Link-Layer Rate of Multiple Access Technologies with Short-Packet Communications for uRLLC

Muhammad Amjad

School of Computer Science and Electronic Engineering,

University of Essex

A thesis submitted for the degree of

Doctor of Philosophy

Sep, 2021

I would like to dedicate this thesis to Dr. Mubashir Husain Rehmani

Abstract

Mission-critical applications such as autonomous vehicles, tactile Internet, and factory automation require seamless connectivity with stringent requirements of latency and reliability. These futuristic applications are supported with the service class of ultra reliable and low-latency communications (uRLLC). In this thesis, the performance of core enablers of the uRLLC, non-orthogonal multiple access (NOMA), and NOMA-random access (NOMA-RA) in conjunction with the short-packet communications regime is investigated.

More specifically, the achievable effective capacity (EC) of two-user and multi-user NOMA and conditional throughput of the NOMA-RA with short-packet communications are derived. A closed-form expressions for the EC of two-user NOMA network in finite blocklength regime (short-packet communication) is derived, while considering transmissions over Rayleigh fading channels and adopting a practical path-loss model. While considering the multi-user NOMA network, the total EC of two-user NOMA subsets is derived, which shows that the NOMA set with users having distinct channel conditions achieve maximum aggregate EC.

The comparison of link-layer rate of NOMA and orthogonal multiple access (OMA) shows that OMA with short-packet communications outperformed the NOMA at low SNR (20dB). However, at high SNR region (from 20dB to 40dB), the two-user NOMA performs much better than OMA. To further investigate the impact of the channel conditions on the link-layer rate of NOMA and OMA, the simulation results with generalized fading model, i.e., Nakagami-m are also presented.

The NOMA-RA with short-packet communications is also regarded as the core enabler of uRLLC. How the NOMA-RA with short-packet communications access the link-layer resources is investigated in detail. The conditional throughput of NOMA-RA is derived and compared with the conventional multiple access scheme. It is clear that NOMA-RA with optimal access probability region (from 0.05 to 0.1) shows maximum performance. Finally, the thesis is concluded with future work, and impact of this research on the industrial practice are also highlighted.

Acknowledgements

First of all, I want to express my deepest gratitude to my supervisor Professor Leila Musavian and my mentor and co-author Professor Sonia Aïssa. This four years jounrey of my PhD was not possible without their support, guidance, and help. Their dedication to high standard research has polished my research approach and changed my way of thinking. They have taught me, how I can think out of the box.

I also want to thank Dr. Mubashir Husain Rehmani for his continuous guidance during my research work. He taught me the lessons of perseverance, hope, and never give up. He is my teacher, co-author as well as my spiritual leader. Because of his relentless efforts, I have improved my writing skills.

I would also like to thank the University of Essex for funding my PhD and providing me opportunity to excel in the field of wireless communications.

All this was not possible without the prayers of my parents and family. My family is everything to me. Their support and love was a source of motivation for my stay at university of Essex. My brothers and sisters always encourage me to pursue higher studies. They were always in my thoughts during the whole journey of my PhD.

Last but not least, I want to thank my wife for making my life beautiful and memorable. She is my best friend and without her continuous support this journey was not possible.

Publications

Journal Papers

1. **M. Amjad**, L. Musavian, and M. H. Rehmani, "Effective Capacity in Wireless Networks: A Comprehensive Survey", IEEE Communications Surveys and Tutorials., vol. 21, no. 4, pp. 3007–3038, 2019.
2. **M. Amjad**, L. Musavian, and Sonia Aïssa, "NOMA versus OMA in Finite Blocklength Regime: Link-layer rate performance," IEEE Transactions on Vehicular Technology, vol. 69, no. 12, pp.16253-16257, 2020.
3. **M. Amjad**, L. Musavian, and Sonia Aïssa, "Effective Capacity of NOMA with Finite Blocklength for Low-Latency Communications", submitted in IEEE Transactions on Wireless Communications (Under Revision).

Conference Papers

1. **M. Amjad**, L. Musavian, "Performance analysis of NOMA for Ultra-Reliable and Low-Latency Communications," in IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, Dec. 2018, pp. 1–5.
2. **M. Amjad**, L. Musavian, and Sonia Aïssa, "Link-Layer Rate of NOMA with Finite Blocklength for Low-Latency Communications", in IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), London, 2020, pp. 1-6

Contents

Contents vi

List of Figures x

Nomenclature xii

1 Introduction 1

1.1 Motivations 1

1.2 Enablers of uRLLC 2

1.2.1 Short-Packet Communications 3

1.2.2 Non-orthogonal Multiple Access 3

1.2.3 Non-orthogonal Multiple Access based on Random Access 4

1.3 Tools to Investigate the uRLLC 4

1.3.1 Effective Bandwidth and Effective Capacity 5

1.4 Thesis Outline and Contributions 5

2 Related Work 11

2.1 Use Cases of 5G and beyond 5G 11

2.1.1 Enhanced Mobile Broadband (eMBB) 11

2.1.2 Massive Machine-type Communications (mMTC) 12

2.1.3 Ultra-Reliable and Low-latency Communications (uRLLC) 12

2.1.3.1 Latency Requirements 13

2.1.3.2 Reliability Requirements 14

2.1.3.3 Applications of uRLLC 15

2.1.3.4 Research Challenges faced by uRLLC 16
2.2 Non-orthogonal Multiple Access 17
2.2.1 Power Domain NOMA: 18
2.2.2 Code Domain NOMA: 18
2.2.3 Advantages of NOMA 20
2.2.3.1 Massive Connectivity: 20
2.2.3.2 Spectrum Efficiency: 20
2.2.3.3 Low Latency: 20
2.2.4 Challenges faced by NOMA 21
2.2.4.1 Imperfect SIC: 21
2.2.4.2 Resource Allocation: 22
2.2.4.3 Proper CSI: 22
2.2.4.4 Practical Implementation of NOMA: 22
2.3 Non-orthogonal Multiple Access based Random Access 23
2.4 Short-Packet Communications 24
2.5 An overview of Effective Capacity and Effective bandwidth 25
2.5.1 Effective Bandwidth 28
2.5.2 Effective Capacity 30
2.6 Effective Capacity analysis with different Fading models 31
2.6.1 Stochastic Fading Models 32
2.6.1.1 Rayleigh Fading Channels 32
2.6.1.2 Nakagami-m 33
2.6.1.3 Rician 34
2.7 Summary 34

3 Effective Capacity of NOMA with Finite Blocklength for Low-Latency Communications 36
3.1 Introduction 36
3.2 System Model 40
3.3 Theory of Effective Capacity 43
3.3.1 Effective Capacity in Finite Blocklength Regime 44
3.4 Effective Capacity of Downlink Two-User NOMA with Finite Blocklength 45

3.4.1 Achievable Effective Capacity of Strong-User NOMA with Finite Blocklength . . . 45
3.4.2 Achievable Effective Capacity of Weak-User NOMA with Finite Blocklength . . . 48
3.4.3 Achievable Effective Capacity of Multiple NOMA Pairs in Finite Blocklength . . . 50
3.5 Effective Capacity of Downlink Two-User NOMA with Finite Blocklength at High Transmit SNRs . . . 52
3.5.1 Effective Capacity of Downlink Two-User NOMA with Finite Blocklength at Extremely High Transmit SNR ($\rho \to \infty$) 53
3.6 Numerical Results . . . 54
3.7 Summary . . . 70

4 NOMA versus OMA in Finite Blocklength Regime: Link-Layer Rate Performance 72
4.1 Introduction . . . 72
4.2 Transmission Framework and Fundamentals . . . 76
4.3 Effective Capacity of NOMA and OMA in Finite Blocklength Regime 78
4.4 Numerical Results . . . 83
4.5 Summary . . . 96

5 Short-Packet Assisted Non-orthogonal Multiple Access Based Random Access 97
5.1 Introduction . . . 97
5.2 System Model and Transmission Framework . . . 102
5.3 Performance Evaluation . . . 107
5.4 Summary . . . 116

6 Conclusions and Future Works 117
6.1 Conclusions and Discussions . . . 117
6.2 Future Research Directions . . . 120
6.3 The Impact of this Research on Industrial Practice . . . 122

Appendix A 124

Appendix B 127

Appendix C 130

References 132

List of Figures

2.1 Two-user NOMA basic architecture. 19

2.2 Basic components involved in the communications of packet switched networks. In this packet-based communications system, different components of physical and link-layer have been illustrated which shows the difference between physical and link-layer channel modelling. 27

3.1 Two-user NOMA operation with finite blocklength with their respective queues: (a) describes the system model with two queues at the BS with their respective receivers, and (b) depicts the equivalent queueing model with the arrival rate and service rate. 42

3.2 Effective Capacity of NOMA weak-user and strong-user versus transmit SNR, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$. 56

3.3 Effective Capacity of NOMA strong-user versus transmit SNR under Nakagami-m fading, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$. . . 57

3.4 Effective Capacity of NOMA weak-user versus transmit SNR under Nakagami-m fading, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$. 58

3.5 Effective capacity of the weak and strong users versus ρ, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$. 60

3.6 Total effective rate of multiple NOMA pairs versus transmit SNR, with $\theta = 0.01$, $\epsilon = 10^{-5}$, and $V = 6$. 61

3.7 Effective capacity of NOMA strong and weak users versus delay exponent θ, with $\epsilon = 10^{-6}$. 62

3.8 Effective capacity of NOMA strong user versus transmission error probability (ϵ), with $n = 400$, and $\theta = 0.01$. 63

3.9 Effective capacity of NOMA weak user versus transmission error probability (ϵ), with $n = 400$ and $\theta = 0.01$. 64
3.10 Queuing delay violation probability versus QoS exponent (θ) for the strong user, with $D_{\max} = 400$, $\epsilon = 10^{-6}$, and $n = 400$. 66
3.11 Queuing delay violation probability versus QoS exponent (θ) for the weak user, with $D_{\max} = 400$, $\epsilon = 10^{-6}$, and $n = 400$. 67
3.12 Queuing delay violation probability versus transmission error probability (ϵ) for the strong user, with $D_{\max} = 100$, $n = 100$, and $\rho = 20$dB. 68
3.13 Queuing delay violation probability versus transmission error probability (ϵ) for the weak user, with $D_{\max} = 100$, $n = 100$, and $\rho = 20$dB. 69

4.1 Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR with Rayleigh fading channel. 84
4.2 Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under severe fading (one-sided Gaussian) with stringent delay, i.e., $\theta = 0.01$. 85
4.3 Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under severe fading (one-sided Gaussian) with less stringent delay, i.e., $\theta = 0.001$. 86
4.4 Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under Rician and lognormal fading with stringent delay, i.e., $\theta = 0.01$. 88
4.5 Total achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under different fading conditions with less stringent delay, i.e., $\theta = 0.001$. 89
4.6 Total achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under different fading conditions with stringent delay, i.e., $\theta = 0.01$. 90
4.7 Total achievable EC of two-user NOMA and two-user OMA versus the transmit SNR. 92

4.8 Total achievable EC of multiple NOMA pairs and multiple OMA users versus the transmit SNR with 6 users out of 12 users. . . . 93
4.9 Achievable EC of two-user NOMA and two-user OMA versus delay exponent (θ). . . . 94
4.10 Achievable EC of two-user NOMA and two-user OMA versus delay exponent (θ). . . . 95

5.1 NOMA-RA Basic operation with the short-packet communications assisted PUSCH-SPT from BS. . . . 103
5.2 Conditional throughput of the NOMA-RA and multi-channel ALOHA short-packet communications versus subchannels. . . . 108
5.3 Conditional throughput of the NOMA-RA short-packet communications versus different power levels. . . . 110
5.4 Conditional throughput of the NOMA-RA short-packet communications versus power levels, while considering different values of access probability. . . . 111
5.5 Conditional throughput of the NOMA-RA and multi-channel ALOHA short-packet communications versus access probability, with different values of power levels. . . . 112
5.6 Conditional throughput of the NOMA-RA short-packet communications versus access probability, with different values of short access error probability. . . . 114
5.7 Conditional throughput of the NOMA-RA short-packet communications versus short access error probability. . . . 115

Chapter 1

Introduction

1.1 Motivations

The ballooning growth of new applications, such as tactile Internet, massive sensing, holographic teleportation, and autonomous vehicles, poses serious challenges in terms of provisioning the spectrum efficiency, higher reliability, and lower latency [1]. The 5th generation (5G) and beyond 5G (B5G) promise seamless connectivity, enhanced capacity, higher reliability, and low-latency, through its use-cases of ultra-reliable and low-latency communications (uRLLC), and massive machine-type communications (mMTC) [2]. To meet the target objectives, research on above mentioned use cases has proliferated, in particular on uRLLC as a promising solution for the applications of overlapping areas of Internet of things (IoT) and tactile Internet. Meeting the stringent requirements of latency and reliability for uRLLC demands both revolutionary and evolutionary changes in the conventional wireless communications paradigm [3, 4]. This is due to the fact that among all the use cases of 5G and beyond 5G, uRLLC pose serious challenges in terms of ultra-reliability and ultra-latency. For example, when an attempt is made to improve the reliability by using redundancy, re-transmission or parity, it results into an increase of the latency.

Provisioning of ultra high reliability (characterized as overall packet loss probability) and very low latency (characterized as the end-to-end delay) for uRLLC use case is even very difficult in very simple communications scenarios, when the

users are not very far from base stations (BSs) and are only adjacent to very limited number of BSs [5]. To meet the stringent latency and reliability requirements of uRLLC, existing work focuses on the reduction of the end-to-end (E2E) delay by reducing the coding delay and the transmission delay. And for the reliability its main focus is the transmission errors. However, the queuing delay and the packet loss resulting from the queuing delay violation probability are not well investigated as part of the latency and reliability constraints for achieving the uRLLC. As a result, the most prominent use case of uRLLC for 5G and B5G under the constraint of queuing delay and queuing delay violation probability is not well understood. Also which tools should be used to analyse the performance of the uRLLC under stringent delay and reliability constraints are not clear [6, 7].

