{\lambda}_t \rceil. \end{cases} \tag{B.26}$$

To gain clarity in the rest of this paragraph, we have recovered the notations with subscripts in $m_{x,t}$ and $w_{x,t}$ where $x = 1$ for HPT and $x = 2$ for LPT, respectively. Consider that, at subframe t, we have $m_{2,t} = U - m_{1,t}$. In the next subframe $t+1$, the gNB will allocate $m_{2,t+1} = U - m_{1,t+1}$ with probability $\hat{P}_{m_{1,t},m_{1,t+1}}$ given by (B.22), i.e., by the transition probabilities of the pseudo-aggregated Markov chain. In other words, the number of slots per subframe allocated to LPT by the gNB in the next subframe $t+1$ only depends on the transitions $m_{1,t} \to m_{1,t+1}$ of HPT. We highlight this fact with the inequality of (B.26). Then, the equivalent expression of (B.13) for LPT devices becomes,

$$\begin{aligned} &\nu = \lceil \max(w_{2,t} + f((h,s,c)_{2,t}), 0) + \hat{\lambda}_{2,t} \rceil; \qquad v = m_{2,t+1} = U - m_{1,t+1}; \\ &p_{2,t+1} = \min\left(\frac{m_{2,t+1}}{w_{2,t+1}}, 1\right) = \min\left(\frac{v}{\nu}, 1\right). \end{aligned} \tag{B.27}$$

By combining (B.25) with the transition probabilities (B.22) of the pseudo-aggregated Markov chain, we obtain the transition probabilities corresponding to the Markov chain for LPT,

$$P_{\mu,u,i;\nu,v,j} = P_{\mu,i;\nu,j} \hat{P}_{U-u,U-v}. \tag{B.28}$$

Note that $P_{\mu,i;\nu,j}$ in (B.28) refers to (B.25), i.e., it is meant for LPT and it differs from (B.15) which refers to HPT. Through (B.28), we claim that 1) the product of both probabilities reflects the 'independence' in the treatment of both types of traffic; and 2) the correlation or dependence between HPT and LPT is taken into account with the transition probabilities (B.22) of the pseudo-aggregated Markov chain.

The steady state regime for LPT is given by the stochastic row vector $\boldsymbol{\pi}$ ($\boldsymbol{\pi}\mathbf{e} = \mathbf{1}$) derived by solving the linear equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ with $\boldsymbol{\pi} = \{\pi_{\mu,u,i}\}$ and $\mathbf{P} = \{P_{\mu,u,i;\nu,v,j}\}$. Then, $\pi_{\mu,u,i}$ is the steady state probability at the start of an arbitrary subframe of the number of active devices estimated by the gNB being μ, the number of active devices being i, and the slot allocated to the LPT flows being u.

B.4.4.2 Throughput, access delay, and packet loss probability for LPT

Similar to the HPT case, we assess the performance of LPT with respect to the same four parameters defined above. In particular, the throughput per subframe for LPT is obtained as follows

$$\gamma_2^{sf} = \sum_{s_t=1}^{u_{2,max}} s_t \sum_{(\mu,u,i)\in\Delta} \pi_{\mu,u,i} \sum_{c_t\in\mathcal{C}} D_{s_t,c_t}^{i,u}(p_t) = \sum_{(\mu,u,i)\in\Delta} i p_t \left(1-\frac{p_t}{u}\right)^{i-1} \pi_{\mu,u,i}. \tag{B.29}$$

The second equality in (B.29) is derived in the same way as what is deduced in (B.17).

In a similar way as for (B.19), the expression of the throughput per slot for LPT is obtained as

$$\gamma_2^{slot} = \gamma_2^{sf} / \sum_{(\mu,i)\in\Delta} u\pi_{\mu,i}. \tag{B.30}$$

Similar to (B.20), the access delay for LPT is obtained by

$$d_2^{sf} = \sum_{(\mu,i)\in\Delta} i\pi_{\mu,i}/\gamma_2^{sf}. \tag{B.31}$$

Lastly, similar to the expression in (B.21), the packet loss probability for LPT is defined as

$$\theta_2 = (M_2 a_2 - \gamma_2^{sf})/M_2 a_2. \tag{B.32}$$

B.5 Simulations and Numerical Results

This section presents the numerical results obtained from both the analytical model and discrete-event based simulations. The proposed DSA-GF scheme has been implemented based on a custom-built simulator constructed in MATLAB which mimics the behavior of the scheme. Extensive simulations are performed under various configurations. The results with respect to the four performance parameters defined in Sec. B.4, i.e., throughput per subframe/slot (in terms of number of packets per subframe/slot), access delay for the successfully transmitted packets, and packet loss probability, are presented below. Two other GF access schemes, known as *complete sharing* and *GF reactive* (see Subsec. B.5.5), have also been implemented and the performance of these three schemes is compared with each other therein. The applicability of DSA-GF to two other numerologies is validated in Subsec. B.5.6.

B.5.1 Simulation Setup and Model Validation

Consider an NR cell with all three GF schemes, i.e., DSA-GF, complete sharing, and GF reactive, enabled. As mentioned earlier in Subsec. B.2.2, although the total number of mMTC devices covered by the cell could be large, the number of devices attempting for

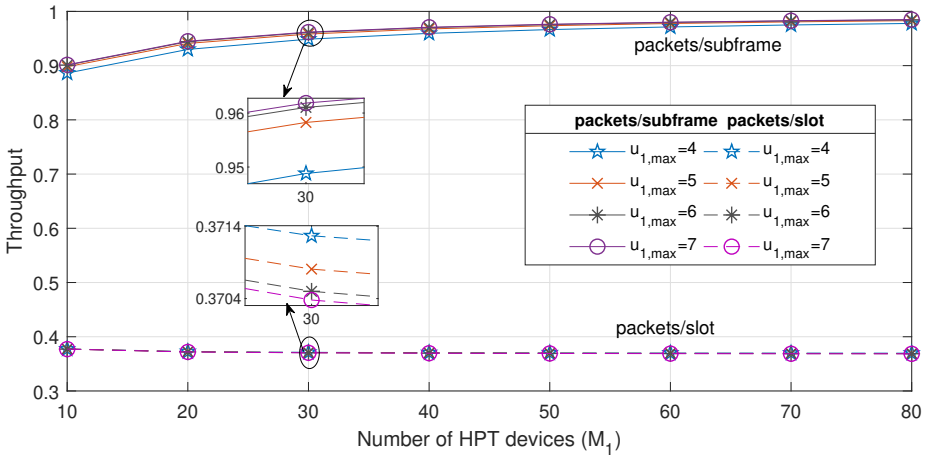


Figure B.3: Throughput of HPT when $M_1 a_1 = 1$ and $u_{1,min} = 1$, $u_{1,max} = 4, 5, 6, 7$.

channel access at a given subframe is considered to be rather limited. In this study, we consider that the device population for each type varies from 10 up to 100, coupled with different activation probabilities. The offered traffic intensities are represented by $M_1 a_1$ and $M_2 a_2$ (in terms of packets per subframe) for HPT and LPT, respectively. Except Subsec. B.5.6 which considers numerology $\beta = 2$ and $\beta = 4$, we adopt $\beta = 3$ for our performance evaluations in all simulations presented below. Note that no matter there are $U = 4, 8$, or 16 slots in each subframe, all of them are available for GF transmissions (discussed in Subsec. B.2.1). For these simulations, we set $u_{1,max} \leq U - 1$. That is, $u_{2,min} \geq 1$. The other parameters like $u_{1,min}$ are configured in favor of HPT performance with the concrete values shown in each figure caption or the corresponding explanations. For all simulation results presented below, we report the average values obtained from multiple runs of simulations.

The accuracy of the developed Markov model is verified through extensive simulations. Under all network configurations, the analytical and simulation results coincide with each other so tightly that the curves obtained from these two methods are largely overlapping. As such, the accuracy of the developed Markov model is validated. As two examples, we plot separately in Figs. B.4 and B.5 the curves obtained from both analysis and simulations. For the sake of illustration clarity, we do not plot both sets of results in other figures.

B.5.2 HPT Performance with Variable Device Population

As explained earlier, the performance of HPT is independent of that of LPT. Accordingly, we evaluate the performance of HPT by varying the number of HPT devices M_1 and the activation probability a_1 while keeping the offered traffic constant as $M_1 a_1 = 1$. Keep in mind that the actual number of slots allocated to HPT per subframe is governed by the DSA-GF scheme where both $u_{1,min}$ and $u_{1,max}$ are tunable parameters but they do not vary on a subframe or frame basis.

The performance of HPT in terms of throughput per slot and per subframe is illustrated in Fig. B.3 where $u_{1,min} = 1$ and $u_{1,max} = 4, 5, 6$, or 7 respectively. It is clear that

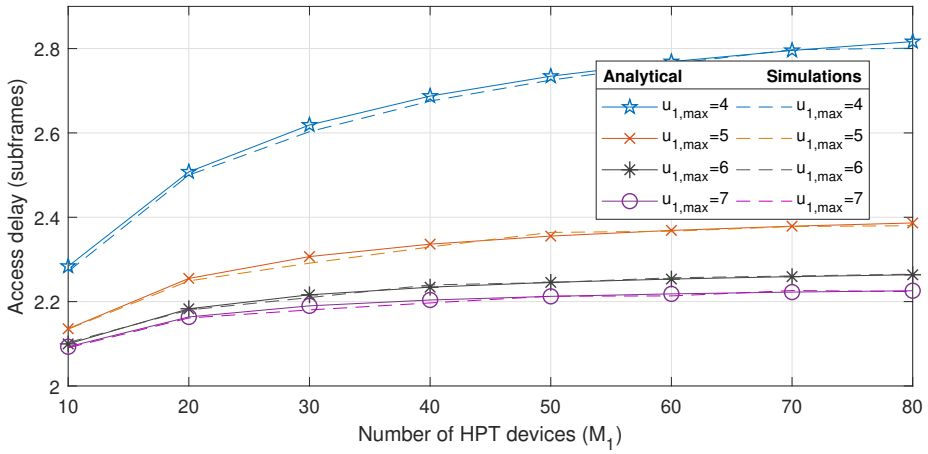


Figure B.4: Access delay of HPT when $M_1 a_1 = 1$ and $u_{1,min} = 1, u_{1,max} = 4, 5, 6, 7$.

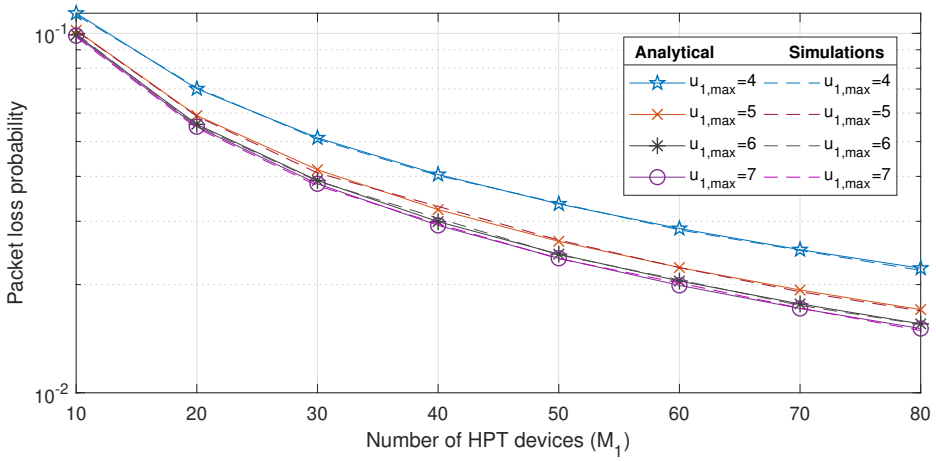


Figure B.5: Packet loss probability of HPT when $M_1 a_1 = 1$ and $u_{1,min} = 1, u_{1,max} = 4, 5, 6, 7$.

the achieved throughput per slot for these configured $u_{1,min}$ and $u_{1,max}$ values is slightly higher than the maximum throughput for slotted ALOHA i.e., $1/e \approx 0.3679$, which is obtained with an infinite population. This is because the number of devices in our simulations is finite. For instance, for a fixed value of $M_1 = (10, 20, 30, ..., 70, 80)$, the resulting successful probability takes the values as $(0.3874, 0.3774, 0.3741, ..., 0.3705, 0.3701)$ respectively, indicating a slightly lower successful probability which approaches the *throughput per slot* for slotted ALOHA as M_1 increases.

On the other hand, we observe that, as M_1 becomes larger, 1) the achieved throughput per subframe increases monotonically towards a maximum value and 2) these values are much higher than that of the throughput per slot. For 1), note first that when a collision occurs, the corresponding packet remains pending to the next subframe. This behavior contributes to an increased number of packets awaiting to be transmitted. Furthermore, the devices that succeeded in the current subframe will also generate with probability a_1 one packet ready for transmission in the next subframe. The net effect is that, when M_1 increases, the mean value of the number of backlogged packets increases slightly and more

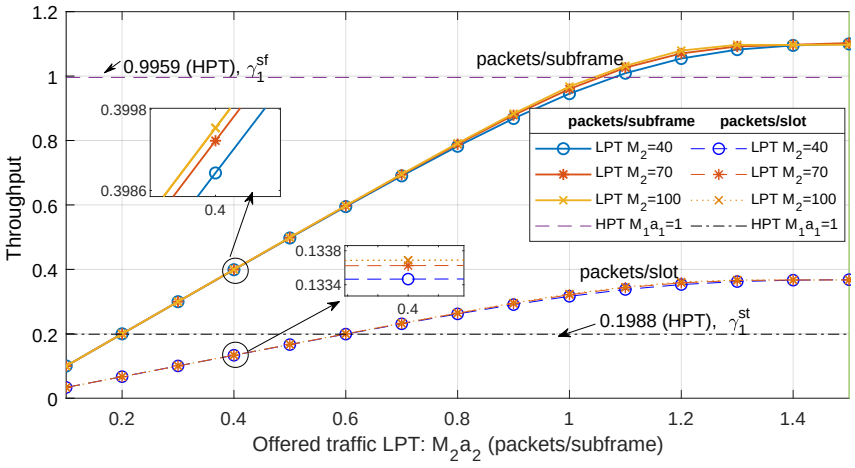


Figure B.6: Throughput per subframe and per slot for LPT under various offered traffic M_2a_2 where $M_1a_1 = 1$ and $M_1 = 100$, $u_{1,min} = 5, u_{1,max} = 7$.

slots will be allocated to HPT, resulting in thus higher throughput per subframe. For 2), it is because multiple slots within the same subframe are utilized by HPT devices. For example, when $(u_{1,min}, u_{1,max}) = (1,4), (1,5), (1,6)$, or $(1,7)$ and $M_1 = 30$, there are on average 2.5555, 2.5842, 2.5939, or 2.5969 number of slots allocated to HPT respectively. Indeed, this result is in accordance with the relationship between per subframe and per slot throughput expressed in (B.19).