It is of capital importance to investigate the challenges faced by the uRLLC to ensure the low latency and ultra reliability for the emerging and futuristic applications. Very limited theoretical work has been done that combines the latency and the reliability. The pioneer work by Polyanskiy [8] provides the bounds on the block error rate for short-packet communications, however still it does not investigate the queuing delay and queuing delay violation probability. This invites the researchers from academia and industry to revisit the enabling technologies and tools to achieve and investigate the stringent latency and reliability requirements for uRLLC. In the next sections, there is the description of the major enabling technologies and tools that are used to ensure the stringent requirements of latency and reliability for uRLLC.

1.2 Enablers of uRLLC

To meet the stringent requirements of latency and reliability for the use case of uRLLC many enabling technologies have been outlined in the literature. To name a few are the short-packet communications (finite blocklength)[1], non-orthogonal multiple access (NOMA), NOMA-random access (NOMA-RA), spatial diversity, machine learning, slicing, network coding, and caching and mobile edge computing [9, 10, 11]. To investigate the performance of all the enabling technologies for

[1]The terms finite blocklength and short-packet communications will be used alternatively throughout the thesis.

uRLLC is beyond the scope of this work. This work only studies the performance of short-packet communications and NOMA, and NOMA-RA technologies, because NOMA and NOMA-RA in conjunction with the short-packet communications not only ensures the low latency and higher reliability requirements but can also provide the massive connectivity, spectrum efficiency, and higher throughput [12, 13]. Below is the brief overview of the short-packet communications and the NOMA technology.

1.2.1 Short-Packet Communications

The usage of the short-packet communications results into the reduction of the transmission delay which makes it a perfect candidate for uRLLC and massive machine type communications (mMTC) [14]. Most of the uRLLC and mMTC applications also require small amount of data for their smooth operations, which further confirms the place of short-packet communications as an enabler technology [15]. Though the shortening of packets can be straight forwardly convincing as a mean to achieve uRLLC, it poses serious challenges such as capacity penalty. Furthermore, to meet the constraints on reliability, the channel codes used for the short-packet communications should also be strong enough [8].

As known, traditional wireless systems are designed based on the Shannon theory without stringent reliability constraint, where long packets are considered in the communications. The Shannon limit, however, is a loose upper bound for the performance of systems with short-packet communications. Therefore, the achievable rate for the short-packet communications should also take into consideration the capacity penalty and transmission error probability. The more detail on the short-packet communications is provided in Chapter 2.

1.2.2 Non-orthogonal Multiple Access

As compared to serving the single user with a dedicated bandwidth resource block as in orthogonal multiple access (OMA), NOMA allows multiple users to occupy the same time and frequency resource block. In this way, NOMA can accommodate multiple users by breaking the conventional orthogonal resource allocations principle [16, 17]. While residing within grant-free transmission, NOMA with

the finite blocklength is also considered as a key solution for provision of low latency, massive connectivity, and higher reliability. The principle of NOMA in finite blocklength regime follows the conventional concept of NOMA, with superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver [12, 13].

The existing NOMA schemes can be categorized into the (i) Power domain NOMA and (ii) Code domain NOMA. In this work, the main focus is on the power-domain NOMA. Using the power domain NOMA or power domain multiplexing, depending on the channel conditions different users are allocated different power levels. These power levels are key for keeping the different users separate and to mitigate the multi-user interference using SIC [18]. A detailed discussion with its basic operation is given in Chapter 2, Chapter 3, and Chapter 4.

1.2.3 Non-orthogonal Multiple Access based on Random Access

NOMA is well suited for the coordinated transmissions where power levels of different signals are used for the multiple access and SIC is used to remove the co-channel interference. However, the benefits of the NOMA can also be exploited for the non-coordinated transmission such as random access. Therefore, NOMA in conjunction with the random access (NOMA-RA) can be used in scenarios where the number of subchannels are limited [19]. So, NOMA-RA can increase the throughput of the existing random scheme without any bandwidth expansion [20]. The more detail of the NOMA-RA is provided in Chapter 2, and Chapter 5.

1.3 Tools to Investigate the uRLLC

Although an extensive work has been done on uRLLC, however, still the ongoing research work on uRLLC is in its early stage. Which tools should be used to investigate the performance of the uRLLC while remaining within the stringent latency and reliability constraint, is a question that requires a thorough discussion. Many tools have been proposed in the literature to characterize the delay

and reliability components of uRLLC. Some are meta distribution, extreme value theory, stochastic network calculus, and effective bandwidth and effective capacity [9]. Employing all the above tools to study the uRLLC is out of the scope of this thesis. In this thesis, effective capacity framework have been employed to investigate the enablers of the uRLLC (NOMA and short-packet communications). Below is the brief description of these two tools:

1.3.1 Effective Bandwidth and Effective Capacity

Effective bandwidth and effective capacity are based on the large deviation principle. Effective capacity is the dual concept of the bandwidth, therefore it is pertinent to first explain the effective bandwidth before effective capacity. Effective bandwidth can be defined as the minimum constant service rate that is needed for a given source rate under the queuing delay constraint. While effective capacity is used to find the maximum arrival rate for a given service rate under the constraint of queuing delay [21]. While other tools do not consider the queuing delay, these tools provide a greater flexibility in investigating the uRLLC for the emerging applications of 5G and beyond 5G networks. A complete section in Chapter 2 is dedicated to the explanation of these two tools.

1.4 Thesis Outline and Contributions

Table 1.1 shows the list of the acronyms and the table 1.2 shows the list of the mathematical notations used in this thesis.

The rest of the thesis is organized as follows:

- In Chapter 2, there is an overview of the existing work on the uRLLC, enablers of uRLLC, i.e.; short-packet communications, NOMA, and NOMA-RA. Provision of higher reliability and low-latency for delay sensitive applications is the major requirement for this critical use case (uRLLC) of 5G. It is clear from the related work that although much work has been done on the uRLLC but a thorough investigation that encompass the delay (more specifically the queuing delay), and reliability, constraints for achieving the uRLLC is still needed. The applications supported by the uRLLC

Table 1.1: List of acronyms and corresponding definitions.

Acronyms	**Definitions**
5G	Fifth Generation
AR	Augmented Reality
AMC	Adaptive Modulation and Coding
AWGN	Additive White Gaussian Noise
BS	Base Station
BLER	Block Error Rate
CSI	Channel State Information
DNNs	Deep Neural Networks
E2E	End-to-End
eMBB	Enhanced Mobile Broadband
EC	Effective Capacity
FD	Full-Duplex
IoT	Internet of things
ITU	International Telecommunication Union
mMTC	Massive Machine-type Communications
MIMO	Multiple-Input and Multiple-output
mURLLC	Massive Ultra Reliable and Low Latency Communications
NOMA	Non-Orthogonal Multiple Access
NOMA-RA	NOMA-random access
OMA	Orthogonal Multiple Access
PDF	Probability Density Function
PMF	Probability Mass Function
PUSCH	Physical Uplink Shared Channel
PBCH	Physical Broadcast Channel
PRACH	Physical Random Access Channel
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
PUSCH-SPT	Physical Uplink Shared Channel for Short-Packet Transmission
QoS	Quality-of-Service
SIC	Successive Interference Cancellation
SNR	Signal-to-Noise Ratio
SISO	Single-Input and Single-Output
SIMO	Single-Input-Multiple-Output
SC	Superposition Coding
uRLLC	Ultra-Reliable and Low-Latency Communications
uMTC	Ultra-Reliable Machine type Communications
UAV	Unmanned Aerial Vehicle
URC	Ultra-Reliable Communications
UE	User Equipment
VR	Virtual Reality

Table 1.2: Summary Of Notations

V	total number of users	ϵ	transmission error probability
α_i	power allocation coefficient of user v_i, $i = \{u, t\}$	r_i	service rate for user v_i
$s_i(\tau)$	message intended for user v_i at time τ	δ_i	channel dispersion for user v_i
P	total transmit power at the BS	$Q^{-1}(.)$	inverse of Gaussian Q-function
n	blocklength	$a_i(\tau)$	number of queue packets at time τ
$h_i(\tau)$	channel coefficient between user v_i and the BS at time τ	θ_i	delay exponent of user v_i
y_i	receive signal at user v_i	$q_i(\infty)$	steady-state of transmit buffer
m_i	AWGN	$D^i_{\max}$	maximum delay
v_u	strong user	C^i_e	effective capacity for user v_i
v_t	weak user	$\mathbb{E}[.]$	expectation operator
ρ	transmit SNR	$\mathrm{H(a,b,z)}$	confluent hypergeometric function of the 2nd kind
N_oB	noise power	$\Gamma(.)$	Gamma function
SNR_u	received SNR at user v_u	$\mathrm{E_i(.)}$	exponential integral
SINR_t	resulting SINR at user v_t	ϕ	combination of all NOMA pairs
$f_{(i:V)}$	PDF of ordered channel gains	T_{ec}	total effective capacity
$B(a,b)$	Beta function	C^i_e	achievable EC of NOMA v_i user at high transmit SNR ($\delta_i = 1$)
V_i^{N}	Channel dispersion for NOMA users	V_i^{O}	Channel dispersion for OMA users
C_i^{N}	Effective Capacity of NOMA user	C_i^{O}	Effective Capacity of OMA user
$P_{AL}(D;K)$	throughput of the multichannel ALOHA	P_L	Power levels used for the NOMA-RA
D	active users	S_ϵ	short access error probability
$U_{NR}(P_L,K)$	conditional throughput of the NOMA-RA	K	subchannels for NOMA-RA

such as autonomous vehicles, factory automation, and smart grid, and the challenges faced by the uRLLC are also explained in detail in this chapter. There is also a thorough explanation for employing the short-packet communication and NOMA and not using the Shannon formula while deriving the achievable rate. The potential benefits of the NOMA such as massive connectivity, spectrum efficiency, and low latency, with their limitations are also highlighted in more detail. A detailed background on the mathematical framework, i.e.; effective capacity (EC), that is used to investigate the performance of NOMA, while residing in short-packet communications regime is provided. Inspired by the large-deviation theory and based on the effective bandwidth, effective capacity is a link-layer model that represents the maximum arrival rate that the given service process can support while residing within delay constraint.

- In Chapter 3, a detailed investigation of the link-layer rate performance of the uRLLC enablers, i.e., short-packet communications and NOMA is performed. Achievable effective capacity of the two-user NOMA (out of the V users) while residing within a short-packet communications regime is formulated. A closed-form expression for the achievable EC of two-users NOMA network with short-packet communications regime under Rayleigh fading channels is also derived. More specifically, the impact of delay exponent, transmit signal-to-noise ratio (SNR), transmission-error probability, and power coefficients on the EC of two-user NOMA has been shown. Through simulations, it is clear that under the stringent delay requirements and due to the dominant factor of transmission error probability, the queuing delay violation probability does not improve below a certain value. In addition to the two-users NOMA, total achievable link-layer rate of the multi-user NOMA network with short-packet communications is also derived and shows that the NOMA set with most distinct channel conditions achieve higher total EC as compared to the users with less distinct channel conditions. The numerical results of the achievable EC of two-users NOMA under finite blocklength regime over Nakagami-m fading channel are also analysed in detail.

- A continuation of Chapter 3, in Chapter 4 a comparative analysis of the orthogonal and non-orthogonal multiple access techniques with short-packet communications is provided. A link-layer rate of the two-users NOMA and OMA while residing within short-packet communications regime is derived. Deriving the achievable EC of two-users NOMA (out of V users), and deriving the achievable EC of only the two-users NOMA are completely different and complex due to short-packet communications. Therefore, in Chapter 3, the achievable EC of two-users NOMA (out of V users) is derived, while in Chapter 4, only the two-user NOMA set is considered and its achievable EC is derived. Under the delay and reliability constraint, this Chapter 3 provides further insights about the enablers of the uRLLC, i.e., short-packet communications and NOMA. Considering the transmission over Rayleigh fading channel, a closed-form expression for the achievable EC of the two-user NOMA and OMA network is provided. A comparison of the achievable EC of NOMA and OMA two-users shows that OMA strong user (user with better channel conditions) outperforms the NOMA users at low SNR under Rayleigh fading channel. Further, the total EC of the NOMA and OMA users is derived under different fading models. More generally, the Nakagami-m fading is used to investigate the achievable EC of two-users NOMA and OMA and the total link-layer rate (total achievable EC) of both transmission. Which shows that the NOMA outperforms the OMA at high SNRs under loose delay constraint. On top of that, the link-layer rate comparison for two-user NOMA and OMA while considering the impact of the fixed power coefficients, delay exponent and transmit SNR are also studied. Numerical results are investigated in detail and accuracy of the proposed closed-form expression is verified using the Monte-Carlo simulations.

- In Chapter 5, the benefits of the multi-channel slotted ALOHA (a conventional random access scheme) are combined with the NOMA. For this purpose, NOMA in conjunction with the random access scheme named as NOMA-random access (NOMA-RA) is proposed while residing within short-packet communications regime. The average throughput of the conventional multi-channel ALOHA with short-packet communications and the

conditional throughput of the NOMA-RA short-packet communications is derived. NOMA-RA users work with the different power levels and selects the power levels from the pre-determined set of power levels. It is shown that as the number of power levels increases the conditional throughput of the NOMA-RA increases while keeping the sub-channels constant. The impact of the access probability and the short-access error probability (due to the short-packet communications) on the conditional throughput is also investigated. More specifically, this analysis shows that increasing the access probability does not increase the throughput of the NOMA-RA and multi-channel ALOHA, but there exists an optima area of the access probability (from 0.05 to 0.1), where the NOMA-RA with short-packet communications shows the maximum throughput.

- In Chapter 6, a detailed summary of the thesis with conclusion and future research directions regarding the enablers of uRLLC and its related performance estimation tools is provided. In this chapter, the potential impact of this research on the industrial practices is also highlighted, which establishes the importance of the NOMA, NOMA-RA, and short-packet communications as a core enabler of the uRLLC.

Chapter 2

Related Work

2.1 Use Cases of 5G and beyond 5G

5G and beyond 5G (B5G) networks promise seamless connectivity, higher throughput, massive connectivity, and low latency. The emerging applications such as augmented reality, virtual reality, tactile Internet, holographic teleportation, and autonomous vehicles require ultra low latency and very high reliability [4]. Different applications of the 5G and B5G networks have different target objectives regarding latency, reliability, availability, and bandwidth. Depending upon these requirements, the work on the 5G and B5G networks is divided into three major service classes or use cases such as uRLLC, massive machine-type communications (mMTC), and enhanced mobile broadband (eMBB). These three use cases proposed by the International Telecommunication Union (ITU) specifically cope the stringent performance requirements such low latency and reliability, higher data rates, and massive connectivity. Below is the detailed discussion regarding each use case:

2.1.1 Enhanced Mobile Broadband (eMBB)

To provide a best user experience and higher data rates, eMBB use case of the 5G networks is considered as the natural evolution to the existing cellular networks

[22]. This use case has the potential to provide a seamless connectivity[1], better user experience, higher throughput, and greater user mobility. The challenging futuristic applications such as virtual reality (VR), augmented reality (AR), and 360 degree video streaming are well supported through this service category. Researchers from academia and industry are revisiting the diverse technologies to achieve the above mentioned stringent performance requirements for eMBB. Some of the proposed enablers for eMBB are millimeter wave communications, massive MIMO, and spectrum sharing [23, 24].

2.1.2 Massive Machine-type Communications (mMTC)

Through the use case of the machine type communications (MTC), massive connectivity for large number of low power devices for sending small data packets can be achieved [25]. This has been further classified into the massive machine type communications (mMTC) and ultra-reliable machine type communications (uMTC) [7]. Millions of sensors and other low complexity devices that fall under the umbrella of Internet of things (IoT), industrial IoT, Internet of everything, and smart cities are well investigated through this use case.

2.1.3 Ultra-Reliable and Low-latency Communications (uRLLC)

Of all the use cases of the 5G networks discussed above, uRLLC is the most challenging one due to its most stringent requirements for provision of very low latency and higher reliability. uRLLC supports a diverse range of applications such as autonomous vehicles, tactile Internet, Industry automation, smart grid, and tele-robotic control are few to name [9]. In addition to these, the use case of uRLLC is also of capital importance for provisioning of low latency and higher reliability for other applications of eMBB and uMTC use case such as VR, AR, and industrial IoT. Each application scenario has its own specific key performance

[1]Seamless connectivity means "smooth and continuous Internet connectivity". The main feature of the all the use cases and especially the enhance mobile broadband is to provide the seamless connectivity or smooth and continuous Internet connectivity to all the connected users at all time.

indicator and challenges associated with them [10]. The main focus of this thesis is on the use case of uRLLC, due to its broad coverage of applications and extensive challenges faced by it as compared to the mMTC and eMBB. Below is the description of the basic definitions and minimum requirements of latency and reliability, research challenges faced by the uRLLC, and its major application scenarios with key performance indicator.

2.1.3.1 Latency Requirements

To meet the stringent latency requirements for the uRLLC, the researchers from the academia and industry are redefining the conventional wireless network architecture. Overall latency faced by a packet in a wireless network is the combination of the end-to-end latency, control plane latency, and the user plane latency [9, 11, 23]. These are explained below in detail:

- End-to-End Latency: Following the 3GPP standard [23] the requirements for the end-to-end latency to achieve the uRLLC service class is 1ms. End-to-end latency not only includes the transmission latency, but also the queuing latency, decoding latency, processing latency, and the retransmission latency.

- Control Plane Latency: This is the state transition latency, i.e, latency in the system caused during the transition from idle to active state. According to the 3GPP standard, the control plan latency of 20ms is the stringent requirement for achieving the service class of uRLLC [9, 11].

- User Plane Latency: Latency faced by the packet while moving across a physical layer is termed as the user plan latency. This latency mostly depends on the number of users involved in the transmission. The 3GPP sets a standard of 1ms minimum requirements of user plane latency for uRLLC in case of one user [9, 11].

The above mentioned minimum requirements of latency for uRLLC can better be realized by revisiting the different existing enabling technologies and proposing the new ones. According to Mehdi Bennis et al, the main enabler for low latency

are NOMA, short-packet communications, machine learning, network slicing and coding, and grant-free access. NOMA and short-packet communications are explained in more detail in the next section of this chapter. The brief overview of the other enablers of low latency are discussed below:
Machine Learning: Machine learning framework more specifically the distributed ML concept provides the scalable and distributed solution on interconnected nodes for reducing the latency. This concept is termed as the AI on the edge.

Network Slicing and Coding: Network slicing and coding at the edge also reduces the latency significantly. This not only reduces the latency but also helps in allocating the network resources such as bandwidth and caching.

Grant-Free Access: With the use of the grant-free access the resource assignment phase for the uplink transmission could be shortened or skipped, that can be resulted in reduction of the latency. With this grant-free access the challenging applications requiring low latency for their operation can be achieved.

2.1.3.2 Reliability Requirements

In conventional terms, reliability is the successful transmission of packets while guaranteeing a delay constraint. To satisfy the minimum requirements of reliability for the uRLLC service class, various enablers and diverse changes in network architecture have been proposed [9, 11, 23]. Different definitions of reliability with their minimum requirements for achieving the uRLLC are described below:

- The main definition of the reliability is set by the 3GPP [23] as the "transmission of data with high successful probability within given time". To achieve the uRLLC the requirement of the conventional reliability is 10^{-5}.
- Control Channel Reliability: This reliability is all about the decoding errors or successful probability for decoding the metadata [9, 11].
- Per node Reliability: This is the combination of the different reliability scenarios such as probability of proactive packet drop, queuing delay violation probability, and probability of transmission errors [9, 11].

If only the reliability is concerned from the uRLLC, then there are multiple enablers of reliability to achieve the stringent requirements of reliability. According to Mehndi Bennis et al, the key enablers of reliability are network slicing and coding, short-packet communications, spatial diversity, and packet duplication.

2.1.3.3 Applications of uRLLC

Many of the mission critical services can be accomplished with the help of uRLLC. Below is the detail of the some of the applications of uRLLC with their potential challenges:

- Autonomous Vehicles: Autonomous vehicles are the key applications of the 5G and B5G networks covered by the uRLLC service class. uRLLC with its capability of 1ms latency and 99.999 % reliability enables the vehicles to communicate with other vehicles, pedestrians, and road side units in real time. Smart parking, collision avoidance, automated overtake, and smooth driving can be achieved with much precision and without human intervention. This application of the uRLLC will ultimately help in improving the road safety and driving efficiency. However, the high mobility and scalability are the key challenges faced by this application [26, 27].

- Factory Automation: Through this application of uRLLC, the concept of fourth generation industrial revolution can be better achieved. Reliability and latency related critical tasks in the factory such as tactile interaction and robot motion control can be perfectly automated. The use case of the uRLLC through this application can reduce the operational cost [26, 27].

- AR and VR: AR and VR pose serious challenge in terms of stringent latency and reliability requirements. AR is the enhanced interactive experience of the physical world, where through VR the virtual experience is created. uRLLC is regarded as the enabler for the AR and VR applications by reducing the latency and increasing the user comfort by satisfying the stringent latency and reliability requirements. However, processing and transmission of the 3D videos is the challenge faced by these applications of the uRLLC [26, 27].

- Smart Grid: The operation of the traditional grid is transferred into intelligent smart grid by introducing the smart decisions related to fault isolation and fault diagnosis. These smart real-time operations can now be controlled with the help of the uRLLC. The costly fibre and cable based fault diagnosis approach for any delay and reliability sensitive services for the remote power distribution lines can now be replaced with the help of the uRLLC use case [26, 27].

2.1.3.4 Research Challenges faced by uRLLC

The use case of the uRLLC is in its infancy stage. There are still a lot of challenges faced by the uRLLC. The details of some of them are explained below:

- Cross Layer Design: Current communications technologies are developed by taking into consideration the different layers of the open systems interconnection model. The changes in one layer significantly impact the other layers and ultimately on the E2E delay and reliability. Most of the current developments in communication do not take into consideration, the cross layer E2E delay and reliability [9]. This poses a most serious challenge in achieving the very low latency and ultra high reliability for the uRLLC.

- High Computing Overhead: The existing systems are extremely complex and dynamic. They have to cope with the varying wireless channel conditions under strict delay and reliability constraints. Sometimes, they have to solve the complex optimization problem for adjusting their resource allocation schemes. This results into high computing overhead, which is the hurdle in achieving the low latency and higher reliability [27].

- QoS Guarantee: The recent advances in the machine learning has enabled the existing systems to deploy the deep learning approached for uRLLC. However, still the QoS guarantee under stringent latency constraint becomes challenging when using the deep learning approach such as deep neural networks (DNNs) [27].

- Scalability: The emerging applications of 5G such as VR, AR, and factory automation are covered by the use case of uRLLC. However, in these ap-

plications, the density of the devices increases exponentially. Therefore, scalability is a serious challenge in guaranteeing the stringent latency and reliability requirements for uRLLC [26].

As already mentioned, among all the use cases of the 5G and B5G networks, uRLLC is the most challenging one due its stringent reliability and latency constraints. This use case is still in its infancy, however there is always an increasing attraction from the industry and academia about this service class of the 5G and B5G networks.

There are plenty of challenges faced by this research topic, i.e, uRLLC. As mentioned above major research challenges faced are the cross-layer design, high computing overhead, QoS guarantee, and scalability. Out of the above mentioned research challenges, the QoS guarantee for the uRLLC has been addressed in this work. The QoS has been modelled using the effective capacity tool and the delay exponent which is the characterization of the delay has been investigated in more detail.

In this thesis, the major focus is on this use case, i.e., uRLLC, due its immense applications in a wide variety of mission critical tasks. To achieve the uRLLC, there are multiple enablers such as NOMA, NOMA-RA, short-packet communications, proactive packet dropping, edge caching and computing, and multicasting [9]. To investigate all the enablers of the uRLLC is beyond the scope of this work. In this thesis, the NOMA, NOMA-RA, and short-packet communications due to their advance feature for supporting not only the low latency and higher reliability but also the massive connectivity, and spectrum efficiency are considered for further investigation as an enabler of uRLLC. For this purpose, the link-layer rate for the multiple access technique (NOMA and OMA) and conditional throughput for NOMA-RA is derived to understand the performance of the uRLLC.

2.2 Non-orthogonal Multiple Access

Satisfying the stringent requirement of latency and reliability for uRLLC is a challenging task. As mentioned above it requires a thorough revisiting of the existing conventional wireless architecture and enabling technologies. To meet

these objectives, NOMA has been proposed as the key enabler of uRLLC for the 5G and beyond B5G applications [16] . NOMA has attracted a lot of attention from the industry and academia for meeting the network level and data level quality of experience requirements.

As compared to the OMA schemes, in NOMA multiple users are accommodated on the same time and frequency resource block. This ultimately results into achieving the spectrum efficiency and massive connectivity. The basic operation of the NOMA system comprises of superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver. This basic principle of the NOMA proved to be completely compatible with other multiple access techniques and other physical layer emerging technologies such as mmWave and MIMO. Depending upon its operation, there are various types of NOMA such as

2.2.1 Power Domain NOMA:

In power domain NOMA, different users are superimposed on the resource block using the different power levels [18]. Depending on the channel conditions, different users are assiged the different power levels. This results into the interference in systems which requires the complete revisiting of existing interference mitigation schemes. Fig. (2.1) shows the operation of the power domain NOMA [28]. This is the basic architecture of the NOMA with two users (far user and near user). The far user is denoted by the v_t, and v_u is the near user, where the h_t and h_u are the channel gains for the far user and near user respectively. Far user (user with bad channel condition) and the near user (user with good channel condition) are allocated different power levels. Then the users are superimposed on the resource block at the transmitter. Then at the receiver (SIC) is used to decode and to remove the user's messages.

2.2.2 Code Domain NOMA:

As compared to the power levels, that are used in power domain NOMA, random Gaussian codes are used in code domain NOMA to make the distinction among users at the transmitter. Gaussian codes or Gaussian spreading is a design criteria for assigning the power levels to the users in Code-domain NOMA.