Furthermore, the obtained access delay and packet loss probability performance is depicted in Figs. B.4 and B.5 respectively. With a larger device population, DSA-GF needs more subframes to accommodate HPT packets, leading to an increasing trend for access delay. On the other hand, the achieved access delay decreases significantly with a larger $u_{1,max}$ due to the fact that more slots are available for HPT. With a larger $u_{1,max}$ value, a competing device obtains a higher probability of selecting a unique slot for successful transmission, resulting in a lower delay. In Fig. B.5, it is shown that a larger $u_{1,max}$ leads to a lower loss probability. With a lager number of HPT devices, the activation probability decreases in order to maintain constant offered traffic. Hence, the impact of buffer limitation is reduced. Correspondingly, the packet loss probability decreases with an increasing M_1. Moreover, one may notice a decreasing gap between two adjacent curves in these two figures with a larger $u_{1,max}$. This is because the performance acceleration rate declines when $u_{1,max}$ increases.

B.5.3 LPT Performance with Variable Offered Traffic

To evaluate the performance of LPT devices, we vary the offered traffic load by LPT devices M_2a_2 given that $M_1a_1 = 1$ with $M_1 = 100$ and $(u_{1,min}, u_{1,max}) = (5,7)$. Under such traffic conditions, the average number of slots allocated to HPT is $\overline{m}_{1,t} \approx 5.0115$. Accordingly, LPT obtains $\overline{m}_{2,t} \approx 2.9885$ slots on average. Figs. B.6-B.8 illustrate the performance in terms of the four parameters defined above.

Fig. B.6 illustrates the obtained throughput per subframe/slot for LPT as M_2a_2 in-

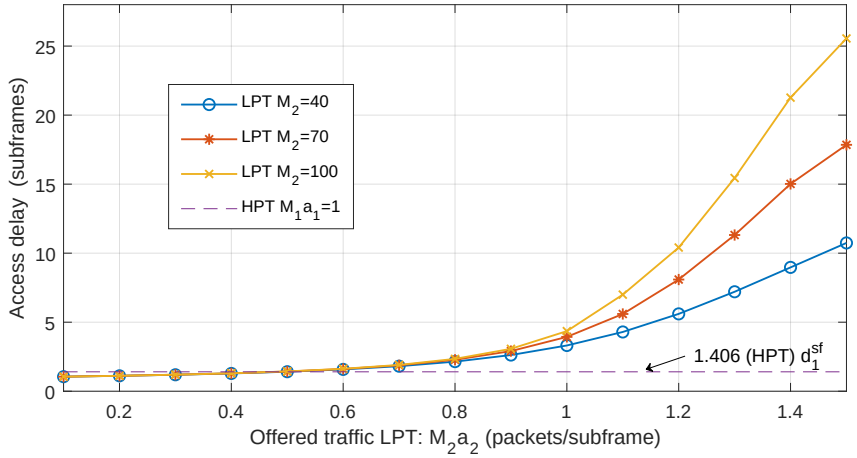


Figure B.7: Access delay for LPT under various offered traffic M_2a_2 where $M_1a_1 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

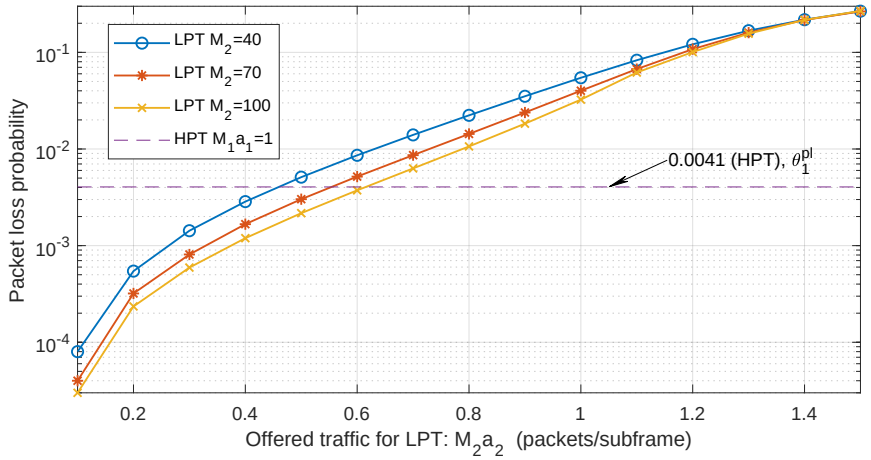


Figure B.8: Packet loss probability for $M_2 = 40, 70, 100$ and various M_2a_2 values in LPT where $M_1a_1 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

creases. Initially, the throughput per subframe increases linearly with M_2a_2 and gradually the behavior reaches a stable limit when the network approaches saturation. A similar trend is observed for the behavior of throughput per slot. The reason is as follows. Since our scheme follows the principle of ALOHA, the highest throughput that can be achieved is $\overline{m}_{2,t}/e = 2.9985/2.7183 = 1.1031$. Therefore, as long as $M_2a_2 < 1.1031$, LPT will exhibit a linear throughput response corresponding to the offered LPT traffic load. The more we increase the offered traffic M_2a_2, the closer we are approaching to the theoretical limit. When M_2a_2 approaches the value of 1.1031, the curve starts to bend and in an asymptotic way it reaches the maximum throughput value.

Fig. B.7 reveals the access delay for successful LPT packet transmissions. When the LPT traffic load increases, a higher number of collisions occur, causing packets to wait for a longer period of time in the buffer. Accordingly, the average delay increases. Recall that devices are equipped with a buffer of unit size. When a new packet arrives and finds the buffer full, it is rejected. This is indeed the implementation of the packet rejection

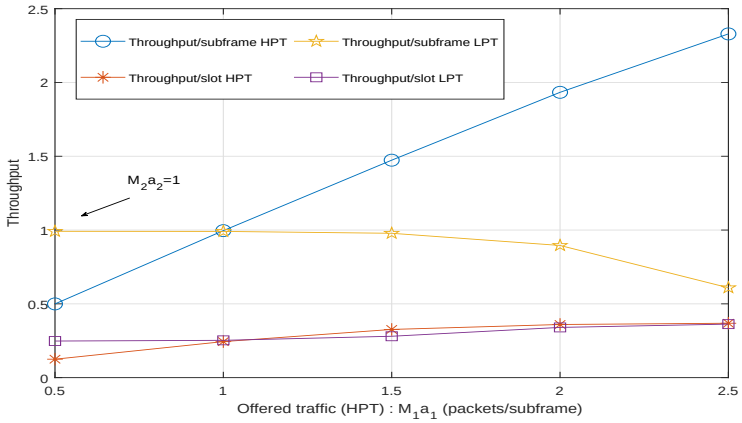


Figure B.9: Throughput per subframe and per slot for HPT/LPT under various offered HPT traffic M_1a_1 where $M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

mechanism [26]. It causes a higher packet loss probability when the offered LPT traffic increases, as shown in Fig. B.8.

Although a loss probability higher than 1% is out of interest, it is worth studying the asymptotic behavior of the loss probability for LPT, i.e., when $a_2 \to 1$. Under the principle of blocking a new packet when the buffer is occupied, the asymptotic loss probability can be expressed as the fraction $(M_2 - \overline{m}_{2,t}e^{-1})/M_2$. It becomes 0.9725, 0.9843, and 0.9989 for $M_2 = 40, 70$, and 100 devices, respectively. These values match perfectly the results provided by the Markov model. Furthermore, due to the introduction of a single size buffer, the asymptotic behavior of the delay performance can be derived as follows. For a given number of LPT devices M_2, when $a_2 \to 1$ (which is the condition for saturation), all M_2 buffers are full, each with one packet ready for transmission at the beginning of each subframe. Since the mean number of successful transmissions per subframe is given by $\overline{m}_{2,t}e^{-1}$, the mean number of subframes that a given packet has to wait in its buffer is given by $M_2e/\overline{m}_{2,t}$. Following the same illustrative example given above with $M_2 = 40, 70$, and 100 LPT devices, the obtained access delay becomes 36.3832, 63.6706, and 90.9581 subframes, respectively. The same as above, these results are in precise agreement with the ones obtained from the Markov model, as expressed in (B.31).

B.5.4 Impact of Offered HPT Traffic Load on HPT/LPT Performance

To assess the impact of offered HPT traffic load on the performance of both HPT and LPT, we perform two sets of simulations, with a combination of constant or variable traffic loads for HPT or LPT respectively. As already discussed above, the performance of HPT remains constant throughout the whole range of the M_2a_2 variations. In other words, these results confirm that HPT's performance remains intact regardless of the variations of the injected LPT traffic load.

On the other hand, LPT's performance will be dominated by HPT traffic intensity since LPT can only occupy the remaining slots in the same subframe that are not allocated

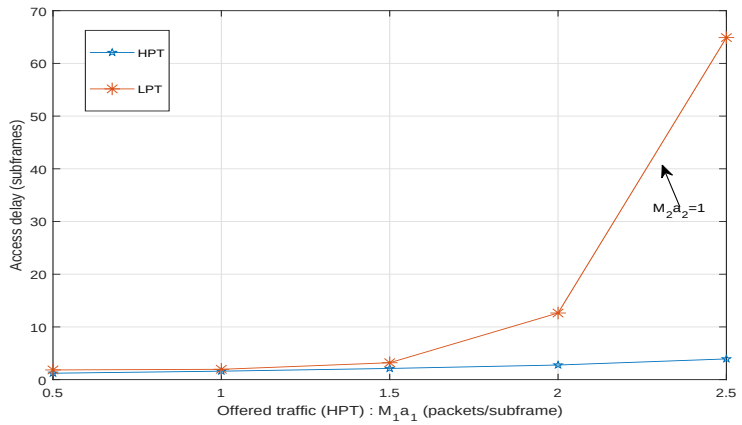


Figure B.10: Access delay for HPT/LPT under various offered HPT traffic M_1a_1 where $M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

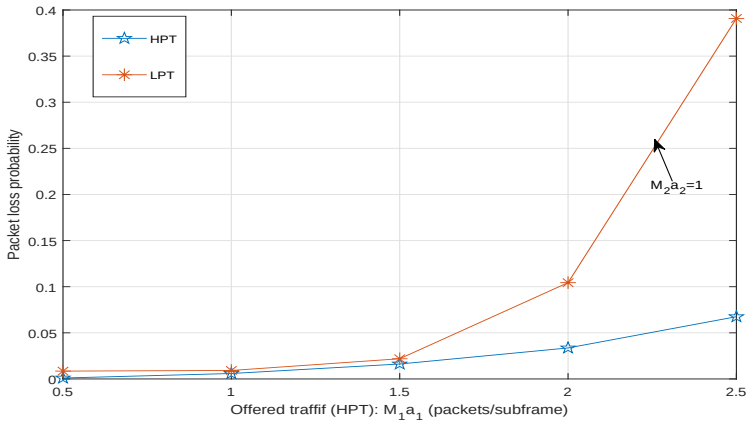


Figure B.11: Packet loss probability for HPT/LPT under various offered HPT traffic M_1a_1 where $M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$.

to HPT packets. As shown in Fig. B.9, while the HPT throughput per subframe increases linearly as M_1a_1 increases (until the saturation point), the LPT throughput per subframe has to sacrifice its performance. With respect to the performance of DSA-GF in terms of access delay and packet loss probability shown in Figs. B.10-B.11, it is convincing that HPT achieves better performance than LPT does.

B.5.5 Performance Comparison with Complete Sharing and GF Reactive

First of all, note that no traffic classification is introduced in these two reference schemes. Before presenting the results, let us outline briefly the principles of these two schemes as follows. 1) Complete sharing works similarly as the proposed scheme. However, the slot allocation and data transmission process in complete sharing does not enable any priorities. Instead of treating HPT and LPT separately, a single class of arrivals will compete for access in all available slots in each subframe. The packet transmission probability is dy-

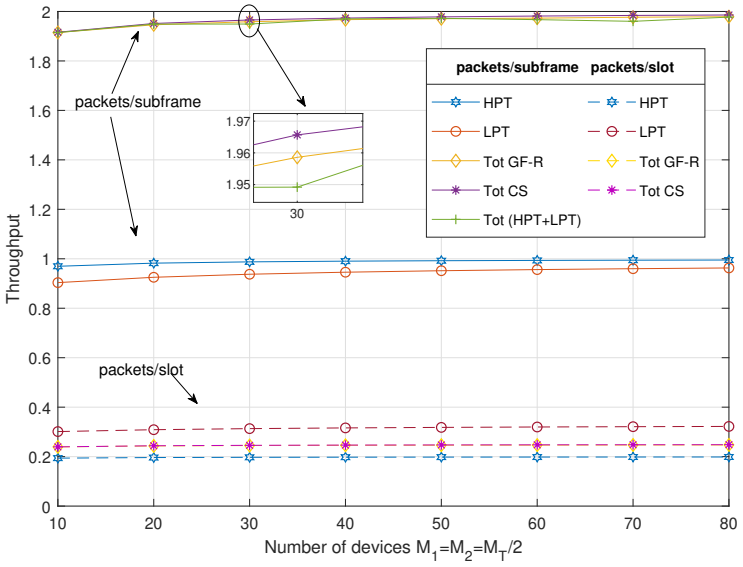


Figure B.12: Throughput comparison with GF reactive and complete sharing ($M_1 a_1 = M_2 a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$).