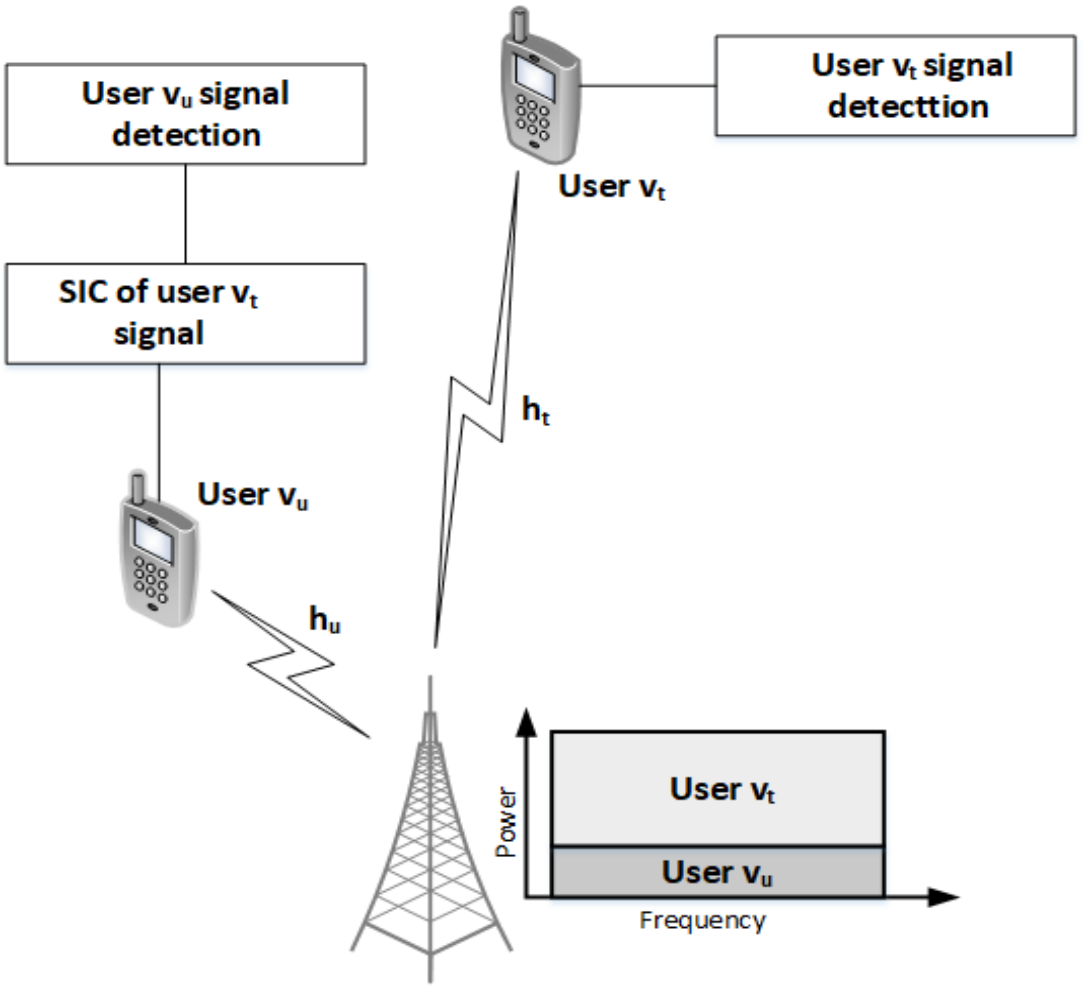

Figure 2.1: Two-user NOMA basic architecture.

These Gaussian codes or Gaussian spreading creates Gaussian like interference in other layers of signals that helps in SIC through multiple layers. And then subsequently, the compressive sensing is used at the receiver to decode the user message. In this way, the probability of the symbol errors is reduced [29]. Code domain NOMA can further be classified into the sparse code multiple access and low-density spreading.

This thesis resolves around the power domain NOMA due to the fact that power domain NOMA are less complex than the code domain NOMA. Below is the description of the main advantages achieved through NOMA.

2.2.3 Advantages of NOMA

NOMA has been extensively researched due to its potential benefits. Following are the core benefits that can be achieved from NOMA:

2.2.3.1 Massive Connectivity:

There is a prediction that the there will be 42 billion of connected IoT devices in 2025. To ensure a seamless connectivity for such a large number of devices, the concept of massive connectivity was introduced. To meet this objective, NOMA provides the massive connectivity while providing the compatibility with the existing access technologies [18]. As compared to the OMA, where only the single user is allocated a complete bandwidth resource block, in NOMA multiple users are allocated (can be accommodated) on the same bandwidth block and at the same time. In this way, the massive connectivity can be ensured via NOMA to provide the connectivity for the billion of devices in the future.

2.2.3.2 Spectrum Efficiency:

Existing radio spectrum is a scarce resource which is approaching its limits. Spectrum efficiency can not be simply achieved by adding more antennas in the system, as it will add more interference. NOMA promises the spectrum efficiency while residing within the existing resources. In NOMA, multiple users are multiplexed on the same bandwidth block. Therefore, multiple users are served without adding extra bandwidth in the system [6].

2.2.3.3 Low Latency:

NOMA has been regarded as the main enabler of the low latency communications. More specifically, grant-free NOMA can reduce the transmission latency. In some cases, NOMA reduces the congestion probability, thereby reducing the latency as well as the congestion in the wireless network. NOMA in conjunction with the grant-free access (grant-free NOMA) ultimately reduces the access delay [9].

Mainly, the advantages of the NOMA are spectrum efficiency, higher throughput, massive connectivity, and low latency. NOMA supports multiple users as

compared to its counterpart OMA. Due to this feature and using the more advance receiver schemes like SIC and grant-free access it reduces the latency and is considered as the enabler of the low latency communications.

2.2.4 Challenges faced by NOMA

Extensive work has been done on NOMA while considering its different dimensions. It is shown that NOMA is compatible with other state-of-the-art technologies such as MIMO, cognitive radio networks, cooperative communications, mmWave [17]. However, there are still some research concerns related to the NOMA. Below is the list of some challenges that need to be addressed to make the NOMA perfect choice for the uRLLC service class.

2.2.4.1 Imperfect SIC:

Downlink NOMA network operates with superposition coding at the BS and SIC at the receiver. The user with the strong channel condition (strong user) performs SIC and, then, detects and removes the weak-user message from its received signal. In some practical scenarios, such as hardware malfunctioning, decoding errors, or channel estimation errors, some interference may exist after the SIC, i.e., imperfect SIC. Imperfect SIC becomes significant when multiple (more than two) NOMA users are considered [30]. Therefore, the imperfect SIC is the serious challenge faced by the NOMA network, as this can significantly reduce the performance of the NOMA network. In this thesis work, the perfect SIC is considered. However, the existing work can be extended by taking into consideration the imperfect SIC.

NOMA supports and serve multiple users per resource block. This invites the excessive interference at the receiver end. SIC is the key player in mitigating the excessive interference due to the multiple users in NOMA operation. In this work, the perfect SIC is used while considering the simplicity of the receiver design. Also, NOMA in conjunction with the short-packet communication and employing the effective capacity tool and then deriving the closed-form expression is a difficult task. So considering the imperfect SIC within NOMA short packet communication scenario will make this system complex and is beyond the scope

of this work.

2.2.4.2 Resource Allocation:

Efficient user pairing and then allocating the power levels among NOMA users requires efficient algorithms to achieve the efficient utilizations of existing resources. Depending on the user pairing and transceiver designs, fixed or dynamic power allocation is another choice which the system designers have to make. In this thesis, the fixed power allocation is considered, where the users select one of the power levels from the set of the pre-determined power levels.

2.2.4.3 Proper CSI:

Usually in downlink NOMA, transmitter performs the user pairing and then allocates the power to users based on the channel state information (CSI). Therefore, it requires the system to be very efficient in handling the CSI feedback with proper reference signal. This factor if not addressed properly could result into the significant performance loss in power domain NOMA.

In this thesis the CSI is available at the BS. And based on the known CSI, BS performs the user ordering. It is a common practice in NOMA operation that BS performs the user ordering based on the known CSI. Based on the CSI, the BS orders the user into weak user and the strong user. However, if the channel uncertainty was considered in this work, then it became difficult for the BS to perform the user ordering and ultimately the SIC failed due to the excessive interference.

2.2.4.4 Practical Implementation of NOMA:

To verify the theoretical results of NOMA regarding its scalability, massive connectivity, spectrum efficiency, and low latency, the NOMA should be put in practice. Most of the existing work of NOMA is on the performance evaluation, which shows it as the potential enabler for the uRLLC and other 5G user cases. However, this requires some over the air demonstrations to verify it as the future enabler of uRLLC [18].

2.3 Non-orthogonal Multiple Access based Random Access

Random access techniques such as ALOHA and multiple channel ALOHA randomly select the channel and improve the efficiency of the system by reducing the access delay [31] and improving the resource efficiency. As compared to the CSMA technique, ALOHA and multi-channel ALOHA significantly reduces the access delay. In ALOHA technique, each user sends its message without sensing the transmission medium whether it is idle or busy. This ultimately reduces the overall transmission delay. However, the conventional random access schemes do not reduce the overall end-to-end delay. Also, the emerging demand of massive connectivity and spectrum efficiency are also not well supported in the existing random access schemes [32].

To take the advantage of the less access delay of conventional multichannel ALOHA, the NOMA can now be integrated with the multichannel ALOHA scheme, named as the NOMA-RA [33]. The conventional NOMA transmission is a coordinated transmission controlled via the BS. In this coordinated transmission, BS estimates the channel conditions of the users (far or near user) and then based on the channel conditions, BS allocates the resources (mainly power) to the NOMA users. This coordination does come at the cost of the large access delay. To address this issue, NOMA scheme in conjunction with the conventional random access scheme is now being used to achieve the ultra low latency, higher reliability, massive connectivity, and spectrum efficiency [34]. NOMA-RA is the non-coordination transmission and the users themselves pick the power levels either from the pre-determined power levels or using some power allocation scheme. In NOMA-RA, users select the different power levels and then randomly select the channel. BS receives the signal and performs the SIC and then decode and remove the user's signal. The chosen power levels by the users depend on the desired transmission rate required for the communications. However, if the multiple users selects the same power levels then the decoding error occurs during the SIC and the system witnesses the collision, named as the power collision.

In NOMA-RA different users select the different power levels and then on the receiver sides SIC is performed to mitigate the interference and to decode the

respective users messages. However, in some cases if the multiple users select the same power levels then the SIC results into decoding errors, this is called the power collision in NOMA-RA. Power collision significantly reduces the system performance and users messages are not decoded properly. In this case, the power level of one user collides with the same level picked up by the other users. This collision is similar to the collision of users messages in ALOHA technique. One of the algorithm called collision resolution period (CRP) [1] has been proposed in the literature to mitigate the power collision.

Recently NOMA-RA has attracted a lot of attention from industry and academia. Following are the core advantages achieved via NOMA-RA scheme.

- As the NOMA-RA is based on the NOMA and random access scheme, therefore it also shares the same advantages with NOMA like spectrum efficiency, low-latency, and massive connectivity.
- NOMA-RA has been endorsed as the core technology for the emerging networks such as 6G [35]. This is mainly due to the less complexity added when the system is upgraded from the conventional random access technique to the NOMA-RA.
- The transition from the coordinated to the non coordinated transmission makes the NOMA-RA, a multiple access technique of choice. This transition results into less access delay which makes the NOMA-RA suitable for the use case of the uRLLC [34].
- NOMA-RA has also attracted attention from the industrial perspective. For example, It has been investigated in detail for its application in the unmanned aerial vehicle (UAV) use case [36].

2.4 Short-Packet Communications

The use case of the uRLLC works under the stringent latency and reliability constraint. Among the core enablers of the uRLLC, short-packet communications is regarded as the most important and most challenging one [9] . Due to the employment of the short-packet communications, the achieved throughput is less

as compared to the usage of conventional packet. This is quite suitable for the emerging IoT applications that require short data payload for their operation [3]. Also, when the short-packet communication is used, the achievable rate no longer is based on the Shannon formula. Then the question arises, how reliable is the usage of the short-packet communication. The reliability of the short-packet communications can be estimated by finding the relation between the achievable rate and the transmission error probability such as [5]

$$r = \log_2(1+\gamma) - \sqrt{\frac{\delta}{n}} Q^{-1}(\epsilon), \tag{2.1}$$

where γ is the received SNR, δ is the channel dispersion, ϵ is the transmission error probability, n is the blocklength, and Q^{-1} is inverse of Gaussian-Q function. The above approximation is true in case of single-input and single-output (SISO), single-input-multiple-output (SIMO), and multiple-input and single-output (MISO) settings. The first part in the above equation is the Shannon formula and the other part is the penalty which is introduced due to the usage of the short-packet communications. This penalty is the core reliability constraint, and to meet this reliability constraint it is always expected that the used channel codes for the short-packet communications are also strong enough [37].

Due to the reliability constraint and blocklength in short-packet communications based system, the complexity of the systems sometimes increased. It becomes more difficult to find the closed-form expression due to the above mentioned constraints. It is also very challenging to design the optimal resource allocation scheme that can consider the two factors such as power and blocklength optimization [38].

2.5 An overview of Effective Capacity and Effective bandwidth

Advances in wireless communications have resulted into the emergence of a wide range of applications. Emerging wireless networks with advanced technologies such as full-duplex (FD) communications, non-orthogonal multiple access (NOMA), multiple input and multiple output (MIMO) antennas, and millimeter wave promise

higher data rates [39]. With provision of this higher data rate and seamless connectivity, multimedia applications, which are regarded as delay-sensitive applications, have gained a lot of attention [40]. This requires an efficient modeling of wireless channel that can take into consideration quality-of-service (QoS) metrics such as delay-violation probability, data rate, and end-to-end delay [41].

Packet switched networks can be analysed with the help of physical and link-layer channel models depicted in Figure 2.2. Using physical-layer channel models for analysing the performance of delay-limited applications can be complex and inaccurate in some cases [21]. Hence, a new link-layer channel model named as "effective capacity (EC)" has been introduced [21]. With the help of EC, the channel can be modeled in terms of link-layer related QoS-metrics, such as probability of having non-empty buffer and delay violation probability. Concept of this link-layer channel model was first introduced in [21], which modeled a wireless link using two EC functions named as QoS exponent and probability of non-empty buffer. The developed link-layer channel model provides advantages such as ease of implementation and translation into the QoS guarantee, i.e., delay violation probability. Main motivations involving EC metric for various performance evaluations are highlighted below:

- EC modelling is based on an in-depth queueing analysis which can be used to characterize a relation between the source rate and the service rate taking into consideration both link-layer and physical layer parameters. Through this characterization, advance validation of communications system performance such as efficient admission control can be achieved [42].

- EC is the dual concept of effective bandwidth [43, 44] and shows the maximum constant arrival rate for a wireless channel while satisfying a delay outage probability constraint [45]. This feature can be exploited to achieve the required QoS for some applications with specific QoS requirements.

- With the help of the EC concept, QoS provisioning over wireless links and efficient bandwidth allocation can be achieved in closed-form while satisfying certain delay guarantee constraints [21].

- The EC performance of well-known physical layer-based resource allocation

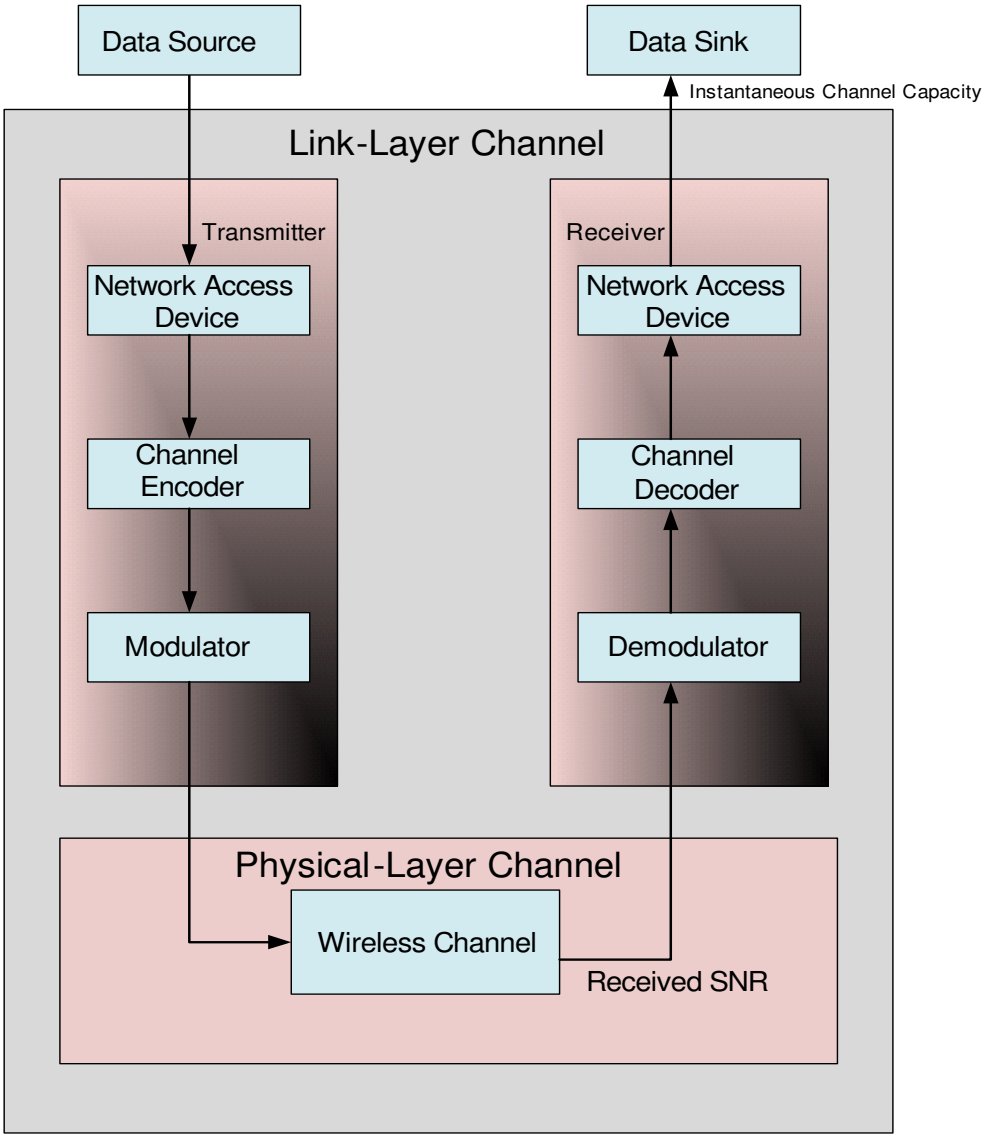

Figure 2.2: Basic components involved in the communications of packet switched networks. In this packet-based communications system, different components of physical and link-layer have been illustrated which shows the difference between physical and link-layer channel modelling.

algorithms, e.g, water filling, can be investigated. Performance of various proposed adaptive modulation and coding (AMC) schemes can be tested by using the EC metric [46].

- Using the EC model, the performance of adaptive resource allocation techniques for a specific QoS-aware connection can be analysed in closed-form in various cases. This, in turn, will pave the way for designing efficient resource allocation algorithms, hence optimizing the system performance.

- Provision of QoS guarantee with support for a variety of traffic flows requires efficient scheduling techniques. Using the EC concept, efficient delay constrained scheduling approaches can be investigated and designed [47].

2.5.1 Effective Bandwidth

The concept of effective-bandwidth is derived through the large-deviation principle and can show the minimum constant service rate that is needed to satisfy a given queueing delay requirement for a given source rate [9]. Effective bandwidth has been used extensively for obtaining optimal resource allocation schemes. Since effective bandwidth is based on large-deviation principle, it is traditionally used in systems with large delay bounds. However, effective bandwidth has also been recently used in scenarios where delay bound is short. For example, in [48], the concept of effective bandwidth is used to design an adaptive resource allocation scheme for a system with ultra low latency requirements. Below is the description of the effective-bandwidth approximation:

Consider a first-in-first-out (FIFO) queue with packets arrival rate at t as $\mu(t)$, the number of packets in the queue as $q(t)$, capacity of the link at time t as $c(t)$. Consider $q(t)$ to converge to a steady state $q(\infty)$ and define $A(t_1, t_2) = \sum_{t=t_1+1}^{t_2} \mu(t)$ as the total number of arrivals at $(t_1, t_2]$ and $C(t_1, t_2) = \sum_{t=t_1+1}^{t_2} c(t)$.

Authors in [49, 50] have proposed a theorem to derive the theory of effective bandwidth. For this purpose, the following assumptions are used as presented in [49, 50].

Let assume

- Arrival rate $\mu(t)$ and the service rate $c(t)$ are both ergodic and stationary. Also $\mathbb{E}[\mu(t)] < \mathbb{E}[c(t)]$, where $\mathbb{E}[.]$ shows the expectation operator.
- Arrival rate and source rate ($\mu(t)$ and $c(t)$) are independent.
- Using the Gartner-Ellis theorem, for all $\theta \in \mathbb{R}$, $\Lambda_A(\theta) \triangleq \lim_{t\to\infty} \frac{1}{t}\log(\mathbb{E}[e^{\theta A(0,t)}])$ and $\Lambda_C(\theta) \triangleq \lim_{t\to\infty} \frac{1}{t}\log(\mathbb{E}[e^{\theta C(0,t)}])$, where θ is the delay exponent, and $\Lambda_A(\theta)$ and $\Lambda_C(\theta)$ are assumed differentiable.

Now if there exists a unique $\theta^* > 0$, such that the below equation holds between the effective bandwidth and EC:

$$\Lambda_A(\theta^*) + \Lambda_C(-\theta^*) = 0, \tag{2.2}$$

then, mathematical derivations are provided in [49, 50] that relates the value of θ^* (found from (2.2)) to the buffer overflow probability according to

$$\lim_{x\to\infty} \frac{\log(\Pr\{q(\infty) \geq x\})}{x} = -\theta^*, \tag{2.3}$$

where $\Pr\{a \geq b\}$ shows the probability of a being greater than or equal to b.
Proof. For proof please refer to [49, 50].
Let x be the buffer size of the link. Packets are usually dropped when buffer becomes full. From (2.3) packet loss probability ϵ can be approximated as [50]

$$\epsilon = e^{-\theta^* x}. \tag{2.4}$$

Assuming that the link has constant capacity such that $c(t) = c$, for all t, $\Lambda_C(-\theta^*)$ can be simplified as

$$\Lambda_C(-\theta^*) = \lim_{t\to\infty} \frac{1}{t}\log\left(e^{-\theta^* ct}\right) = -\theta^* c. \tag{2.5}$$

Using (2.2), one can get $\frac{\Lambda_A(\theta^*)}{\theta^*} = c$. To have a small packet loss probability, a capacity of the link equal to $\frac{\Lambda_A(\theta^*)}{\theta^*}$ is required where the value for θ^* comes from the unique solution of $\theta^* = -(\log\epsilon)/x$ (derived from 2.4).

2.5.2 Effective Capacity

Authors in [21] introduced the concept of EC by taking motivations from the theory of effective bandwidth. EC is the dual concept of effective bandwidth. Recall that effective bandwidth shows the minimum service rate that is needed to guarantee a delay requirement for a given source traffic. The EC model, on the other hand, can be used to find the maximum source rate that the channel can handle (service rate) with the required delay constraint. As has been discussed above, the concept of EC can be used when a delay bound is large. However, it can also be used to test the performance of a system when delay bound is small, as has been discussed in [48]. Analytical framework for deriving EC has been discussed briefly below:

The service process is assumed as $c(t), t = 0, 1, 2, ..,$ with a partial sum $C(t1, t2) = \sum_{t=t_1+1}^{t_2} c(t)$ is ergodic and stationary. Further, the Gartner-Ellis limits for this service process is expressed as

$$\Lambda_C(\theta) \triangleq \lim_{t\to\infty} \frac{1}{t} \log(\mathbb{E}[e^{\theta C(0,t)}]). \tag{2.6}$$

From (2.2), it becomes

$$E_c(\theta^*) = -\frac{\Lambda_C(-\theta^*)}{\theta^*} = \mu. \tag{2.7}$$

Recall that (2.7) is EC of service process, while θ^* is the QoS exponent. The delay outage probability can now be formulated as

$$\lim_{x\to\infty} \frac{\log(\Pr\{q(\infty) \geq x\})}{x} = -\theta^*. \tag{2.8}$$

A more stringent QoS requirements can be represented by a larger value of θ^* with a faster decay rate. However, smaller values of θ^* represent slower decay rates and provide looser QoS guarantees.

Now, an expression for the delay $(D(t))$ experienced by a packet at any time t can also be approximated as follows

$$\Pr\{D(t) > D_{\max}\} \approx \Pr\{q(\infty) > 0\}\, e^{-\theta^* \mu D_{\max}}, \tag{2.9}$$

where $\Pr\{q(\infty) > 0\}$ is the probability of non-empty buffer and $D_{\max}$ is the maximum delay bound. EC is the combination of two functions, namely, QoS exponent and probability of non-empty buffer.

The probability of non-empty buffer can be achieved by considering the

$$\Pr\{q(\infty) > 0\} \approx \frac{\mathbb{E}[\mu(t)]}{\mathbb{E}[c(t)]}. \tag{2.10}$$

The above analytical explanation of effective-bandwidth and EC can be summarized as follows:

- The value of EC at θ^*, $E_c(\theta^*)$, shows the maximum constant arrival rate. Hence, $\mu \leq E_c(\theta^*))$ should hold.
- The solution for θ^* can be found when $E_b(\theta^*) = E_c(\theta^*)$ (for the arrival and source processes) holds.
- Using (2.9), the delay-violation probability can be estimated by using the pre-determined value of delay bound, probability of non-empty buffer, and obtained value of θ^*.
- Using (2.10), the probability of non-empty buffer can be estimated.

2.6 Effective Capacity analysis with different Fading models

In this section, a survey of existing work with their achievable EC under different fading models used in various wireless networks is provided. It is noted that channel variability can cause long delays in the transmission buffers. Hence, indicating the importance of using a suitable mathematical framework for testing the performance of the networks. The multipath propagation in wireless signal can well be described by the wireless channel models namely, Rayleigh, Nakagami-m, and Rician models. These models provide the characterization of the probability distribution function of the received signal power. Rayleigh model considers the random non line of sight paths while Rician assumes atleast one line of sight signal. As compared to the Rayleigh and Rician, Nakagami is the generalized fading

model that with the help of parameter m encompasses multiple fading patterns. In this thesis, most of the work has been done using the Rayleigh fading model. Rayleigh fading model provides the good picture of the fading scenario when there are multiple non line of sights signals are considered. It is close to the real-world or practical scenario. However, the simulations have also been verified using the generalized fading model, i.e, Nakagami-m fading model. The EC model can indeed be used in designing the adaptive resource allocation and scheduling schemes [51] that are specifically suitable for applications with constraints on the buffer size. The main advantages of utilizing the EC model with different fading models are the provision of a general mathematical framework and simplification of QoS-aware metrics.

2.6.1 Stochastic Fading Models

Stochastic fading models cover the fading in a channel that results from scattering and multipath propagation. In these models, a random variable is added to show the additional fading. Recall that EC provides a generalized link-layer mathematical framework for testing the performance of the channel under delay constraints. On that basis, different fading models can be analyzed with their distinct characters. Existing work on the EC model mostly takes into consideration the stochastic fading models for an assessment of QoS-awareness in wireless networks. Among the stochastic fading models, Rayleigh and Nakagami-m fading channels have been used extensively with EC concept. Current work in wireless networks considers different versions of stochastic fading models including Rayleigh, Nakagami-m, Rician, log-normal, and Weibull fading channels with EC metric. Below is the description of each fading channel:

2.6.1.1 Rayleigh Fading Channels

Most of the existing work on EC in wireless communications considers Rayleigh fading channels. Rayleigh fading is more prominent when there is no line of sight communications between the transmitter and receiver. Following works [52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70] consider Rayleigh fading with the EC model in different wireless networks. More prominent wireless

networks that have been investigated with Rayleigh fading channels with the EC model, are CRNs, cellular networks, and cooperative networks including the FD-relay networks. In cellular networks, with statistical QoS provisioning, Rayleigh fading has been extensively evaluated with EC metric. In CRNs with multiple channels, prediction related to multiple interference has also been studied with Rayleigh fading channels. Achievable EC in CRNs with multiple channels and Rayleigh fading as the physical channel model has been extensively investigated to find the maximum arrival/source rate with the required delay-outage probability.