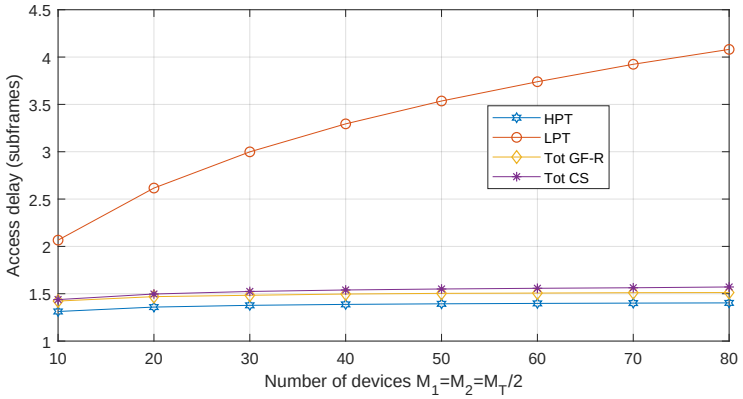


Figure B.13: Access delay comparison with GF reactive and complete sharing ($M_1 a_1 = M_2 a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$).

namically adjusted on a subframe-by-subframe basis following the same pseudo-Bayesian estimation process. 2) The GF reactive scheme discussed in Subsec. B.1.1. No permission probability exists in this scheme, i.e., a failed transmission attempt shall for sure try again in the next subframe. To avoid the situation that an 'unlucky' packet could attempt to transmit forever, a retry limit of 10 is configured in our simulations for GF reactive. In this study, we do not include any proactive GF scheme due to the consideration that high collision could occur for GF proactive with $K > 1$ since two or more packet replicas from the same device will compete for slot access inside the same subframe in GF proactive schemes.

The numerical results obtained from the three studied schemes are compared in Figs. B.12-B.14 where GF-R and CS in the legends stand for GF reactive and complete sharing, respectively. With respect to the achieved throughput per subframe, the values obtained

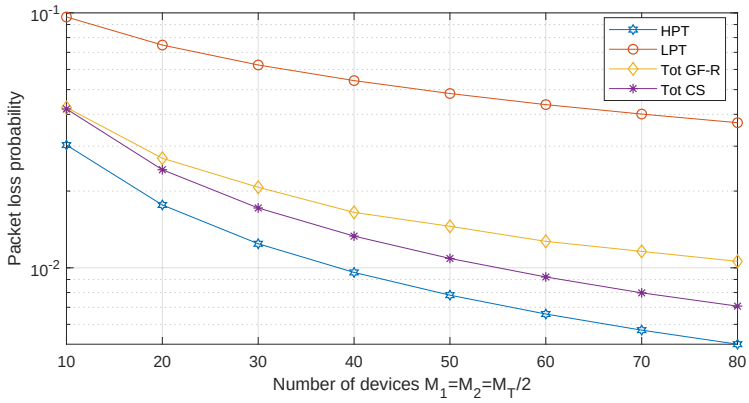


Figure B.14: Packet loss probability comparison with GF reactive and complete sharing ($M_1a_1 = M_2a_2 = 1$, $u_{1,min} = 5, u_{1,max} = 7$).

from all three schemes (for DSA-GF, it is meant for the sum of HPT and LPT throughput) are very close to each other (the curves for throughput per slot for GF-R and CS are indeed overlapping). This is because the offered traffic in all cases is high enough so that the highest slot utilization has been reached. Thanks to the privilege given to HPT with $(u_{1,min}, u_{1,max}) = (5, 7)$, the throughput per subframe for HPT exhibits the highest values, at the cost of reduced LPT throughput.

When it turns to access delay, one may observe a similar trend. That is, HPT achieves the lowest delay across the whole range of device populations, obtained after a small sacrifice of LPT's delay. On the other hand, the reason that GF reactive reaches lower access delay than complete sharing does is that more access opportunities are given to GF reactive devices due to the fact that there is no permission probability as well as the constraint of the retry limit.

Let us now compare the performance in terms of packet loss probability for those three schemes. It is convincing that HPT under the DSA-GF scheme achieves the lowest packet loss probability thanks to its access privilege. This result reveals once again the benefit brought by introducing priority for dynamic slot allocation. On the other hand, when comparing the packet loss probabilities for complete sharing and GF reactive, the results meet our intuition that complete sharing performs better. This is because complete sharing imposes access control via a permission probability when collisions are detected in the previous subframe, thus limiting the number of competing devices in the current subframe. Given that the number of slots in each subframe is fixed, the lesser the number of competing devices, the lower the packet loss.

B.5.6 Applicability of DSA-GF to Numerology $\beta = 2$ and $\beta = 4$

Considering that the subframe duration is fixed for all numerologies as 1 ms, we keep *the offered traffic per subframe constant*, however, with different combinations of device populations and activation probabilities. More specifically, for $\beta = 4$, we configure four sets of device populations as $M_1 = M_2 = 40, 60, 80$, and 100 for HPT and LPT, each set coupled with an activation probability of $a_1 = a_2 = 1/20, 1/30, 1/40$, and $1/50$ respec-

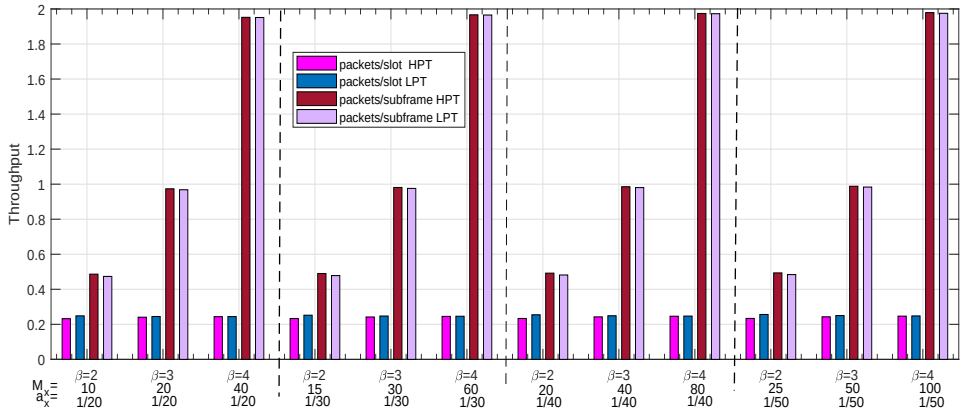


Figure B.15: Applying DSA-GF to three numerologies, $\beta = 2$, 3, and 4: Throughput per subframe/slot.

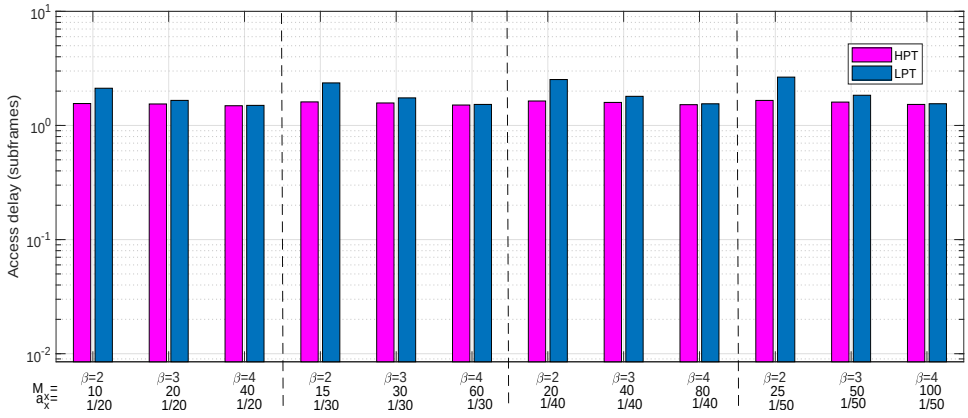


Figure B.16: Applying DSA-GF to three numerologies, $\beta = 2$, 3, and 4 respectively: Access Delay.

tively. In this way, the offered traffic per subframe equals to $M_x a_x = 2$ for each type of traffic, i.e., 2 packets/subframe. For $\beta = 3$ and 2, devices are split into 2 and 4 groups respectively due to resource allocation explained in the next paragraph. Accordingly, we have $M_x a_x = 1$ and 0.5 per subframe for $\beta = 3$ and 2, since there are 2 and 4 parallel subframes respectively. Detailed configurations on M_x and a_x for each numerology can be found in Figs. B.15-B.17.

To accommodate the offered traffic, the gNB can either allocate one subframe for $\beta = 4$, or two parallel subframes over the frequency domain for $\beta = 3$. This configuration is reasonable since the subcarrier spacing in $\beta = 4$ is twice as much as in $\beta = 3$. Following the same logic, there will be four parallel subframes over the frequency domain when $\beta = 2$ is adopted. As such, the total number of slots in all three numerologies is the same as 16 slots, however, grouped into 4, 8, or 16 slots per subframe for $\beta = 2, 3$, or 4 respectively. Accordingly, we configure the tunable parameters as $(u_{1,min}, u_{1,max}) = (2, 3), (4, 6)$, and $(6, 12)$ respectively.

In Figs. B.15-B.17, the performance of the DSA-GF scheme is illustrated as a histogram for the four sets of device population and activation probability configurations

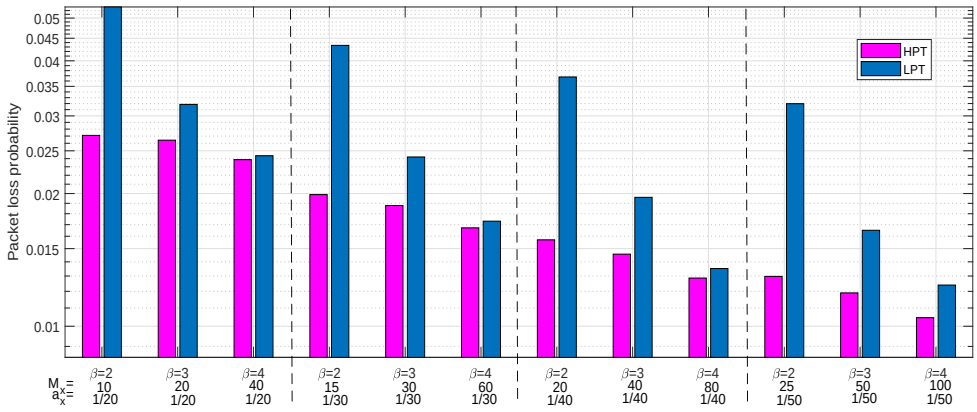


Figure B.17: Applying DSA-GF to three numerologies, $\beta = 2$, 3, and 4 respectively: Packet loss probability.

respectively. As shown in Fig. B.15, the achieved throughput per slot remains almost constant in all three numerologies. On the other hand, the throughput per subframe is doubled when a higher-level numerology is adopted. As expected, the achieved throughput per subframe reaches 0.4865, 0.9736, and 1.9522 for $\beta = 2$, 3, and 4 respectively. This is due to the fact that more slots per subframe are available with a higher-level numerology. When observing the achieved packet loss probability in Fig. B.17, it is evident that a higher-level numerology leads to lower packet loss. This is because more slots are aggregated in one subframe as resources for devices to share when a higher-level numerology is adopted. In addition, for a fixed numerology, the probability of packet loss decreases as M_2 increases, since a_2 reduces when $M_2 a_2$ is constant.

Finally, let us compare the performance of HPT and LPT. Although the offered HPT and LPT traffic is the same, the achieved throughput per subframe for HPT is slightly higher that of LPT. As a consequence, better performance has been achieved for HPT than for LPT in terms of access delay and packet loss probability. This benefit is brought by the priority enabled DSA-GF adaptive algorithm which performs well in all studied numerologies and network configurations. As shown in Fig. B.16, the access delay for HPT is slightly decreasing with a higher-level numerology whereas (much) higher delays are experienced by LPT. The same trend applies to the packet loss probability performance as well, as illustrated in Fig. B.17. Obviously, better performance for HPT can be achieved by increasing the values of $u_{1,min}$ and $u_{1,max}$, at the expenses of slight penalties for the performance of LPT.

B.5.7 Further Discussions

The DSA-GF scheme considers the distinctive characteristics of two co-existing traffic types in a 5G NR network. Although it is unavoidable to sacrifice the performance of LPT in order to ensure the high performance of HPT, serious access congestion for LPT can be avoided or minimized through proper parameter configurations. In general, there is a tradeoff between the performance of these two traffic classes when deciding the values for $u_{1,min}$ and $u_{1,max}$.

Furthermore, $u_{1,min}$ and $u_{1,max}$ are two configurable parameters. Their values are considered to be pre-configured based on gNB's observations as well as service requirements and do not change over a short term (i.e., neither on a subframe-by-subframe nor on a frame-by-frame basis).

B.6 Conclusions and Future Work

This paper presents a priority enabled GF access and data transmission scheme which enables dynamic slot allocation for heterogeneous GF traffic in 5G NR networks. Based on the NR frame structure, the proposed scheme grants access privilege for slot occupancy to high priority traffic based on traffic estimation and the observed transmission status and allocates the remaining slots in each subframe to low priority traffic. While the performance of high priority traffic is guaranteed through proper configuration of relevant parameters, low priority traffic also enjoys satisfactory performance. Furthermore, the precedence of high priority traffic and the dependence between two heterogeneous traffic classes are captured through a Markov model which derives a pseudo-aggregated process to bridge the aforementioned dependency. Through both analysis and simulations, we demonstrate the elegance and effectiveness of the scheme with respect to four performance parameters, i.e., throughput per subframe and per slot, access delay, and packet loss probability, as well as its applicability. To achieve optimal performance, proper parameter tuning is needed based on network setup and traffic conditions. How to adjust $u_{1,min}$ and $u_{1,max}$ configurations periodically, e.g., in the order of seconds, over a long term, or reactively depending on real-time traffic measurements, and how to deal with estimation error are left as our future work.

References

[1] I. Leyva-Mayorga, C. Stefanovic, P. Popovski, V. Pla, and J. Martinez-Bauset, "Random access for machine-type communications," *Wiley 5G Ref: The Essential 5G Reference Online*, 2019.

[2] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4-16, Feb. 2014.

[3] 3GPP TS 38.213, "NR; Physical layer procedures for control," v16.3.0, Sep. 2020.

[4] 3GPP TS 38.214, "NR; Physical layer procedures for data," v16.3.0, Sep. 2020.

[5] 3GPP TS 38.912, "Study on new radio (NR) access technology," v16.0.0, Jul. 2020.

[6] N. Abramson, "The AlohaNet – Surfing for wireless data," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 21–25, Dec. 2009.

[7] 3GPP TR 38.824, "Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)," R16, v16.0.0, Mar. 2019.

[8] A. T. Abebe and C. G. Kang, "Comprehensive grant-free random access for massive & low latency communication," in *Proc. IEEE ICC*, May 2017, pp. 1–6.