Most of the delay-sensitive applications with Rayleigh fading in wireless networks have also been evaluated with EC metric. Rayleigh fading has been used extensively because it helps the researchers to understand the radio signals in heavily urban environment. Closed-form expression of achievable EC with Rayleigh-fading is less complex as compared to the Nakagami-m fading channel. Therefore, maximization in EC of different wireless networks with Rayleigh-fading has been investigated extensively in the existing works. Another fading channel, that has been used extensively after Rayleigh fading is Nakagami-m fading channel.

2.6.1.2 Nakagami-m

For EC-based delay performance estimation of wireless networks, where the large delay-time spreads are going to be estimated, Nakagami-m fading channels are used by clustering different reflected ways. Authors in [71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89], have considered Nakagami-m fading distributions in different wireless networks using the EC model. Nakagami-m channel model is often regarded as the general fading channel and can be used to investigate the mobile and indoor-mobile scenarios. Depending upon the parameter m, where $m \in \left\{\frac{1}{2}, +\infty\right\}$, different fading conditions can be achieved. For example, when $m = \frac{1}{2}$, this represents the severe fading case, while $m = 1$ is the Rayleigh-fading, $m > 1$ approximates the Rician channel, and $m = \infty$ represents additive white Gaussian noise (AWGN).

The main advantage of using Nakagami-m fading distribution with the EC model in wireless networks is better matching of its empirical data as compared to other distributions such as Rayleigh and Rician. Authors in [90], have investigated

the EC model with Nakagami-m fading channel. This study reveals that uncorrelated Nakagami-m flat fading conditions can well be analyzed with EC-based QoS-aware model. Complementary cumulative distribution function (CCDF) of delay has also been approximated by the EC model in this work. This analytical approximation based on EC leads to understanding the delay statistical behavior, which is not possible with the physical layer channel models. Works of CRNs [76, 81, 91] and relay networks [78, 90] show that queueing behavior can well be evaluated with EC metric under Nakagami-m fading conditions.

2.6.1.3 Rician

As compared to Rayleigh fading channel, in Rician fading, out of several different paths one must be the line of sight path. In this fading conditions, amplitude of the propagated signals are modelled by using Rician distributions. Achievable EC of wireless networks with Rician fading conditions has been discussed in [92, 93, 94, 95]. As Rician fading conditions consider one strong component, this strong component can be the line of sight path, therefore Rician fading can be employed in some advance networks such as satellite communications [96] which studies the EC model of channel. Note that the satellite communications suffer from long delays in signal transmission due to the very long distance between the satellite and the users. Hence, the concept of EC can be very useful for analyzing the performance of these communications systems. In addition to satellite communications, the EC model with Rician fading has also been studied in cellular communications, indoor networks, and vehicular networks.

2.7 Summary

This chapter provided an depth theoretical discussion of the core enablers of uRLLC and mathematical tools to investigate it. The core enablers such as NOMA, NOMA-RA, and short-packet communications was discussed with their potential advantages and limitation. The latency and reliability requirements with their numerical values for the uRLLC was also discussed with their potential kinds. Research challenges faced by the uRLLC was also highlighted in this

chapter. The basic operation of the NOMA was discussed from the theoretical perspective and its advantages such as massive connectivity, low latency and higher throughput was also discussed in detail. Some challenges faced by NOMA transmission such as imperfect SIC and proper SIC was also highlighted. NOMA-RA with its basic operation and advantages was also introduced in detail. Finally, the link-layer channel model EC and EB was discussed in more detail with their mathematical derivations.

Chapter 3

Effective Capacity of NOMA with Finite Blocklength for Low-Latency Communications

3.1 Introduction

Non-orthogonal multiple access (NOMA), cloud radio access networks, massive MIMO, and full-duplex, are key enabling technologies for 5G and beyond 5G (B5G) networks [97]. Major use cases in these networks require the provision of ultra reliable and low-latency communications (URLLC). NOMA in conjunction with finite blocklength, a.k.a. short-packet communications is considered as a key enabler for URLLC [3, 9]. In fact, NOMA has gained much attention in academia and industry due to its potential to achieve higher throughput, massive connectivity, low latency, and higher reliability in favorable circumstances, as compared to its orthogonal multiple access (OMA) counterpart [16, 18]. NOMA with finite blocklength follows the basic operation of traditional NOMA, with superposition coding (SC) at the transmitter, and successive interference cancellation (SIC) at the receiver. However, the conventional Shannon formula to approximate the attainable rate (with almost no errors) is not applicable when considering short packets in the communications [98].

To achieve latency as low as 1ms, and reliability as high as 99.999%, com-

munications in finite blocklength regime is very promising [37, 98, 99]. In the leading work [8], the achievable rate of finite blocklength communication link constrained by a given error probability was investigated in additive white Gaussian noise (AWGN) channels. Therein, the blocklength was taken as small as 100 bits, and it was shown that the maximal achievable rate with finite blocklength could not be approximated with the Shannon formula. In the same work, a penalty factor, which is function of the channel dispersion and error probability, was introduced to obtain the achievable rate in finite blocklength regime. The study work was further extended for the case of Rayleigh block fading channels in [100], wherein a trade-off between reliability, latency, and throughput in finite blocklength regime was investigated to establish the importance of short-packet communications for low latency. Furthermore, upper and lower bounds on the received signal-to-noise ratio (SNR) while considering finite blocklength for a given error probability were obtained in [100].

To analyze the suitability of short-packet communications for low latency, the effective capacity (EC) framework was used in [9, 101]. EC is the dual concept of effective bandwidth, and is used to find the maximum arrival rate for a given service rate while satisfying a certain delay constraint [21, 102]. The performance of short-packet communications to achieve URLLC was also investigated in [103, 104] using the EC concept. For example, in [103], the performance of point-to-point communications under latency constraint and considering finite blocklength transmission was investigated while. In that work, three transmission strategies (fixed-rate, variable-rate, and variable-power) were studied with focus on short-packet communications. Later, the closed-form expression for the achievable EC with short-packet communications over Rayleigh fading channels for machine-type communications was found in [104]. The latter work solely considered the ultra-reliable communications (URC) use case, but did not consider URLLC. In [101], the performance of short-packet communications for achieving low latency was investigated with the EC concept.

As the combination of NOMA with finite blocklength is considered as an enabling technology for low-latency communications, several investigations have been done so far. In this regard, the performance of NOMA with finite blocklength was investigated in [13], which showed the amount of physical-layer trans-

mission latency that NOMA with finite blocklength can reduce under reliability constraint as compared to OMA. In [13], a closed-form expression for the block error rate of a two-user NOMA was derived and validated with simulations. Authors in [105] considered NOMA with short-packet communications, and looked into the trade-off between the decoding error probability, the transmission rate, and the blocklength. More specifically, the challenges associated with the SIC and transmission rate while using finite blocklength were highlighted in [105]. The latency performance of NOMA in finite blocklength regime as compared to its OMA counterpart was investigated in [106]. The work showed the improved performance of NOMA in terms of throughput and reducing latency as compared to OMA with finite blocklength. Another work on the comparative analysis of NOMA and OMA in short blocklength regime under reliability and latency constraints was done in [107]. The latter work was focused on energy-efficient transmission with NOMA, and showed improved performance as compared to OMA. On the other hand, a detailed statistical delay analysis of NOMA using EC was conducted in [108], including closed-form expressions for the achievable EC of a two-user NOMA system when the users are chosen from a set of V users. The work in [108], however, did not consider the latency performance of NOMA with short-packet communications. NOMA with short-packet communications was investigated with the concept of effective bandwidth in [109], where the required SNR for a given delay exponent and transmission error probability was obtained. Link-layer rate of two-user NOMA and OMA in finite blocklength was investigated in [110], which showed the improved performance of NOMA as compared to OMA when the delay exponent is loose. However, detailed delay analysis of NOMA in finite blocklength, i.e., multiple NOMA users, and the impact of the delay exponent on the queueing delay violation probability, are yet to be investigated.

In this chapter, achievable EC of a two-user NOMA network is derived, when the users are chosen from a set of V users in the cell, with finite blocklength regime to investigate the low-latency communications. The impact of a given transmission error probability, delay exponent, and transmit SNR, on the achievable EC with finite blocklength is investigated in detail. The major contributions of this chapter can be summarized as follows:

- The achievable EC of downlink two-user (out of V users) NOMA, and a multi-user NOMA network, with finite blocklength, are derived.
- Closed-form expressions for the achievable total EC of two-user NOMA subset, as well as the achievable individual EC of each user, in finite blocklength regime, are derived (cf. Section 3.4), and also validated using Monte-Carlo simulations (cf. Section 3.6).
- Realizing the complexity of the proposed closed-form expressions for the two-user NOMA, a simplified closed-form expressions is also derived to approximate the EC of the two-user NOMA network at high transmit SNRs.
- The total EC of multiple NOMA pairs in finite blocklength regime is also investigated by taking into consideration the different pairing sets of multiple users. These findings show that a NOMA set with users having more distinct channel conditions achieve a higher total EC as compared to one where users have less distinct channel conditions.
- While considering short-packet communications, i.e., communications in finite blocklength regime, the impact of different power coefficients, transmit SNR, delay exponent, transmission error probability, and queuing delay violation probability, on the achievable EC of NOMA networks are investigated. The impact of practical path-loss model on the achievable EC of two-user NOMA network is further investigated. In particular, it is shown that when the delay exponent becomes stringent, the queueing delay violation probability cannot be reduced below a certain value due to the dominant effect of the transmission error probability.
- To provide a thorough performance evaluation, simulation results of two-user NOMA network with short-packet communications over generalized fading channels, i.e., with Nakagami-m model, are also investigated.

In detailing these contributions, the remainder of the chapter is organized as follows. First, communication model is discussed in Section 3.2. Concepts related to the theory of EC are presented in Section 3.3. Then, Section 3.4 provides the achievable EC of two-user NOMA with finite blocklength. Numerical results with

their insights are discussed in Section 3.6, and the chapter is concluded in Section 3.7.

3.2 System Model

In this work, power-domain downlink NOMA network with short-packet communications is considered. The network consists of one base station (BS) and V single-antenna users. The upper-layer packets of each user are assembled into frames, then stored at the transmit buffer of the BS, and later transmitted over the wireless channel as bit streams. It is assumed that each user is provided an individual buffer at the BS. Following the NOMA operation, the BS will send a broadcast signal, $\sum_{i=1}^{V} \sqrt{\alpha_i P} s_i(\tau)$, to the destination nodes, where α_i is the power allocation coefficient of user v_i, $s_i(\tau)$ is the message intended for v_i at time τ, and P is the total transmit power at the BS. The channels between the BS and the destination nodes are assumed to be block fading, i.e., the fading remains constant during each fading block, but changes independently from one fading block to another. Meanwhile, the blocklength is assumed to have the same size as that of the block fading and is taken as n. In this work, the channel gains follow Rayleigh distribution with unit variance. Users in this NOMA operation are classified based on their channel conditions. The channel coefficient between user v_i and the BS is referred to by $h_i(\tau)$. Without loss of generality, it is assumed that $|h_1(\tau)|^2 \leq |h_2(\tau)|^2 \leq ... \leq |h_V(\tau)|^2$. A practical path-loss model is adopted such that the channel between a user v_i and the BS is denoted by $L_i(\tau) = h_i(\tau) / d_i^{\frac{\alpha_L}{2}}$, where d_i is the distance between user v_i and the BS, and α_L is the path-loss exponent. Following the NOMA operation, the respective power coefficients are ordered as $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_V$.

In our system model, the users are ordered according to perfect instantaneous channel state information (CSI). The instantaneous CSI is available at the BS. BS selects the users based on the known CSI and performs the user ordering. The user ordering is random and BS randomly selects the users and paired them for the superposition coding at the transmitter. The BS knows the ordering of the users, based on which it allocates the power from the set of a fixed power coefficients. Fig. 3.1 shows the basic operation of the system model.

The BS broadcasts a message to users. The receive signal at user v_i can be formulated as:[1]

$$y_i = L_i \sum_{i=1}^{V} \sqrt{\alpha_i P} s_i + m_i, \tag{3.1}$$

where y_i is the received signal at user v_i, and m_i represents the AWGN.

In this network with V users, It is assumed that only two users (out of V) share the same resource block, using the NOMA operation. Refer to these users by v_u and v_t. When $u > t$, user v_u (the strong user) performs the SIC and detects the message of user v_t (the weak user). The so-called strong user will then remove the weak-user's message from its received message. In this case, the received SNR at user v_u can be formulated as

$$\mathrm{SNR}_u = \alpha_u \rho \left| L_u \right|^2, \tag{3.2}$$

where ρ is the transmit SNR, i.e., $\rho = \frac{P}{N_o B}$, with $N_o B$ the noise power. On the other hand, the message of user v_u at the weak-user's receiver will be considered as noise, therefore, user v_t will only decode its own message. The resulting SINR at the weak user is hence given by

$$\mathrm{SINR}_t = \frac{\alpha_t \rho \left| L_t \right|^2}{\alpha_u \rho \left| L_t \right|^2 + 1}. \tag{3.3}$$

In NOMA operation, the users are ordered according to their ordered channel gains. Using $\rho \left| h_i \right|^2 / d_i^{\frac{\alpha_L}{2}} = \gamma_i$, and denoting its probability density function (PDF) by $f(\gamma_i)$, the PDF of the ordered γ_i, $i = \{1, ., ., V\}$, can be obtained from ordered statistics [111, 112], and is given by

$$f_{(i:V)}(\gamma_i) = \xi_i f(\gamma_i) F(\gamma_i)^{i-1} (1 - F(\gamma_i))^{V-i}, \tag{3.4}$$

where $f_{(i:V)}$ is the PDF of ordered γ_i from a set of V users, $\xi_i = \frac{1}{B(i, V-i+1)}$, and $B(a, b)$ is the Beta function [113].