[9] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free random access solutions for URLLC services in 5G new radio," in *Proc. IEEE ISWCS*, Aug. 2019, pp. 607-612.

[10] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182-185, Apr. 2018.

[11] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[12] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Simple semi-grant-free transmission strategies assisted by non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4464–4478, Jun. 2019.

[13] J. Ding, D. Qu, and J. Choi, "Analysis of non-orthogonal sequences for grant-free RA with massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 150–160, Jan. 2020.

[14] E. Casini, R. De Gaudenzi, and O. R. Herrero, "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access schemefor satellite access packet networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1408–1419, Apr. 2007.

[15] V. Casares-Giner, J. Martinez-Bauset, and C. Portillo, "Performance evaluation of framed slotted ALOHA with reservation packets and succesive interference cancelation for M2M networks," *Comput. Netw.*, vol. 155, pp. 15-30, Mar. 2019.

[16] F. Lázaro, Č. Stefanović, and P. Popovski, "Reliability-latency performance of frameless ALOHA with and without feedback," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6302–6316, Oct. 2020.

[17] J.-F. Frignon and V. C. M Leung, "A pseudo-Bayesian ALOHA algorithm with mixed priorities," *Wireless Netw.*, vol. 7, no. 1, pp. 55-63, Jan. 2001.

[18] R. L. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. 33, no. 3, pp. 323-328, May 1987.

[19] M. H. Habaebi, B. M. Ali, and M. R. Mukerjee, "Wireless adaptive framed pseudo-Bayesian ALOHA (AFPBA)," *Int. J. Wireless Inf. Netw.*, vol. 8, no. 1, pp. 49-59, Jan. 2001.

[20] 3GPP TS38.211, "NR; Physical channels and modulation," R16, v16.1.0, Mar. 2020.

[21] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. of the Communications Society*, vol. 55, pp. 179–189, Jan. 2020.

[22] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.

[23] 3GPP TS 22.261, "Service requirements for the 5G system," R18, v18.0.0, Sep. 2020.

[24] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1-7.

[25] T. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Priority-based initial access for URLLC traffic in massive IoT networks: Schemes and performance analysis," *Comput. Netw.*, vol. 178, 107360, Sep. 2020.

[26] B. T. Doshi and H. Heffes, "Overload performance of several processor queueing disciplines for the M/M/1 queue," *IEEE Trans. Commun.*, vol. 34, no. 6, pp. 538-546, Jun. 1986.

[27] G. Rubino and B. Sericola, "Sojourn times in finite Markov processes," *J. Appl. Prob.*, vol. 27, no. 4, pp. 744-756, Dec. 1989.

[28] V. Casares-Giner, V. Sempere-Payá, and D. Todolí-Ferrandis, "Framed ALOHA protocol with FIFO-blocking and LIFO-push out discipline," *Netw. Protocols and Algorithms*, vol. 6, no. 3, pp. 82-102, Aug. 2014.

Appendix: A List of Notations

Table B.4: Summary of notations and descriptions

Notation	Description
x	Traffic type index where $x = 1$ and $x = 2$ represent HPT and LPT respectively
β	Numerology index in the NR frame structure where $0 \leq \beta \leq 4$
$\mathbb{N}$ ($\mathbb{Z}$)	The set of natural (integer) numbers
M_x	Number of HPT and LPT devices
a_x	Probability of generating one data packet per subframe for traffic type x, i.e., activation probability
$M_x a_x$	Offered traffic of type x
U	The number of slots per subframe (with its value decided by the adopted NR numerology)
$u_{x,min}$	The min. number of slots allocated to traffic type x
$u_{x,max}$	The max. number of slots allocated to traffic type x
$W_{x,t}$	r.v. of the number of active devices estimated by the gNB for traffic type x at subframe t
$U_{x,t}$	r.v. of the number of slots allocated by the gNB to traffic type x at subframe t
$N_{x,t}$	r.v. of the number of active devices of type x according to the packet arrival and departure processes
$w_{x,t}$	Number of packets of type x estimated by the gNB ready for transmission at subframe t
$E_{x,max}$	The maximum value of $W_{x,t}$ estimated by the gNB
$m_{x,t}$	Number of slots allocated to traffic type x at subframe t ($m_{1,t} + m_{2,t} = U$)
$\overline{m}_{x,t}$	Average number of slots allocated to traffic type x at subframe t
$n_{x,t}$	Number of active devices of type x at subframe t according to the packet arrival and departure processes
(μ, u, i)	Simplified notation for $(w_{x,t}, m_{x,t}, n_{x,t})$
(ν, v, j)	Simplified notation for $(w_{x,t+1}, m_{x,t+1}, n_{x,t+1})$
$P_{\mu,u,i;\nu,v,j}$	The set of transition probabilities from subframe t to subframe $t+1$
$\pi_{\mu,u,i}$	Steady-state probability that, at the beginning of a subframe, the number of active devices estimated by the gNB is μ, the number of allocated slots is u, and the number active terminals is i
$\hat{P}_{u,v}$	The set of transition probabilities from subframe t to subframe $t+1$ for the pseudo-aggregated process
$\hat{\pi}_u$	Steady-state probability of number of slots occupied by HPT
$p_{x,t}$	Permission probability for packet transmissions in subframe t
$h_{x,t}$	Number of unused slots (holes) in subframe t
$s_{x,t}$	Number of successful slots in subframe t
$c_{x,t}$	Number of collided slots in subframe t
$(h, s, c)_{x,t}$	The set of observations at subframe t
$\hat{\lambda}_{x,t}$	The estimated number of *new* devices that have become active during subframe t
$\hat{w}_{x,t+1}$	Number of estimated backlogged devices which will be active in subframe $t+1$
$D^{i,u}_{s_t,c_t}(p)$	Probability that within subframe t, z out of i active devices transmitted with permission probability p, and the result is s_t successes and c_t collisions
$B^i_z(p_t)$	Probability that z out of i active devices ($0 \leq z \leq i$) transmit in subframe t following a binomial distribution
Ω_x	The set of (h, s, c) values observed in subframe t for traffic type x
γ^{sf}_x	Total throughput per subframe for traffic type x
θ_x	Packet loss probability for traffic type x
γ^{slot}_x	Total throughput per slot for traffic type x
d^{sf}_x	Delay for traffic type x, in number of subframes
Δ	The set of all possible values in μ and i, (μ, i)
$\mathcal{C}$	The set of all possible collisions such that $h_t + s_t + c_t = u$
$E = \{(\mu, i)\}$	The set of states of the Markov chain for HPT

Paper C

Title:	Supervised Learning based Arrival Prediction and Dynamic Preamble Allocation for Bursty Traffic
Authors:	Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, and Frank Y. Li
Affiliation:	Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway
Conference:	*IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2019
URL:	https://ieeexplore.ieee.org/document/9093789
Copyright ©:	IEEE

Supervised Learning based Arrival Prediction and Dynamic Preamble Allocation for Bursty Traffic

Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, and Frank Y. Li

Abstract - Achieving ultra-reliable low latency communications (URLLC) in massive machine type communication networks requires novel medium access mechanisms to accommodate a huge number of traffic arrivals. Random access based on LTE-A suffers from collisions and long latency when two or more devices select the same preamble to initiate channel access simultaneously and this problem becomes severe in mMTC networks. In this paper, we propose a machine learning based scheme that allows an eNB to predict the number of arrivals at each random access slot and allocate preambles accordingly. We demonstrate that, by combining arrival prediction with group based dynamic preamble reservation, the grouped devices are able to achieve URLLC under bursty traffic conditions and meanwhile the performance of non-grouped devices is also improved.

C.1 Introduction

Driven by various novel application scenarios, the development of 5G mobile and wireless communication standards is focusing on three main technological directions, i.e., enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable low latency communications (URLLC). While mMTC is expected to accommodate 1+ million connections per square kilometer, URLLC is targeted at providing ultra-reliability levels with very low latency for certain types of, e.g., mission-critical, services. When the number of devices attempting to access the network is large, it often results in access congestion due to competitions among devices for scarce resources and thereby deteriorating their performance. This situation becomes even more challenging for bursty traffic. Therefore, such problems need to be addressed to a satisfactory degree in order to achieve URLLC for mMTC. However, the long-term evolution advanced (LTE-A) based random access (RA) process is not designed to facilitate a very large number of devices due to its limited number of available preambles. This limitation may increase collision probability and lead to long latency.

In [2], the 3rd generation partnership project (3GPP) specifies several possible solutions to address LTE RA congestion. One popular solution is access class barring (ACB) based schemes according to which devices are classified into access categories with different access probabilities and barring times. Moreover, approaches like dynamic resource allocation, MTC specific backoff, slotted RA, and pull based (i.e., eNB initiated) access procedures were also considered in [2]. In addition to 3GPP based solutions, there exist also other approaches proposed to reduce RA congestion (see e.g., [2] and the references therein). Moreover, group based access schemes have also been proposed to reduce collision probabilities [3]. However, while these solutions contribute towards reducing random

access channel (RACH) congestion, they do not adequately address the URLLC requirements. Consequently, these solutions often provide high levels of reliability at the cost of long latency. For URLLC, the tradeoff between reliability and latency needs to be addressed.

To estimate traffic arrivals at an eNB is one potential technique that can be utilized to prevent the occurrence of congestion at an early stage which leads to reduced latency. For ACB based random access, there exist some studies which focus on estimating random access load and then adjusting ACB parameters using different methods. For example, [4] proposed a congestion-aware admission control mechanism in which MTC signaling traffic is rejected at the radio access network with a probability p that represents the level of congestion. It utilizes a proportional integrative derivative controller to derive the value of p. A Markov chain based traffic load estimation scheme was proposed in [5]. As a machine learning based effort, [6] proposed a reinforcement learning based approach to dynamically adjust the ACB barring rate. While most of these solutions focused on parameter tuning for the ACB scheme, it is imperative to investigate how the observed data at an eNB can be combined with machine learning techniques to achieve real-time predictions of arrivals data which enables URLLC.

In this paper, we propose a supervised learning based random access scheme which first predicts bursty traffic arrivals at an eNB and then allocates preambles for group based access. The prediction is based on the detected number of preambles at the eNB. According to its prediction, the eNB is able to dynamically evaluate traffic arrival conditions and allocate preamble resources to different types of user groups. By combining a group based access phase along with bursty traffic prediction, the eNB is able to provide URLLC services to a set of mMTC devices.

The remainder of this paper is organized as follows. Sec. C.2 provides the background information and problem statement for this study. Thereafter, Sec. C.3 presents a machine learning based approach for predicting the number of arrivals, followed by Sec. C.4 which proposes one scheme for achieving URLLC based on arrival predictions. The numerical results are provided in Sec. C.5. Finally, the paper is concluded in Sec. D.6.

C.2 Background and Problem Statement

In this section, we first present some related background for this study and then introduce the problem statement.

C.2.1 LTE-A RACH Process

Consider that multiple MTC end user devices are covered by the same eNB. When an uplink communication is required, an LTE-A device needs to follow a 4-step random access procedure [7], as illustrated in Fig. C.1. Each device needs to randomly select a preamble from a set of R available preambles which are periodically advertised by the eNB. These R preambles are orthogonal to each other and the selected preambles are transmitted in the next available RA slot which appears every fifth subframe [7]. If two or more devices select the same preamble to transmit in the same RA slot, a collision of preambles may be

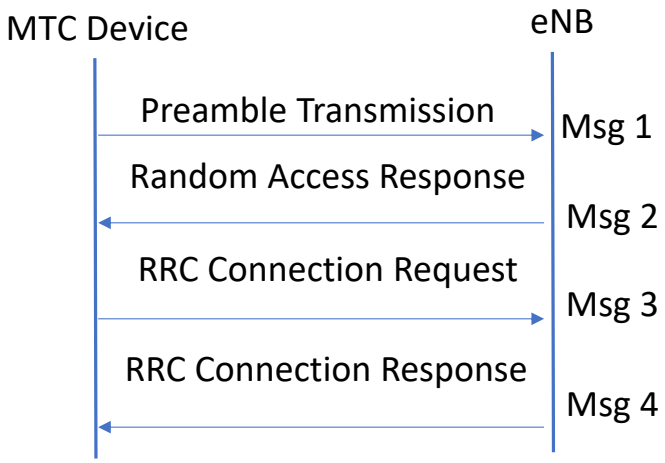


Figure C.1: 4-step LTE-A random access for MTC devices.

detected and the collided devices will not receive a Msg 2, i.e., a random access response (RAR) message, from the eNB. If no Msg 2 is received within a timeout period, the devices will retransmit Msg 1 up to a maximum number of times, N_{max}. A retransmission follows the same procedure as mentioned above but happens after a backoff time selected randomly from the range of $[0, W_{BO} - 1]$, where W_{BO} is the backoff window size.

If a preamble is successfully received at the eNB, it will reply with Msg 2 in Fig. C.1. If two or more devices select the same preamble and transmit within the same RA slot but the eNB cannot detect a collision at Step 1, then Msg 3 and Msg 4 will be exchanged to resolve the contention.

C.2.2 RACH Limitations

A main constraint of the LTE-A RA process is the limited number of preambles available for a cell. According to [7], 64 preambles can be allocated for a particular cell and out of these a certain amount, typically 10, is reserved for non-contention based transmissions like handover traffic. The rest are considered to be available for the competing devices for random channel access.

When a large number of devices attempt to access the channel at the same time, the preamble collision probability increases. Additionally, this will result in longer latency for successfully accessed devices. This problem is even more serious for mMTC scenarios where the number of competing devices could be much larger. To demonstrate this effect, we illustrate in Fig. C.2 the access success probability for a bursty mMTC traffic scenario with an increasing number of devices based on the analysis in [8] and the simulations we performed. It can be observed that when the number of devices is very large, i.e., 30000 or more, the access success probability decreases sharply to an unsatisfactory level.

Generally, traffic arrivals for periodic data reporting are considered to follow an uniform arrival process. On the other hand, event-driven data reporting which often leads to bursty traffic arrivals is assumed to follow a Beta distribution based arrival function as expressed below

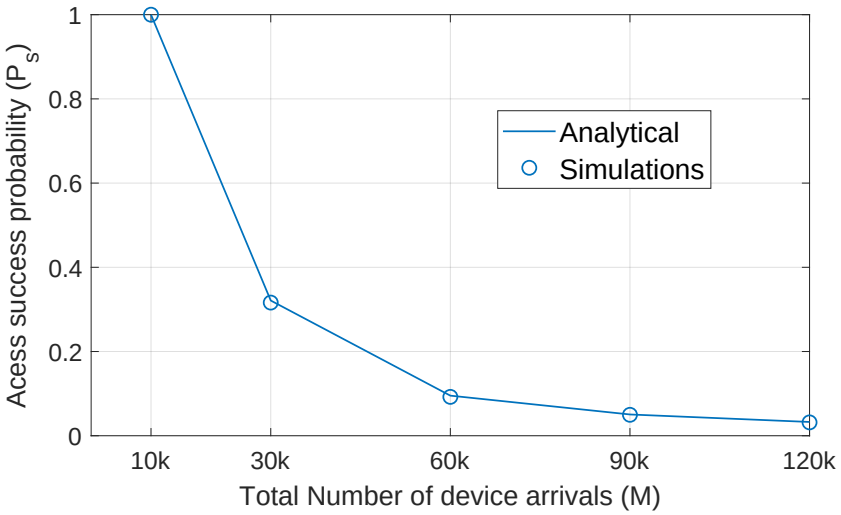


Figure C.2: Access success probability for a varying number of total devices.

$$A(i) = M \int_{t_i}^{t_{i+1}} p(t)dt, \tag{C.33}$$

where $A(i)$ represents the access intensity for a total number of M devices accessing an RA slot i between time t_i and t_{i+1}. In (C.33), $p(t) = (t^{\alpha-1}(T-t)^{\beta-1})/(T^{\alpha+\beta-1}Beta(\alpha,\beta))$ with $Beta(\alpha,\beta)$ being the Beta function with $\alpha = 3$ and $\beta = 4$. T is the total observation time for traffic arrivals [2].

C.2.3 Problem Statement

Considering the drawback in the LTE-A RA process as discussed above, it is imperative to introduce novel solutions to enable URLLC based applications. Since an eNB does not have sufficient information on the number of devices attempting random access at a particular time, it is difficult to implement real-time dynamic preamble allocation based solutions which satisfy the needs of URLLC.

Assuming that a preamble is correctly received by the eNB, collision detection at the eNB depends on several factors like the delay spread and the signal strength received from the competing devices. The eNB may detect a preamble even though multiple devices are transmitting the same preamble, if the devices are separated sufficiently far away from each other. On the other hand, when multiple devices are closer to each other, the preamble transmissions overlap with each other and the eNB cannot distinguish whether there are two or more users transmitting using the same preamble. Hence, Msgs 3 and 4 are needed to resolve the collision. Generally, it is difficult to use the previous collision data to predict the future MTC arrivals accurately.

In this work, we resolve this issue by proposing an arrival prediction based preamble allocation (APPA) scheme which consists of two phases. 1) We first propose a machine learning based technique to predict the number of arrivals in a given RA slot based on the number of successful detections at the eNB; and 2) we then propose dynamic preamble allocation to achieve URLLC based on the estimated arrivals. In the following two sections, we present these two phases in details.

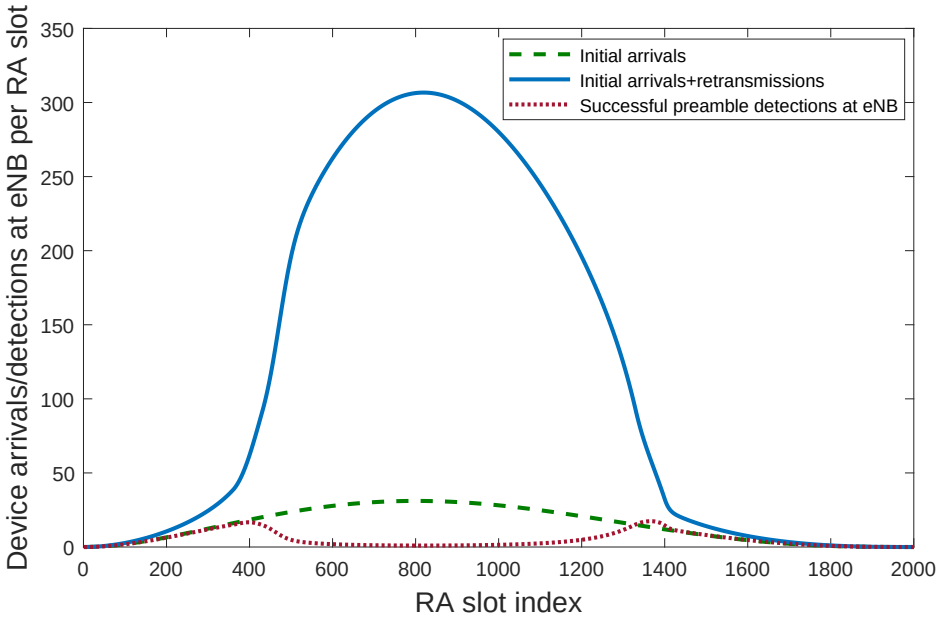


Figure C.3: Number of arrivals and detections in LTE-A random access for 30000 devices with 54 preambles following a bursty arrival process.

C.3 APPA Phase 1: Arrivals Prediction

In this section, the proposed machine learning based prediction technique is presented. It corresponds to the first phase of APPA as shown in Algorithm 1.

C.3.1 Arrivals versus Successful Access: A Dilemma

For a given RA slot, the total number of arrivals consists of new arrivals and the retransmissions from previously failed devices. Due to collisions and detection failures, only a few number of such arrivals are correctly decoded at the eNB. In some cases, not all the detected devices will receive Msg 2 due to the limit on the number of devices that can be responded in a given RAR message.

A substantial distinction between the numbers of arrivals and detections within an RA slot under a bursty traffic scenario can be observed in Fig. C.3, which is obtained based on 30000 devices in a period of 10 sec with $R = 54$ preambles. The numerical results presented in this figure are generated following the LTE-A RA process [2] and the analytical model proposed in [8]. It is evident that the actual number of arrivals consisting of the initial arrivals and the retransmissions is much higher than the initial arrivals itself. Hence, the number of retransmissions caused by collisions is a major cause for further collisions. Furthermore, it is observed that the number of successfully detected preambles initially increases with the increasing number of arrivals and then decreases to a lower value as the arrival traffic reaches its peak around the 850^{th} RA slot. Thereafter, the success rate increases when the number of arrivals decreases, and then it reaches null corresponding to zero arrivals.

In reality, not all the information illustrated in Fig. C.3 is available at the eNB. Instead, the eNB has only knowledge on how many preambles are successfully detected at each RA slot. From the above observations, we argue that there is a clear relationship

Algorithm 1 Algorithm for arrival prediction and group based dynamic preamble allocation

Input for Phase 1: Training data including Detections, Arrivals, and Current arrival type; Validation data including Number of detected preambles (real time); Bursty arrival threshold;

Input for Phase 2: Initial arrival type; Number of levels (L); Priority level for each group (l); Number of groups per level (N_l); Device population M; Number of available preambles at eNB R; Number of priority levels (L); Assigned priority level l for each group; Number of groups with priority level l (N_l), priority level l group threshold η.

Output: Predicated number of arrivals at each RA slot; Dynamic preamble allocation and group enabling

Phase 1: Prediction of Arrivals:

1: **Training:** Smooth input data via Savitzky-Golay filtering;
2: Train the model using smoothed training data;
3: Select the model with minimum RMSE.
4: **Prediction:** Input real detection data at each RA slot and current arrival type to the trained model.
5: **if** predicted arrivals > Bursty arrival threshold, **then**
6: current arrival type = Bursty
7: **else**
8: current arrival type = Normal
9: **end if**

Phase 2: Dynamic Group Preamble Allocation:

10: **if** predicted arrivals > priority level l group threshold **then**
11: Upgrade N_l groups in level l to highest priority
12: Reserve preambles for N_l groups in priority level l
13: Inform NGDs about preamble reservation
14: Update device population and number of available preambles for next prediction and preamble allocation.
15: **else**
16: Go to Line 4 **Prediction**.
17: **end if**

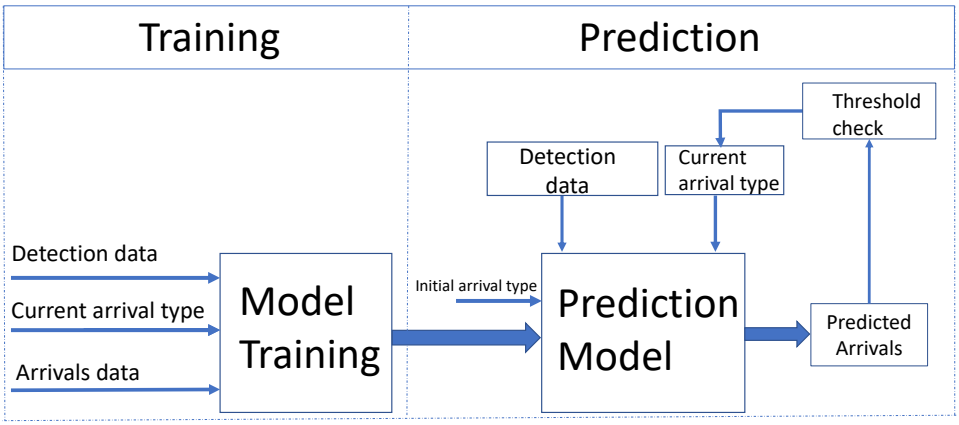


Figure C.4: A machine learning based prediction model.

between the number of detections and the corresponding number of arrivals for a given device population attempting to access the same channel. If the eNB is able to predict the number of arrivals based on the local available information, i.e., the detected preamble, it would be helpful to allocate on the fly a number of preambles to reduce RA congestion. In the following, we propose a machine learning based arrival prediction technique that utilizes the detected data to predict the number of arrivals at each RA slot.

C.3.2 Arrival Prediction using Supervised Learning

Supervised learning is one type of machine learning algorithms that maps an input to an output based on labeled training input-output data pairs. It is a widely used technique that can be utilized for various regression and classification tasks. In this work, we aim at applying supervised learning to predict the number of arrivals at the eNB.

Fig. C.4 denotes a block diagram which illustrates the main idea of the proposed model for arrival prediction. The input data available at the eNB is the number of successful detections at each RA slot. The eNB has also information about the initial arrival type, i.e, the traffic type is bursty or normal. These two features and the corresponding arrival data are used as input to train the model to predict the number of arrivals. The trained model is then validated with the test data. For arrival prediction, we are interested in identifying the point at which the predicted level of arrivals crosses a certain pre-defined threshold. This criterion is evaluated with each new data arrival and the result is fed back to the model to update its information about the current status of the traffic arrivals. In what follows, we further elaborate the aforementioned process.

C.3.2.1 Input data preparation

For model training, a simulation based data set following the LTE-A RA process is generated. For initial training, a burst of arrivals with $M = 30000$ devices for a duration of 10 sec is considered. These devices compete for 54 RA available preambles following the RA process described in Subsec. C.2.1. The losses due to channel impairment are represented using a detection probability p_n, where the preamble detection probability of the n^{th} preamble transmission is given by $p_n = 1 - 1/e^n$. In order to increase the accuracy of learning, the input data is filtered to smooth out any abrupt changes. This

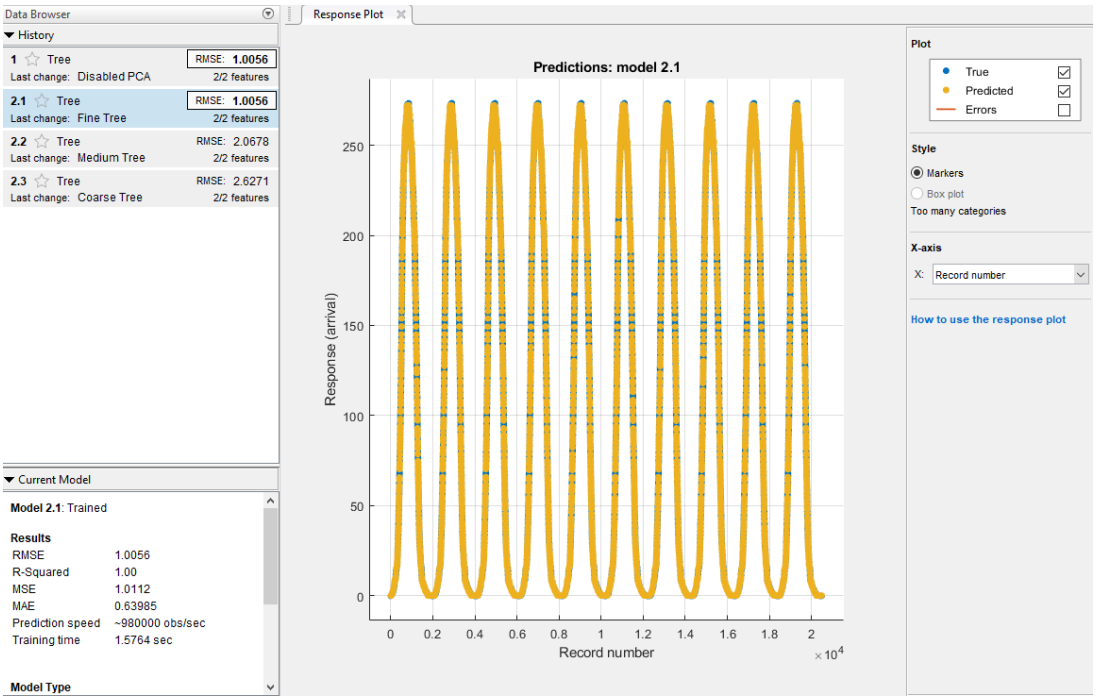


Figure C.5: Responses from the learning model with RMSE.

is achieved using the Savitzky-Golay (SG) filtering method which enables to increase the signal-to-noise ratio without significantly distorting the signal. The SG method achieves this by minimizing the least-square errors in fitting a polynomial to frames of noisy data.

C.3.2.2 Model training

The model training process is carried out through a 5-fold cross validation procedure that ensures protection against over fitting by partitioning the data set into different folds and estimating the accuracy of each fold. Different models are evaluated based on root mean square errors (RMSE) which indicate how close the observed data points are with respect to the values predicted by the model. Among the models available in the simulation tool, the tree based models generate the minimum RMSE error. Therefore, the fine tree model with RMSE ≈ 1.0 is selected for our validation. In Fig. C.5, we illustrate the response plots based on the fine tree model. It reveals that the predicted values (in yellow) match the real values (in blue which are largely overlapping with the yellow curve) precisely. In Fig. C.6, we demonstrate the accuracy of the training model by plotting the real response and the prediction response in the x- and y-axis respectively. With both sets fitting the diagonal line, the accuracy of the trained model is verified.