In this work, finite blocklength transmission is considered, hence the achievable rate cannot be represented by the Shannon formula, as proven in [8]. The

[1] Hereafter, the time index τ is removed for simplicity, whenever it is clear from the context.

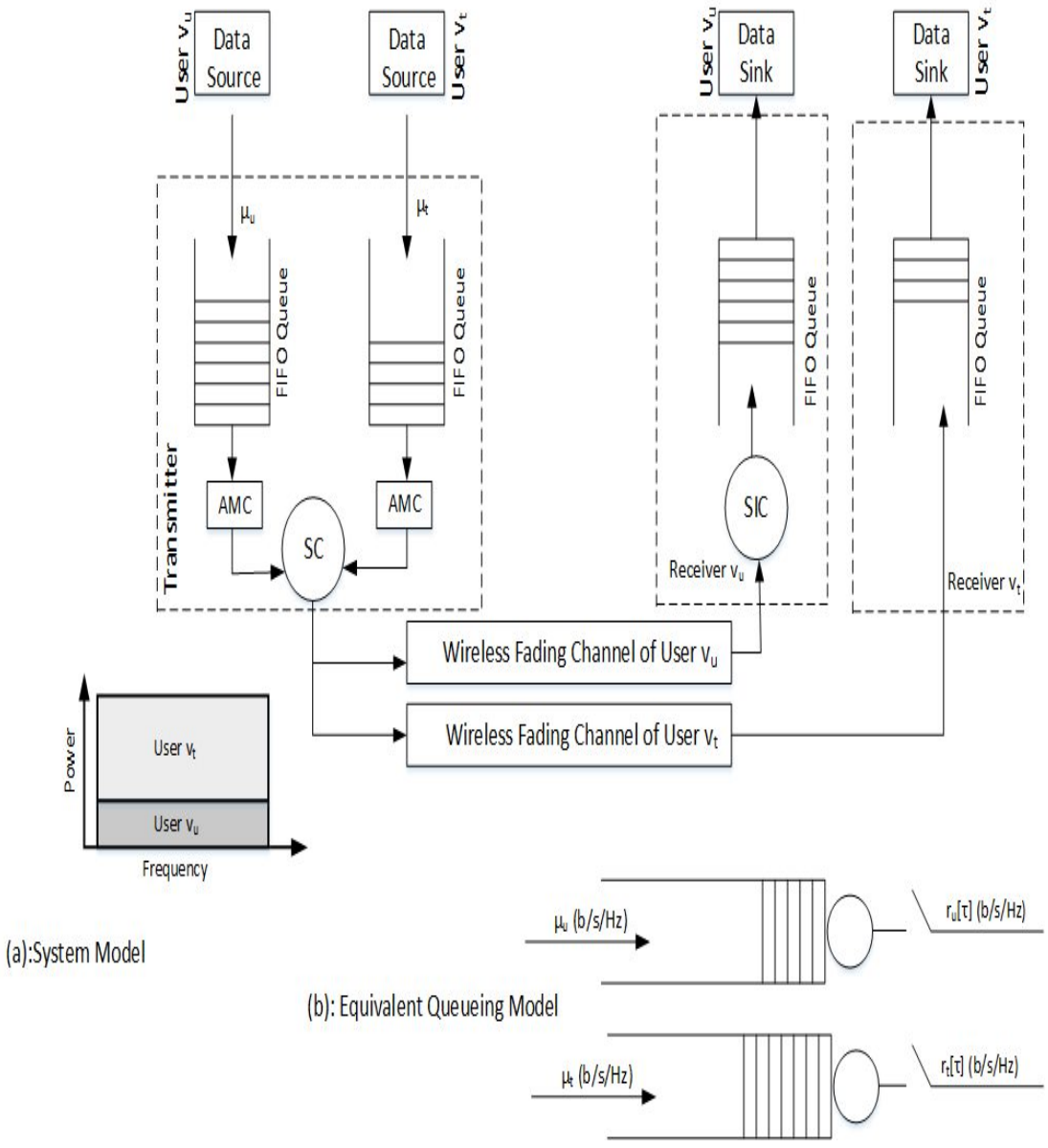

Figure 3.1: Two-user NOMA operation with finite blocklength with their respective queues: (a) describes the system model with two queues at the BS with their respective receivers, and (b) depicts the equivalent queueing model with the arrival rate and service rate.

results in [8] show that the achievable rate is a function of not only the received SNR (or SINR), but also of the transmission error probability (ϵ) and the transmission blocklength (n). Using [8], the achievable rate for user v_u and user v_t with finite blocklength can be approximated in bit/s/Hz as,

$$r_u = \log_2\left(1+\alpha_u\gamma_u\right) - \sqrt{\frac{\delta_u}{n}}Q^{-1}(\epsilon), \tag{3.5}$$

$$r_t = \log_2\left(1+\frac{\alpha_t\gamma_t}{\alpha_u\gamma_t+1}\right) - \sqrt{\frac{\delta_t}{n}}Q^{-1}(\epsilon), \tag{3.6}$$

where δ_u and δ_t are the channel dispersions for users v_u and v_t, respectively, which can be approximated as $\delta_u = \sqrt{1-\left(1+\alpha_u\gamma_u\right)^{-2}}$, $\delta_t = \sqrt{1-\left(1+\frac{\alpha_t\gamma_t}{\alpha_u\gamma_t+1}\right)^{-2}}$, $Q^{-1}(.)$ is the inverse of Gaussian Q-function with $Q\left(x\right) = \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}dw$, ϵ is the transmission error probability, and n is the blocklength.

3.3 Theory of Effective Capacity

In this section, the basic concept for the theory of EC is explained. EC is the dual concept of effective bandwidth and has been proposed in [21] to introduce the link-layer QoS metrics, such as queuing-delay violation probability and the probability of non-empty buffer. Assume infinite-size buffer at the BS for the messages of users v_i, $i = \{1,.,.,V\}$, and the link capacity (service process) as $r_i(\tau)$ at time τ. At time τ, the number of packets arriving and the number of packets in queue are represented by $a_i(\tau)$ and $q_i(\tau)$, respectively. Let the arrival rate and the link-layer capacity be ergodic and stationary processes and $E[a_i(\tau)] < E[r_i(\tau)]$, so $q_i(s)$ converges to a steady state denoted by $q_i(\infty)$ [114]. In practice, buffer overflow will occur if $q_i(\infty)$ exceeds the maximum length of the buffer. Assume x is a maximum threshold on $q_i(\infty)$, then using the large deviation theorem,

$$-\lim_{x\to\infty}\frac{\ln\left(\Pr\left\{q_i(\infty)>x\right\}\right)}{x} = \theta_i, \tag{3.7}$$

where θ_i is the delay exponent of user v_i, $q_i(\infty)$ is the steady state of the transmit buffer, and Pr{a>b} is the probability that $a>b$ holds. Now, using EC, the buffer overflow probability, given in (3.7), with a certain target θ_i can be satisfied if

$$\Lambda_{a_i}(\theta_i) + \Lambda_{r_i}(-\theta_i) = 0, \tag{3.8}$$

where $\Lambda_{a_i}(\theta_i) = \lim_{T\to\infty} \frac{1}{T}\log(\mathbb{E}[e^{\theta_i \sum_{\tau=1}^{T} a_i(\tau)}])$ is the Garntner-Ellis limits of the source process (arrival rate), and $\Lambda_{r_i}(\theta_i) = \lim_{T\to\infty} \frac{1}{T}\log(\mathbb{E}[e^{\theta_i \sum_{\tau=1}^{T} r_i(\tau)}])$ is the Garntner-Ellis limits of the service process [115]. Suppose that the source rate $a_i(\tau)$ is constant such that $a_i(\tau) = a_i$. From (3.8), one can get the maximum arrival rate (effective capacity) for some unique θ_i (delay QoS exponent), which is named EC and can be approximated by $-\frac{\Lambda_{r_i}(-\theta_i)}{\theta_i}$ [21]. From (3.7), the delay experienced by the source packets in buffer at time τ can also be estimated in terms of the queuing delay violation probability using [21]

$$\Pr\left\{D_i(\tau) > D^i_{\max}\right\} \approx \Pr\left\{q_i(\infty) > 0\right\} e^{-\theta_i \mu_i D^i_{\max}}. \tag{3.9}$$

The above expression is the queuing delay violation probability for user v_i, where $\Pr\{q_i(\infty) > 0\}$ represents the probability of non-empty buffer, and $D^i_{\max}$ is the maximum delay. It is important to note that $\mu_i = C^i_e$ is the effective capacity satisfying a certain QoS metric for user v_i [21]. The value for θ_i ($\theta_i > 0$) from (3.9) is the decay rate of the outage probability corresponding to user v_i. A more stringent delay requirements can be represented with a larger value of θ_i, while a smaller value of θ_i shows a less stringent delay requirement.

3.3.1 Effective Capacity in Finite Blocklength Regime

In this section, the major aim is to investigate the latency performance of a two-user (out of V users) NOMA with the short-packet communications using the EC framework. The traditional stochastic model for finding the achievable EC using the Shannon limit as the service rate is not suitable when considering finite blocklength transmissions. With short-packet communications in NOMA, use r_u and r_t as provided in (3.5) and (3.6) for the service rates. The stochastic model for the achievable EC with short-packet communication is provided in [103]. By

following the derived service rate from (3.5) and (3.6), the EC for the two-user NOMA with finite blocklength can be approximated as,

$$C_e^i = -\frac{1}{\theta_i n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\, e^{-\theta_i n r_i}\right]\right), \tag{3.10}$$

where C_e^i and θ_i are the EC and the QoS constraint for user v_i respectively, and $\mathbb{E}[.]$ is the expectation operator.

3.4 Effective Capacity of Downlink Two-User NOMA with Finite Blocklength

In this section, the achievable EC of a two-user NOMA and multi-user NOMA networks with short-packet communications is derived. Focusing on a two-user NOMA network, a closed-from expressions for the EC of the strong and weak users in finite blocklength regime is also provided.

3.4.1 Achievable Effective Capacity of Strong-User NOMA with Finite Blocklength

Out of the two users, i.e., the strong user (v_u) and the weak user (v_t), first the achievable EC of the strong user is provided and then provide its corresponding closed-form expression is provided. Following (4.14), the achievable EC of user v_u (C_e^u) is formulated as

$$C_e^u = -\frac{1}{\theta_u n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(1+\alpha_u\gamma_u\right)^{2\zeta_u} e^{\beta_u\delta_u}\right]\right), \tag{3.11}$$

where $\zeta_u = -\frac{\theta_u n}{2\ln 2}$ and $\beta_u = \theta\sqrt{n}Q^{-1}(\epsilon)$.
The above expression can be simplified by deriving its closed form. In this regard, after applying the order statistics from (3.4), the achievable EC of the strong user, given in (3.11), is expanded as

$$C_e^u = -\frac{1}{\theta_u n}\ln\left(\int_0^\infty \left(\epsilon + (1-\epsilon)\left(1+\alpha_u\gamma_u\right)^{2\zeta_u} e^{\beta_u\delta_u}\right) f_{(u:V)}\left(\gamma_u\right) d_{\gamma_u}\right). \tag{3.12}$$

This expression can further be simplified by expanding the order statistics from (3.4) and changing $e^{\beta_u \delta_u}$ into a fraction. In this vein, using the Maclaurin series for the expansion of $e^{\beta_u \delta_u}$ such that $e^{\beta_u \delta_u} \approx 1+\beta_u \delta_u + \frac{(\beta_u \delta_u)^2}{2}$, the achievable EC of the strong user is expressed as

$$C_e^u = -\frac{1}{\theta_u n} \ln\Bigg(\int_0^\infty \left(\epsilon + (1-\epsilon)\left(1+\alpha_u \gamma_u\right)^{2\zeta_u} \left(1 + \beta_u \delta_u + \frac{(\beta_u \delta_u)^2}{2} \right) \right) \times \xi_u f\left(\gamma_u\right) F\left(\gamma_u\right)^{u-1} \left(1 - F\left(\gamma_u\right)\right)^{V-u} d_{\gamma_u} \Bigg). \tag{3.13}$$

After inserting $\delta_u = \sqrt{1-\left(1+\alpha_u \gamma_u\right)^{-2}}$ in the above equation, the EC of the strong user becomes

$$\begin{aligned} C_e^u = -\frac{1}{\theta_u n} \ln\Bigg(\int_0^\infty \Bigg(& \epsilon + (1-\epsilon)\left(1+\alpha_u \gamma_u\right)^{2\zeta_u} + \beta_u \\ & \times \left(1+\alpha_u \gamma_u\right)^{2\zeta_u} \sqrt{1-\left(1+\alpha_u \gamma_u\right)^{-2}} + \frac{{\beta_u}^2}{2} \\ & \times \left(1+\alpha_u \gamma_u\right)^{2\zeta_u} \left(1-\left(1+\alpha_u \gamma_u\right)^{-2}\right) \Bigg) \xi_u f\left(\gamma_u\right) \\ & \times F\left(\gamma_u\right)^{u-1} \left(1-F\left(\gamma_u\right)\right)^{V-u} d_{\gamma_u} \Bigg), \end{aligned} \tag{3.14}$$