C.3.2.3 Prediction model

To validate the training model, another set of simulation data that represents a traffic arrival burst is generated. The initial traffic arrival type is considered to be normal and we assume that the eNB can detect the transmitted preambles successfully at each RA slot. The eNB performs live data detection and relies on the trained model to predict new arrivals. The data detection result from the current RA slot is added to the existing

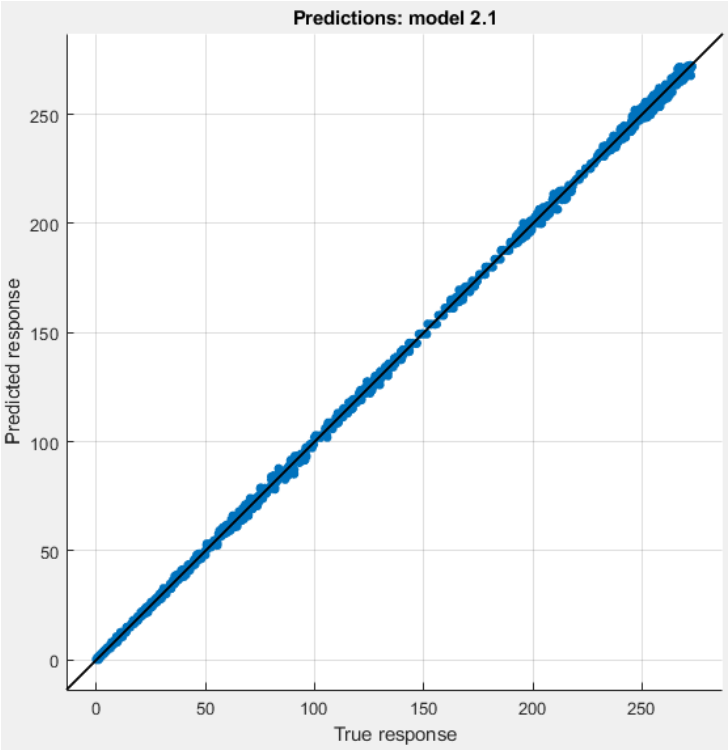


Figure C.6: Responses from the learning model.

data to predict arrivals in the next slot.

When the predicted number of arrivals within an RA slot exceeds the bursty arrival threshold, the traffic type is assessed to having changed from normal to bursty. Furthermore, different thresholds are configured to measure the bursty level of traffic arrivals. We will further elaborate this procedure in Sec. C.4. On the other hand, when the predicted arrival intensity drops below a certain threshold, the traffic type is considered to having changed back to normal traffic.

Moreover, providing a higher number of training data samples results in a lower RMSE in the training model. Hence, several iterations of bursty traffic with the same 10 sec duration are provided for training. Recall that the detection data alone may not provide sufficient information for accurate arrival prediction. Therefore, we assume that the traffic type is known initially. With the knowledge on both traffic type and detection data, precise arrival prediction can be achieved at the eNB.

C.4 APPA Phase 2: Preamble Allocation

The knowledge on the number of arrivals can be exploited in several ways for URLLC applications. In this section, we present two group based preamble allocation schemes, one static which serves as a baseline scheme and another one dynamic which is based on arrival prediction presented above.

C.4.1 Static Group based Preamble Allocation (SGPA)

In this scheme, MTC devices are grouped based on their URLLC priority levels, location, and applications. Devices with URLLC access requirements, e.g., those for monitoring mission-critical information in smart grids or industrial processes, belong to grouped devices (GDs). Other MTC devices which are covered by the same eNB are regarded as non-grouped devices (NGDs).

Each group has a dedicated preamble managed by the group leader which has higher processing and memory capability in comparison with its members. The eNB stores information about group members and their leaders during the initial registration process. When a triggering event occurs, the group devices establish an uplink communication with their associated eNB through a collision-free preamble transmission initiated by their group leader. By decoding the dedicated preamble, the eNB identifies the group and its members based on the stored information during the initial registration process. Then, the eNB will allocate an appropriate amount of radio resources for all the devices in that particular group. For NGDs, the access process follows the standard RA procedure as explained in Subsection II-A. However, for a given device population, allocating a suitable number of preamble to NGDs is not an easy task, especially for bursty traffic. With static preamble allocation without real-time traffic intensity awareness, the performance of GDs and/or NGDs may be deteriorated.

C.4.2 Arrival Prediction based Preamble Allocation

In the APPA scheme, we consider that GDs have L different priority levels based on their URLLC requirements, denoted as $1, 2, ..., L$ in an descending priority order. The devices belonging to a higher priority level have more stringent latency requirements than those in a lower level group. Regardless of the arrival traffic type, the highest priority, i.e., level-1, GDs always have their dedicated preambles reserved for communication with the eNB. Under *normal* traffic conditions, other priority level group members follow the legacy LTE-A RA process the same as NGDs. However, when a traffic burst is observed in the first phase of APPA, the eNB will dynamically allocate more preambles as dedicated to a lower level GDs. In this way, more devices will experience collision-free transmissions based on dynamic preamble allocation.

Whenever a group belonging to a certain priority level is upgraded to the level with dedicated preambles, its member devices will be able to enjoy collision-free preamble transmissions through their group leaders. We assume that the eNB transmits the updated reservation information immediately to all devices. With this information, these GDs will not compete with the NGDs using the common RA preambles. Accordingly, the number of preambles available for the NGDs is reduced. Meanwhile, the number of competing devices is also reduced. Recall, however, that such a preamble allocation update is performed dynamically based on the predicted traffic arrival. Hence, in comparison with the SGPA scheme, the preamble utilization efficiency is improved in APPA.

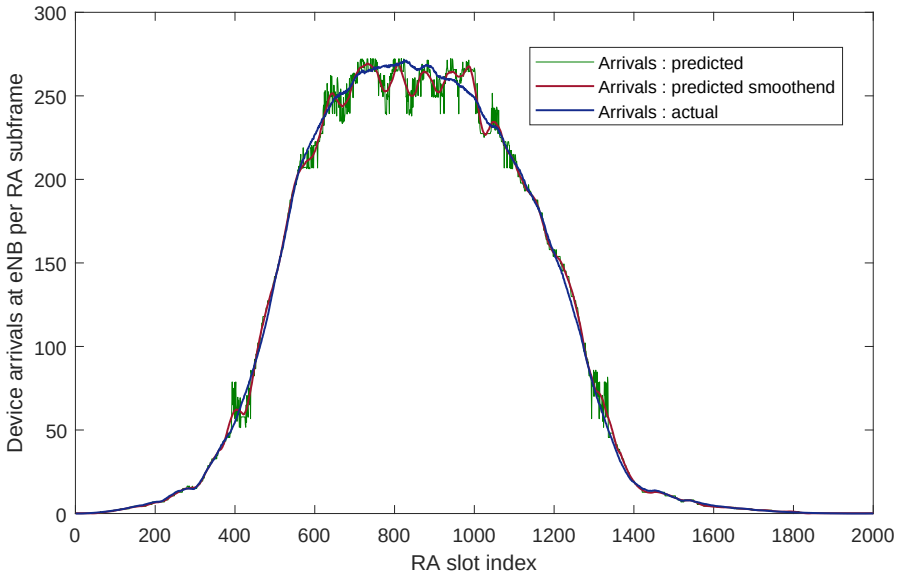


Figure C.7: Bursty traffic arrivals: Prediction versus actual.

C.5 Simulations and Numerical Results

To evaluate the performance of the proposed schemes, we perform extensive simulations in MATLAB. Consider a cell with $M = 30000$ devices and a traffic burst in a period of 10 sec. The devices are categorized into GDs (with 40% of M devices) and NGDs (with 60% of M devices). These GDs are further classified into $L = 3$ levels, as 10%, 10%, and 20% of M devices for level-1, -2, and -3 respectively. For each level, there are multiple groups each with a number of member devices. As mentioned earlier, each group has only one dedicated preamble which is managed by the group leader.

For performance comparison, three schemes are studied, i.e., 1) preamble allocation without grouping (PAWG), 2) SGPA, and 3) APPA. The following three metrics recommended by 3GPP [2] are used for our performance evaluation.

- Collision probability, P_c, defined as the ratio between the number of occurrences when two or more devices transmit the same preamble during the same RA slot and the overall number of RA opportunities within this slot.
- Access success probability, P_s, defined as the probability that a device successfully completes the RA procedure within $N_{max} + 1$ transmissions.
- Average delay for successfully accessed devices, D_a, calculated from the first preamble transmission attempt to the successful completion of the access process.

C.5.1 Validation of the APPA Scheme

To assess the performance of the proposed APPA scheme, we plot two figures to illustrate the results obtained based on the two phases of APPA, i.e., Fig. C.7 for Phase 1 and Fig. C.8 for Phase 2, respectively.

In Fig. C.7, we illustrate both the actual traffic arrivals generated by simulations and the predicted number of arrivals which is obtained based on the supervised learning

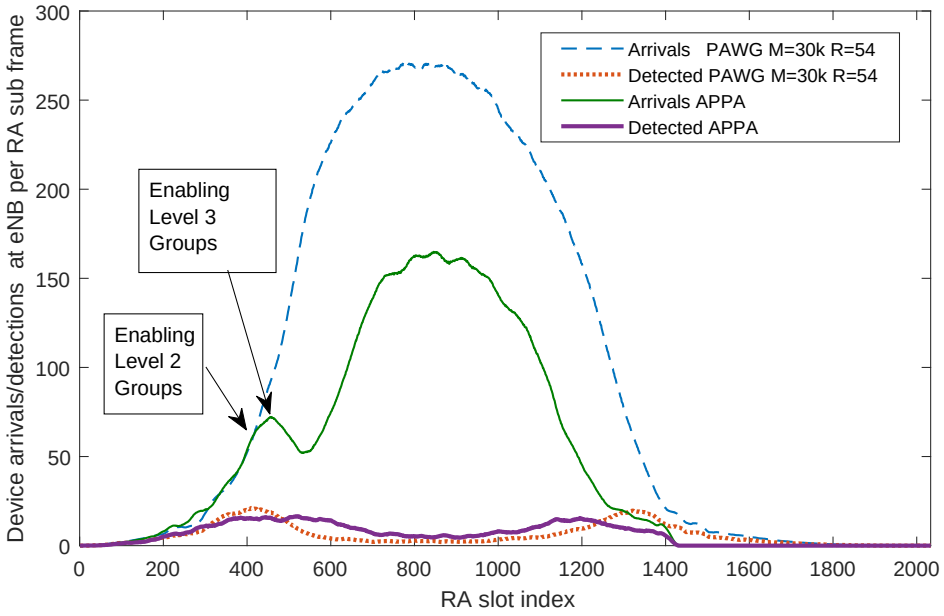


Figure C.8: Enabling dynamic grouping according to traffic arrival prediction.

algorithm presented in Subsection III-B. It is evident that, despite the sparks occurred in simulated arrivals, the predicted number of arrivals represents closely the actual data arrivals.

Let us now explain how the second phase of APPA works using Fig. C.8 which illustrates how to allocate the number of dedicated preambles based on the predicted number of arrivals per RA slot. Following the LTE-A RA process, there are $R = 54$ preambles which are available for all $M = 30000$ devices. However, how to allocate these preambles depends on the adopted scheme. In PAWG, all devices follow the standard process based on 54 preambles. In SGPA, certain number of preambles are allocated to GDs beforehand but the eNB is not aware of traffic arrival patterns.

When APPA is employed, we initially enable only level-1 groups with dedicated preambles. In other words, there are 27k devices, including NGDs and level-2 and -3 GDs, competing for 51 preambles. At around the 407^{th} RA slot, the number of predicted arrivals exceeds the threshold, which is 50 according our network configuration. Then level-2 groups are upgraded with contention-free preambles. Correspondingly, the total number of competing devices is reduced to 24k while the number of preambles reduces to 48. At the 466^{th} RA slot, the predicted arrivals exceed the second threshold, which is 75. Immediately, the level-3 groups are upgraded with contention-free preambles. This causes a greater reduction of the number of competing devices to 18k, competing for 44 preambles.

At each stage when more GDs are allocated with dedicated preambles, the number of NGD arrivals decreases significantly since only leaders generate preambles on behalf of each group. This behavior can be observed in the figure when comparing the original arrival curve based on PAWG in which all 30k devices compete for these 54 available preambles. Accordingly, the RA performance could be improved as presented below.

Table C.5: Performance evaluation of grouped and non-grouped devices

SchemeMetric	P_c	P_s	D_a
GDs	0	1	11.0
NGDs with PAWG	0.4288	0.3153	67.35
NGDs with SGPA	0.2909	0.5875	67.0144
NGDs with APPA	0.2684	0.6816	73.5183

C.5.2 Performance Comparison of PAWG, SGPA, and APPA

Table C.5 illustrates the numerical results for the studied three schemes. For grouped devices with dedicated preambles, denoted as GDs in the table, the collision probability is zero and the access success probability equals to one. The delay associated with these devices is calculated to be approximately 11 subframes. In comparison with the other results shown in the same table, this means that for GDs ultra reliability is achieved together with very low access delay.

For NGDs, as well as level-2 and/or -3 GDs that do not yet have a dedicated preamble, all denoted as NGDs in Table C.5, we investigate the benefits brought by APPA based on the performance metrics defined earlier. It is evident that APPA outperforms SGPA and APPA in terms of P_s and P_c. This is because in APPA preamble resources are dynamically allocated depending on the predicted traffic arrivals. Compared with the significant improvements for P_s and P_c, the extra delay cost introduced by APPA is low.

C.5.3 Further Discussions

The delay results shown in Table C.5 have a unit as the duration of a subframe which is 1 ms in LTE-A. In 5G new radio (NR), the transmission time interval (TTI) is shortened down to 125 μs or even 62.5 μs. Accordingly, much shorter delay can be achieved, meeting the requirements for URLLC.

Another potential application of the proposed APPA scheme Phase 1 is to apply it for dynamic frame structure configuration as suggested in [9]. If a flexible 5G NR frame structure is implemented, where the TTI size is configurable on a per-user basis according to its specific service requirement, different TTI sizes can be configured on the fly depending on the predicted traffic arrival load at eNB. When traffic load is predicted to be light, a short TTI appears to be more pragmatic for achieving low latency, and vice versa.

C.6 Conclusions

In this paper, we have proposed a machine learning based traffic prediction and preamble allocation scheme which encompasses two phases. The first phase relies on local information available at the eNB to estimate the number of arrivals within one RA slot. Based on the prediction, dynamic allocations of preamble resources are performed to enable URLLC applications. By combining group based preamble reservation with the proposed traffic

prediction and preamble allocation scheme, we demonstrate, through extensive simulations, that URLLC for grouped devices can be achieved while improving the performance of non-grouped devices.

References

[1] 3GPP TR37.868, "Study on RAN improvements for machine type communications," v11.0.0, Sep. 2011.

[2] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, Jan. 2017.

[3] K. Lee, J. Shin, Y. Cho, K. Ko, D. Sung, and H. Shin, "A group-based communication scheme based on the location information of MTC devices in cellular networks," in *Proc. IEEE ICC*, Jun. 2012, pp. 4899–4903.

[4] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: overload control," *IEEE Network.*, vol. 26, no. 6, pp. 54–60, Nov. 2012.

[5] H. He, Q. Du, H. Song, W. Li, Y. Wang, and P. Ren, "Traffic-aware ACB scheme for massive access in machine-to-machine networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 617–622.

[6] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Reinforcement learning-based ACB in LTE-A networks for handling massive M2M and H2H communications," in *Proc. IEEE ICC*, May 2018, pp. 1–7.

[7] 3GPP TS36.321, "Evolved universl terrestrial radio access (e-UTRA)," v9.4.0, Sep. 2011.

[8] C. Wei, G. Bianchi, and R. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.

[9] G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen, and P. Mogensen, "On the impact of multi-user traffic dynamics on low latency communications," in *Proc. IEEE ISWCS*, Sep. 2016, pp. 204–208.

Paper D

Title:	Time-Space Domain Availability Analysis Under Reliability Impairments
Authors:	Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, and Frank Y. Li
Affiliation:	Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway
Journal:	*IEEE Networking Letters*, vol. 1, no. 3, pp. 103-106, May 2019
DOI:	10.1109/LNET.2019.2916909

Time-Space Domain Availability Analysis under Reliability Impairments

Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, and Frank Y. Li

Abstract - Availability and reliability are two essential metrics for the design, deployment, and operation of future ultra-reliable low latency communication (URLLC) networks. Despite a vast amount of research efforts towards URLLC, very little attention has been made on the ultra-reliable communication (URC) aspect of URLLC from a dependability perspective. As an effort towards achieving *anytime and anywhere* communication, this paper consolidates a dependability theory based availability concept for individual users by taking into account reliability impairments that affect URC in *both spatial and temporal domains*. To this end, we perform per-user availability analysis by considering channel status and user mobility patterns in a Poisson Voronoi network.

keywords - URC/URLLC, per-user availability, reliability impairments, time and space domains.

D.1 Introduction

Ultra-reliable communication (URC) is an essential requirement for mission-critical and industry automation applications. To support ultra-reliable low latency communications (URLLC) to end users, the fifth generation (5G) wireless networks face various challenges. Many of these challenges for downlink transmissions are related to the reliability requirements for data and control channels [1]. Achieving URC and URLLC, with respect to reliability, requires a paradigm shift in terms of terminologies, methodology, and standards in comparison with earlier generations of wireless networks [2].

From the perspective of network operators, a standard availability metric is highly anticipated to measure the *anytime and anywhere* operation of their 5G networks as a key performance indicator. On the other hand, ensuring the accuracy of reliability and/or availability evaluation requires the capability to adopt to the changes in both time and space domains together with channel conditions. While most prior work focused on reliability or availability analysis in the time domain, we proposed dependability theory based definitions to measure network and user availability by focusing on the space domain [3] or considering both space and time domains [4]. In this letter, we consolidate the proposed URC availability definition in [4] by including factors that may degrade the reliability level of a network, known as reliability impairments (RIs) [5], into our availability analysis. When determining an actual reliability level, RIs like excessive interference, resource constraints, and system failures need to be addressed.

The contributions of this work are as follows. To illustrate the applicability of the advocated definition, we perform an analysis on the URC level experienced by an individual

user in a wireless network. In order to reflect the time and space domain significance on URC availability, a mobility pattern and time varying location of a mobile user (MU) are considered in this work. Moreover, the channel status of the associated cell is modeled as a potential RI and its impact on the overall URC availability level is calculated. Additionally, we investigate different channel selection strategies when an MU is covered by more than one cell.

D.2 Per-user Availability

The per-user availability definition presented below is targeted at an individual end user but it applies to any user [4]. In order to receive services, the user should be located within the cell coverage according to a pre-defined criterion and there should be a sufficient number of vacant channels in the network. Furthermore, the effects of RIs could impair the level of availability experienced by the user even if it is covered.

Let URC-region (UR) denote the region within which URC is achieved despite RIs and $\mathbf{U}(t)$ be the set of coordinates belonging to a UR at time t. Furthermore, denote by $p_k(t)$ the position coordinates of user k at time t. Considering whether the relationship $p_k(t) \in \mathbf{U}(t)$ is true or not, we introduce an indicator random variable $I_{p_k}(t)$ as follows,

$$I_{p_k}(t) = \begin{cases} 1, & \text{if} \quad p_k(t) \in \mathbf{U}(t) \\ 0, & \text{otherwise.} \end{cases} \tag{D.34}$$

From the theory of dependability, availability is defined as the ratio between mean up time and total time which is the sum of mean up time and mean down time. Accordingly, we propose a URC availability definition from a dependability perspective for a single mobile user, k, as follows,

$$A_k = \frac{\int_0^{t_{tot}} I_{p_k}(t)dt}{t_{tot}} \tag{D.35}$$

where A_k is the defined per-user availability and t_{tot} is the total observation time which is assumed to be sufficiently large. t_{tot} can be decided based on a pre-defined criterion like travel time or a fixed observation duration. The integral of the indicator function $I_{p_k}(t)$ over time gives the accumulated time during which $p_k(t) \in \mathbf{U}(t)$. Hence, the ratio between this duration and t_{tot} represents the URC availability that user k experienced during this period of time. When user k is an MU, its location is a function of time. Moreover, $\mathbf{U}(t)$ also varies over time due to the time varying impacts of RIs over UR. Hence, (D.35) provides a general expression to calculate per-user availability, capturing both time and space domain aspects that affect URC.

D.3 System Model

In this section, we introduce the system model for per-user availability analysis based on the advocated definition.

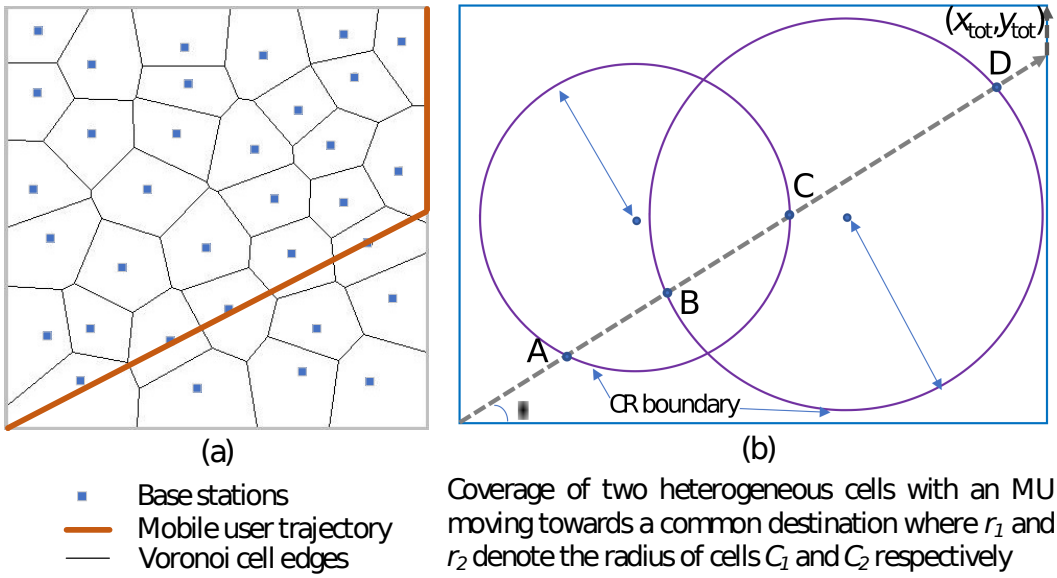


Figure D.1: (a) A PPP distributed homogeneous cellular network consisting of Voronoi cells, (b) Mobile user motion in a two-cell network.

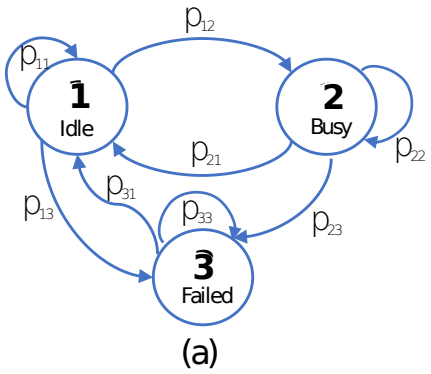
D.3.1 Network Scenario and User Mobility

Stochastic geometry including Voronoi diagrams and Poisson point process (PPP) has been shown to be a powerful mathematical tool for modeling the random distribution of base stations (BSs) [6]. In this study, to model the cellular network with a rectangular area of interest ($x_{tot} \times y_{tot}$), a popular stochastic geometry model is adopted where the infrastructure nodes are Poisson distributed. It consists of M variable size cells forming a Voronoi tessellation. Within each cell, there is a BS located at the center of the cell with an omini-directional antenna. Moreover, a network may be deployed as a homogeneous or heterogeneous network which consists of cells with either identical or different cell sizes.

Two user mobility patterns are considered in this study, denoted as MP1 and MP2 respectively. Mobility pattern MP1 is illustrated in Fig. D.1 where the MU travels from the coordinate of origin $(0, 0)$ to its destination (x_{tot}, y_{tot}). The MU has multiple path options to select, represented by the angle of departure, θ where $0 < \theta < \pi/2$ measured from the x-axis. If the reached boundary is not the destination, the MU will turn towards the destination at the boundary point and then move towards it either counterclockwise or clockwise. MP2 is the random direction model [7] which is a variant of the well-known random way point mobility model.

D.3.2 Cell Coverage and URC Region with RIs

For per-user availability analysis, the URC region needs to be calculated when both cell coverage and reliability impairments are considered. Let coverage-region (CR) denote the region within which the user is covered according to certain criterion. Under an ideal condition, the area of the CR for a single cell is πR^2 where R is the radius of the cell. Note however that *being covered by a BS is a necessary condition for URC provisioning but it is not sufficient.*



$$P = \begin{bmatrix} 1-(p_{12}+p_{13}) & p_{12} & p_{13} \\ p_{21} & 1-(p_{21}+p_{23}) & p_{23} \\ p_{31} & 0 & 1-p_{31} \end{bmatrix}$$

(b)

Figure D.2: (a) A DTMC with three channel states: Idle, busy, and failed; (b) Transition probability matrix of the DTMC.

D.3.3 Channel States and Channel Availability

As mentioned in Sec. D.1, resource constraints and system failures are two types of RIs that could impair URC performance. In wireless networks, channel occupancy and link status may represent such RIs. In what follows, we model error-prone channels together with channel occupancy status as examples of RI using a discrete time Markov chain (DTMC) based approach. For in-depth analysis of channel holding times and modeling of unreliable links, refer to [8] [9].

Consider that a single channel is used in each cell. To model channel status, we adopt a 3-state DTMC. Three states, $\mathbf{1}, \mathbf{2}$, and $\mathbf{3}$, as shown in Fig. D.2(a), represent the idle, busy, and failed states of the channel respectively. The channel is considered to be idle if it is neither occupied by another user nor in a failed state. Let p_{ij} denote the transition probability from state i to state j and the steady state probability of state $\boldsymbol{x}$ is denoted as $\pi_{\boldsymbol{x}}$. The corresponding transition probability matrix of this DTMC is presented in Fig. D.2(b).

The steady state probabilities of this DTMC can be obtained by solving the set of linear equations $\boldsymbol{\pi P} = \boldsymbol{\pi}$ and $\sum_{\boldsymbol{x}\in \boldsymbol{S}} \pi_{\boldsymbol{x}} = 1$ where $\boldsymbol{S}$ is the state space. $\boldsymbol{\pi}$ is the stationary distribution and $\boldsymbol{P}$ is the transition probability matrix. A channel is regarded as available for a new MU when it is in the idle state. Otherwise, the channel is reliability impaired, i.e., when it is either in the occupied or in the failed state. Therefore, channel availability, A, for a new user can be expressed as, $A = \pi_{\mathbf{1}}$.

D.4 Per-User Availability Analysis and Cell Selection Strategies

Considering that an MU is moving across multiple cells, we analyze the achieved per-user availability when channel impairments are regarded as an RI. In addition, we propose three channel selection strategies when a user is located inside a cell intersection region. By following the approach mentioned in Subsec. D.3.3, the probabilities for the three states, i.e., idle, occupied, and failed, can be obtained as $\pi_{\mathbf{1}} = p_{31}(p_{21}+p_{23})/((p_{31}+p_{13})(p_{21}+p_{23})+$

$p_{12}(p_{31}+p_{23}))$, $\pi_{\mathbf{2}} = \pi_{\mathbf{1}} p_{12}/(p_{21}+p_{23})$, and $\pi_{\mathbf{3}} = (\pi_{\mathbf{1}}(p_{13}(p_{21}+p_{23}) + p_{23}p_{12}))/p_{31}(p_{21}+p_{23})$, respectively.

D.4.1 Per-user Availability with RIs

Consider an arbitrarily selected MU. Its per-user availability can be obtained from (D.35). Denote by t_{in} the duration that the MU is located inside the CR. Considering the RI experienced such as channel unavailability discussed above, the per-user availability is obtained by $A_k = (t_{in}/t_{tot})\pi_{\mathbf{1}}$ for a single cell.

For a multi-cell scenario, the transition probabilities vary from one cell to another. To obtain per-user availability in this case, we need to calculate the accumulated time that an MU spends in the availability state across all cells, divided by the total duration of the journey from source to destination, or the observation time. More specifically, it is given by

$$A_k = \frac{\sum_{j=1}^{M} t_{in}(j)\pi_{\mathbf{1}}(j)}{t_{tot}}, \tag{D.36}$$

where $\pi_{\mathbf{1}}(j)$ denotes the steady state probability of being in the idle state of cell j, $t_{in}(j)$ denotes the duration the MU spends inside the CR of cell j during its movement. In an overlapped region across two or more cells, the MU needs to be associated with a specific cell determined by one of the strategies presented in the next subsection.

To calculate $t_{in}(j),\ j = 1, 2, \cdots, M$, the travel distance inside each cell needs to be calculated. Consider the two-cell scenario illustrated in Fig. D.1(b). The intersection points between the coverage circles of these two cells and MU's path can be calculated by substituting the center points and radius values of each cell in the places of (x_0, y_0) and r respectively and jointly solving $x = (x_0 + y_0 \tan\theta \pm \sqrt{(r^2 \sec^2\theta) - (x_0 \tan\theta - y_0)^2})/\sec^2\theta$ and $y = x\tan\theta$. Once the intersection points are known, the distances between points A, B, C, and D can be obtained. Thereafter, the corresponding time values are obtained. This process can be generalized to multi-cell scenarios.

The total time t_{tot} is calculated from $t_{tot} = l_{tot}/s$, where l_{tot} is the total travel distance that can be obtained through geometric relationships [4], and s is the speed that the MU is traveling which is assumed to be constant in this study.

D.4.2 Cell Selection Strategies

When an MU enters an area where two or more cells intersect, the MU needs to take a network-assisted decision regarding which cell it will be associated with according to a specific criterion. We propose below three strategies that an MU can use for cell selection when it is covered by two or more cells, referred to as Str_1, Str_2, and Str_3 respectively.

- Str_1: *The cell with the highest steady state probability at the idle state will be selected.*
- Str_2: *The cell with the lowest state holding time at the failed state* will be selected.
- Str_3: *The cell with the lowest occupied-to-failed transition probability will be selected.*

With Str_1, the MU will compare the steady state probabilities of the overlapping cells at the idle state, i.e., $\pi_{\mathbf{1}}(j),\ j = 1, 2, \cdots, M$, for each cell. Then it is associated with the

cell which has the highest $\pi_{\mathbf{1}}$. On the other hand, Str_3 is designed to give higher priority to ongoing communications. When the MU is inside an overlapped area with an ongoing session, it requires higher priority so that its ongoing communication can be completed successfully. To do so, Str_3 selects a cell which has the lowest transition probability from the busy to failed state, i.e., p_{23}.

For Str_2, the MU selects a cell with the aim of minimizing the time spent at the failed state.To do so, a decision variable based on channel state holding times is needed. Denote by $\gamma_{\mathbf{3}}$ the number of time steps the system spent in the failed state during the visit. The distribution of the state holding time $P[\gamma_{\mathbf{3}} = h]$, where h is the number of time steps, follows a geometric distribution with parameter p_{33}. Therefore, from geometric distribution, it is known that $P[\gamma_{\mathbf{3}} = h] = (1 - p_{33})p_{33}^{h-1}$. Denote the expected value of this distribution as $E[\gamma_{\mathbf{3}}]$, then

$$\begin{aligned} E[\gamma_{\mathbf{3}}] &= \sum_{h=1}^{\infty} h(1-p_{33})p_{33}^{h-1} = \sum_{h=1}^{\infty}\left(hp_{33}^{h-1} - hp_{33}^{h}\right) \\ &= \sum_{h=0}^{\infty}\left((h+1)p_{33}^{h} - hp_{33}^{h}\right) = \sum_{h=0}^{\infty} p_{33}^{h} = \frac{1}{1-p_{33}}. \end{aligned} \tag{D.37}$$

From the expected values for $\gamma_{\mathbf{3}}$ for all cells, the MU selects the cell which provides the minimum holding time at the failed state. Furthermore, if two or more cells have exactly the same value for $\pi_{\mathbf{1}}$, $E[\gamma_{\mathbf{3}}]$, or p_{23}, no reliability based handover will take place. Tab. D.6 summarizes the primary selection conditions of these strategies and the selected cell at the intersection based on the network scenario shown in Fig. D.1.

D.5 Obtained Per-user Availability and Discussions

Consider a region of interest as a unit area $(x_{tot}, y_{tot}) = (1, 1)$ within which an MU is moving according to the two mobility patterns discussed in Sec. D.3. In order to obtain per-user availability for a multi-cell scenario as defined in (D.36), we need to calculate the t_{in} duration inside each cell for all cells that the MU traverses through. It is obtained through simulations as follows.

Table D.6: Cell selection for 3 strategies at intersection. In this table, $p_{23}(j),\ j = 1, 2, \cdots, M$ denotes the transition probability from the occupied state to the failed state of cell j

Strategy	**Main criterion for cell association**	**Selected cell at intersection, Cell C_s**
Str_1	Steady state probability of the idle state	$C_s = \underset{j}{\text{argmax}}\ \pi_{\mathbf{1}}(j)$
Str_2	State holding time of the failed state	$C_s = \underset{j}{\text{argmin}}\ E[\gamma_{\mathbf{3}}(j)]$
Str_3	Occupied-to-failed transition probability	$C_s = \underset{j}{\text{argmin}}\ p_{23}(j)$

Observe the movement of the MU at a pre-defined discrete time interval t_{step}. The MU evaluates whether or not its position is inside the coverage area of a cell or multiple cells. If it is inside a single cell, its associated duration to that cell t_{in} will be incremented by t_{step}. If it is covered by two or more cells, the MU needs to select a cell to associate with according to one of the cell selection strategies presented above and obtain t_{in} in that cell accordingly. This process will continue until the MU has reached its destination or the observation duration has elapsed.

The simulation results shown below are the average values obtained based on multiple typologies for a 10-cell PV network. The state transition probabilities were selected randomly based on uniform distribution with two sets of ranges, marked as rng_1 and rng_2 respectively, as shown in Tab. D.7. These ranges are selected to reflect various channel conditions. For instance, a failed channel's recovery probability in rng_2 exhibits a larger value than that of rng_1, leading to higher channel availability when rng_2 is adopted.

D.5.1 Multi-cell Scenario with MP1

Consider a heterogeneous network with cell sizes configured as $r = 0.3$ and $r = 0.4$ respectively. The MU traverses from $(0, 0)$ to (x_{tot}, y_{tot}) according to MP1, at a constant speed of $s = 0.01$ distance per unit time and with a given angle of departure, θ, for each journey. Fig. D.3 illustrates the availability results obtained based on the transition probabilities mentioned in rng_2 of Tab. D.7, as the angle of departure varies. Clearly, the obtained availability considering only the CR is always higher than that of the UR since the presence of RIs degrades availability. For per-user availability with RIs, Str_1 generally provides higher mean availability compared with Str_2 and Str_3. This is because Str_1 directly relates channel availability with per-user availability as defined in (D.36). When comparing Str_2 and Str_3, the former one performs better or equally well. This is because Str_2 focuses on minimizing the time spent in the failed state and the channel has a higher probability to return to the idle state from a failed state. Despite the fact that Str_1 performs best, Str_2 or Str_3 may be employed since they provide greater flexibility to users for faster channel recovery or for protection of ongoing traffic. Furthermore, higher availability is achieved when the angle of departure is not too close to the x- or y-axis since there is a high probability that the MU will be covered by the CR when it travels closer to the center of the region of interest.

Table D.7: Two configuration sets of transition probability ranges

	p_{12}	p_{21}	$p_{23} = p_{13}$	p_{31}
rng_1	$(0.1 \sim 0.2)$	$(0.8 \sim 0.9)$	$(0.0002 \sim 0.0005)$	$(0.8 \sim 0.9)$
rng_2	$(0.01 \sim 0.1)$	$(0.9 \sim 0.99)$	$(0.0002 \sim 0.0005)$	$(0.9 \sim 0.99)$

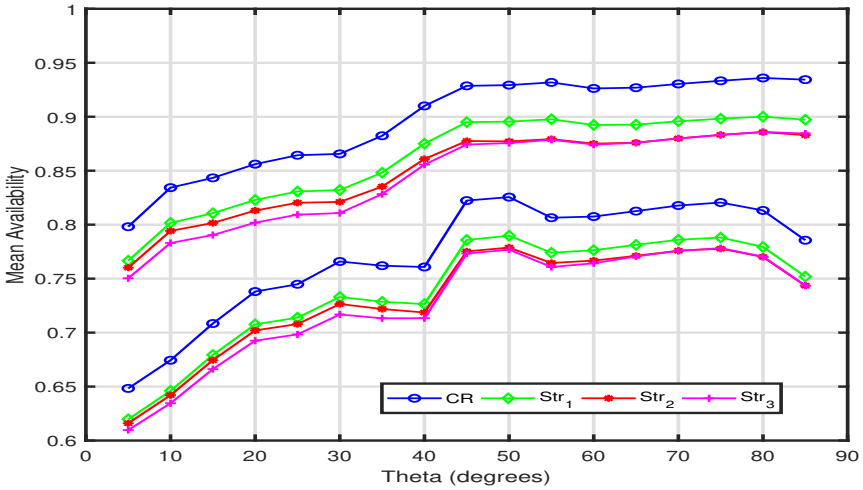


Figure D.3: Mean per-user availability in a 10-cell network with MP1. CR indicates the availability when only the coverage region is considered.

D.5.2 Multi-cell Scenario with MP2

In this case, the MU starts its journey from $(0, 0)$ and traverses through the area according to MP2 until the simulation time ends. The direction and the travel time for each phase of the journey are randomly selected from $(0.1, 0.5)$ time units and $(-\pi, \pi)$ radian ranges respectively. The travel speeds are configured from two ranges, i.e., $s_{high} \in (0.1 \sim 0.15)$ and $s_{low} \in (0.05 \sim 0.1)$ per unit time.

Fig. D.4 illustrates the average availability values obtained from multiple randomly selected trajectories according to the aforementioned configurations. The x-axis in this figure represents the speed and transition probability range combinations. From the results, it can be observed that the transition probabilities related to rng_2 provide higher availability compared with rng_1. Having lower p_{12} values with higher p_{21} and p_{31} values results in higher channel access opportunities for new users. Moreover, a lower speed gives higher availability. This is because MU traveling at a higher speed will traverse across the CR more quickly and stay at the boundaries of a region for a longer period of time. As the network coverage is comparatively poor near the boundaries, the obtained availability becomes lower. Finally, the impact of cell selection strategies is similar to what is observed in the MP1 case.

D.5.3 Further Discussions on Availability in URC/URLLC

D.5.3.1 Availability with RIs

Overall the obtained availability levels presented in this study are generally lower than what is regarded as the URC level in the literature [2] [3]. This is because most prior work on URC/URLLC is performed based on an implicit assumption that MUs are always covered. In this study, we consider that an MU may be located outside the CR boundary for our per-user availability definition. Therefore, the proposed definition (D.35) provides a more generic and accurate metric for evaluating anytime and anywhere communication *quantitatively*, since no assumption on the availability of coverage is made herein.

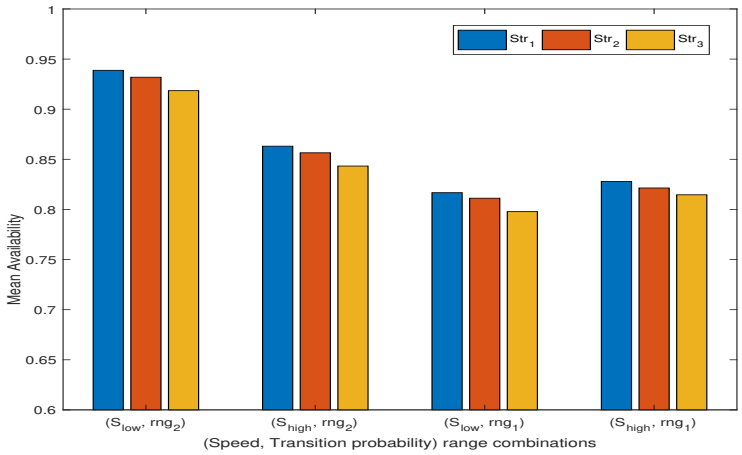


Figure D.4: Mean availability for an MU in a 10-cell network with MP2.

D.5.3.2 Shortest path versus highest availability path

Another interesting scenario to apply the proposed metric is to consider the highest availability route for user mobility. Instead of taking the shortest path towards a destination, an MU could select the path which gives the highest availability in order to satisfy its service requirements for ultra-reliable communication.

D.6 Conclusions and Future Work

This letter presents a time-space domain per-user availability analysis in 5G networks when reliability impairments are taken into consideration. We demonstrate that higher availability could be achieved by designing a proper strategy to compensate the negative effect of RIs on availability. For future work, we plan to enrich our analysis by including different RIs based on real-life measurements.

References

[1] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8–15, Apr. 2018.

[2] 3GPP TR38.824, "Study on physical layer enhancements for NR ultra-reliable low latency communication (URLLC)," R16, v0.0.1, Oct. 2018.

[3] H. V. K. Mendis and F. Y. Li, "Achieving ultra reliable communication in 5G networks: A dependability perspective availability analysis in the space domain," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2057–2060, Sep. 2017.

[4] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Per-user availability for ultra-reliable communication in 5G: Concept and analysis," in *Proc. IEEE WCNC*, Apr. 2018, pp. 1–6.

[5] K. Marashi, S. S. Sarvestani, and A. R. Hurson, "Consideration of cyber-physical interdependencies in reliability modeling of smart grids," *IEEE Trans. Sustain. Comput.*, vol. 3, no. 2, pp. 73–83, Apr.-Jun. 2018.

[6] C.-H. Liu and L.-C. Wang, "Random cell association and void probability in Poisson-distributed cellular networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 2816–2821.

[7] M. Garetto and E. Leonardi, "Analysis of random mobility models with partial differential equations," *IEEE Trans. Mobile Computing*, vol. 6, no. 11, pp. 1204–1217, Nov. 2007.

[8] B. Ku, Y. Ren, J. Weng, J. Chen, and W. Chen, "Modeling and analysis of channel holding time and handoff rate for packet sessions in all-IP cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3331-3344, Apr. 2017.

[9] J. Long, M. Dong, K. Ota, A. Liu, and S. Hai, "Reliability guaranteed efficient data gathering in wireless sensor networks," *IEEE Access*, vol. 3, pp. 430-444, Apr. 2015.