Then, using $\rho_l = \rho / {d_i}^{\frac{\alpha_L}{2}}$ and inserting $f\left(\gamma_u\right) = \frac{1}{\rho_l} e^{-\frac{\gamma_u}{\rho_l}}$ and $F\left(\gamma_u\right) = 1 - e^{-\frac{\gamma_u}{\rho_l}}$ into

(3.14), the achievable EC becomes

$$C_e^u = -\frac{1}{\theta_u n}\ln\Bigg(\frac{\xi_u}{\rho_l}\int_0^{\infty}\Bigg(\epsilon+(1-\epsilon)\left(1+\alpha_u\gamma_u\right)^{2\zeta_u}+\beta_u \times \left(1+\alpha_u\gamma_u\right)^{2\zeta_u}\sqrt{1-\left(1+\alpha_u\gamma_u\right)^{-2}}+\frac{{\beta_u}^2}{2} \times \left(1+\alpha_u\gamma_u\right)^{2\zeta_u}\left(1-\left(1+\alpha_u\gamma_u\right)^{-2}\right)\Bigg)e^{-\frac{(V-u+1)\gamma_u}{\rho_l}} \times \left(1-e^{-\frac{\gamma_u}{\rho_l}}\right)^{u-1} d_{\gamma_u}\Bigg). \tag{3.15}$$

After solving the above integral, the final closed-form expression for the achievable EC of the NOMA strong user in a finite blocklength can be approximated as,

$$C_e^u \approx -\frac{1}{\theta_u n}\ln\Bigg(\epsilon+(1-\epsilon)\Bigg(\frac{\xi_u}{\rho_l\alpha_u}\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i \times \mathrm{H}\Big(1,2+2\zeta_u,\eta_u\Big)\Big(\mathrm{K_u}+1\Big)-\mathrm{H}\Big(1,2\zeta_u,\eta_u\Big) \times\left(\mathrm{K_u}-\frac{\beta_u}{2}\right)\Bigg)\Bigg), \tag{3.16}$$

where $\eta_u = \frac{V-u+1+i}{\rho_l\alpha_u}$, $K_u = \frac{\beta_u^2}{2}+\beta_u$, and $\mathrm{H}(a,b,z)$ is the confluent hypergeometric function of the second kind defined by

$$\mathrm{H}\left(a,b,z\right) = \frac{1}{\Gamma\left(a\right)}\int_0^{\infty} e^{-zt}t^{a-1}\left(1+t\right)^{b-a-1}dt \quad \text{for Re [a], Re [z]} > 0, \tag{3.17}$$

where $\Gamma\left(.\right)$ is the Gamma function [116].

The details for deriving the closed-form expression for the EC of the strong user are provided in Appendix A. The accuracy of the proposed closed-form expression has also been verified using simulations, as will be shown later in Section 3.6.

3.4.2 Achievable Effective Capacity of Weak-User NOMA with Finite Blocklength

Following the steps for deriving the achievable EC of user v_u and its closed-form expression, the achievable EC of user v_t is formulated and its closed-form expression is derived. Using (4.14), the achievable EC of the weak user is given as

$$C_e^t = -\frac{1}{\theta_t n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} e^{\beta_t\delta_t}\right]\right), \tag{3.18}$$

where $\zeta_t = -\frac{\theta_t n}{2\ln 2}$ and $\beta_t = \theta\sqrt{n}Q^{-1}(\epsilon)$.

The above expression can be simplified by deriving its closed form. In this sense, applying the order statistics from (3.4) in (3.18), the EC of v_t becomes

$$C_e^t = -\frac{1}{\theta_t n}\ln\left(\int_0^\infty \left(\epsilon + (1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} e^{\beta_t\delta_t}\right) f_{(t:V)}(\gamma_t)\, d_{\gamma_t}\right). \tag{3.19}$$

Equation (3.19) can further be simplified by expanding the order statistics from (3.4) and using the Maclaurin series for the $e^{\beta_t\delta_t}$ expression such as $e^{\beta_t\delta_t} \approx 1 + \beta_t\delta_t + \frac{(\beta_t\delta_t)^2}{2}$. Accordingly, the achievable EC of user v_t is re-written as

$$\begin{aligned} C_e^t = -\frac{1}{\theta_t n}\ln\Bigg(\int_0^\infty &\left(\epsilon + (1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}\left(1+\beta_t\delta_t+\frac{(\beta_t\delta_t)^2}{2}\right)\right) \\ &\times \xi_t f(\gamma_t) F(\gamma_t)^{t-1}(1-F(\gamma_t))^{V-t}\, d_{\gamma_t}\Bigg), \end{aligned} \tag{3.20}$$

Using $\delta_t = \sqrt{1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}}$ in the above expression, the achievable EC of the

weak user can then be reformulated as

$$C_e^t = -\frac{1}{\theta_t n}\ln\left(\int_0^\infty\left(\epsilon+(1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}+\beta_t\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}\right.\right.$$
$$\times\sqrt{1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}}+\frac{\beta_t^2}{2}\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} \tag{3.21}$$
$$\left.\left.\times\left(1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}\right)\right)\xi_t f(\gamma_t)F(\gamma_t)^{t-1}(1-F(\gamma_t))^{V-t}\,d_{\gamma_t}\right).$$

Now, using $f(\gamma_t)=\frac{1}{\rho_l}e^{-\frac{\gamma_t}{\rho_l}}$ and $F(\gamma_t)=1-e^{-\frac{\gamma_t}{\rho_l}}$, the above expression for C_e^t is further simplified to

$$C_e^t = -\frac{1}{\theta_t n}\ln\left(\frac{\xi_t}{\rho_l}\int_0^\infty\left(\epsilon+(1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}+\beta_t\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}\right.\right.$$
$$\times\sqrt{1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}}+\frac{\beta_t^2}{2}\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} \tag{3.22}$$
$$\left.\left.\times\left(1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}\right)\right)e^{-\frac{(V-t+1)\gamma_t}{\rho_l}}\left(1-e^{-\frac{\gamma_t}{\rho_l}}\right)^{t-1}d_{\gamma_t}\right).$$

Finally, taking some further mathematical simplification in (3.22) and solving the integrals, the closed-form expression for the achievable EC of the NOMA weak user in a finite blocklength is obtained as

$$C_e^t \approx -\frac{1}{\theta_t n}\ln\Bigg(\epsilon+(1-\epsilon)\Bigg(\Bigg(\frac{\alpha_u^{-2\zeta_t}\xi_t}{\rho_l}\Bigg(\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r\frac{1}{\eta_t\alpha_u}+\frac{\theta_t n(\alpha_u-1)}{\alpha_u \ln 2}$$
$$\times\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r e^{\eta_t}\mathrm{E_i}(-\eta_t)+\sum_{s=2}^{\infty}\binom{2\zeta_t}{s}\Big(\frac{\alpha_u-1}{\alpha_u}\Big)^s\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r$$
$$\Bigg(\frac{\sum_{r=1}^{s-1}\frac{(r-1)!}{\alpha_u^{-r}}(-\alpha_u\eta_t)^{s-r-1}}{(s-1)!}-\frac{(-\alpha_u\eta_t)^{s-1}}{(s-1)!}e^{\eta_t}\mathrm{E_i}(-\eta_t)\Bigg)\Bigg)\Bigg)(K_t+1)$$
$$-\Bigg(\Bigg(\frac{\alpha_u^{-(2\zeta_t-2)}\xi_t}{\rho_l}\Bigg(\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r\frac{1}{\eta_t\alpha_u}+\frac{\theta_t n(\alpha_u-1)}{\alpha_u\ln 2}\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r e^{\eta_t}$$
$$\times\mathrm{E_i}(-\eta_t)+\sum_{s=2}^{\infty}\binom{2\zeta_t-2}{s}\Big(\frac{\alpha_u-1}{\alpha_u}\Big)^s\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r$$
$$\times\Bigg(\frac{\sum_{r=1}^{s-1}\frac{(r-1)!}{\alpha_u^{-r}}(-\alpha_u\eta_t)^{s-r-1}}{(s-1)!}-\frac{(-\alpha_u\eta_t)^{s-1}}{(s-1)!}e^{\eta_t}\mathrm{E_i}(-\eta_t)\Bigg)\Bigg)\Bigg)\Big(K_t-\frac{\beta_t}{2}\Big)\Bigg)\Bigg)\Bigg), \tag{3.23}$$

where $\eta_t=\frac{V-t+1+r}{\rho_l\alpha_u}$, $K_t=\frac{\beta_t^2}{2}+\beta_t$, and $\mathrm{E_i}(.)$ is the exponential integral with $E_i(x)=-\int_{-x}^{\infty}\frac{e^{-w}}{w}dw$. The details for deriving the close-form expression of the weak-user's EC is provided in Appendix B.

3.4.3 Achievable Effective Capacity of Multiple NOMA Pairs in Finite Blocklength

In this part, the total achievable EC of multiple NOMA pairs with finite is investigated. Specifically, it is considered that there are V users in total, which are divided into $\frac{V}{2}$ pairs, with $\mathbb{M}=\left\{1,2,...,\frac{V}{2}\right\}$ denoting the set of group indices. The parameter ϕ is introduced as the combination of all NOMA pairs such that $\phi=\left\{\phi_1,\phi_2,...,\phi_{V/2}\right\}$. By taking the m^{th} NOMA pair with finite blocklength, in which the strong user is denoted by v_{u_m} and the weak user is denoted by v_{t_m}, such that $\phi_m=\left\{(t_m,u_m)\mid t_m\neq u_m,\ |h_{t_m}|^2\leq|h_{u_m}|^2,\ \forall_m\in\mathbb{M}\right\}$.

Next, the m^{th} NOMA pair is considered, and its achievable EC for users v_{u_m} and v_{t_m} in the pair are investigated. It is to be noted that the inter-pair multiple

access is based on frequency-division multiple access (FDMA). Considering the m^{th} NOMA pair with communication in finite blocklength regime, the transmission rate for the strong and weak users can be approximated as

$$r_{u_m} = \frac{2}{V}\left(\log_2\left(1+\alpha_{u_m}\gamma_{u_m}\right) - \sqrt{\frac{\delta_{u_m}}{n}}Q^{-1}(\epsilon)\right), \tag{3.24}$$

$$r_{t_m} = \frac{2}{V}\left(\log_2\left(1+\frac{\gamma_{t_m}+1}{\alpha_{u_m}\gamma_{t_m}+1}\right) - \sqrt{\frac{\delta_{t_m}}{n}}Q^{-1}(\epsilon)\right), \tag{3.25}$$

where $\gamma_{i_m} = \rho\left|hi_m\right|^2 / d_{i_m}^{\frac{\alpha_L}{2}}$. Using (3.24) and (3.25) as the transmission rates with finite blocklength, and applying the Gartner-Ellis theorem, the achievable EC for the strong and weak users can respectively be formulated as

$$C_e^{u_m} = -\frac{1}{\theta_{u_m} n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(1+\alpha_{u_m}\gamma_{u_m}\right)^{\frac{4\zeta_{u_m}}{V}} e^{\frac{2}{V}\beta_{u_m}\delta_{u_m}}\right]\right), \tag{3.26}$$

$$C_e^{t_m} = -\frac{1}{\theta_{t_m} n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(1+\frac{\gamma_{t_m}+1}{\alpha_{u_m}\gamma_{t_m}+1}\right)^{\frac{4\zeta_{t_m}}{V}} e^{\frac{2}{V}\beta_{t_m}\delta_{t_m}}\right]\right). \tag{3.27}$$

The achievable EC of multiple NOMA pairs in (3.26) and (3.27), and the EC of two-user NOMA network shown in (3.11) and (3.18), have similar expressions. Therefore, by following Appendix A and Appendix B, the closed-form expressions for the achievable EC of v_{t_m} and v_{u_m} users in multiple NOMA pairs with finite blocklength can be derived. The total EC, T_{ec}, can be estimated by using $\sum_{m=1}^{V}(C_e^{t_m} + C_e^{u_m})$. Analytical results in Section 3.6 regarding the multiple NOMA pairing will be investigated in detail. The users with more distinct and less distinct channel conditions will be paired together and their T_{ec} will be analyzed with respect to ρ.

3.5 Effective Capacity of Downlink Two-User NOMA with Finite Blocklength at High Transmit SNRs

The performance of two-user NOMA in finite blocklength regime can be investigated by leveraging the closed-form expressions for the EC of the strong user and weak user presented in (3.16) and (3.23), respectively. The obtained closed-form expressions are somehow complex with insights being difficult to get from. Here, the achievable EC formulation is simplified by using the approximations for the two-user NOMA network and their corresponding closed-form expressions. In this regard, the channel dispersion (δ_i) is approximated as $\delta_i \approx 1$, for high SNRs [8]. When the transmit SNR is above the 20dB, the transmit SNR is taken as high. This high transmit SNR results into the channel dispersion as $\delta_i \approx 1$. Considering this approximation at high SNR, the achievable EC of the strong and weak users can now be simplified as

$$\bar{C}_e^{ru} = -\frac{1}{\theta_u n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(1+\alpha_u\gamma_u\right)^{2\zeta_u} e^{\beta_u}\right]\right), \tag{3.28}$$

$$\bar{C}_e^{t} = -\frac{1}{\theta_t n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} e^{\beta_t}\right]\right), \tag{3.29}$$

where $\bar{C}_e^{ru}$ and $\bar{C}_e^{t}$ are the achievable ECs of the NOMA strong user and weak user at high transmit SNR ($\delta_i = 1$), respectively. The above equations can be further simplified by deriving their closed-form expressions. Using the order statistics from (3.4), the achievable ECs of the strong and weak users at high transmit SNR can be expanded as

$$\bar{C}_e^{u} = -\frac{1}{\theta_u n}\ln\left(\int_0^\infty \left(\epsilon + (1-\epsilon)\left(1+\alpha_u\gamma_u\right)^{2\zeta_u} e^{\beta_u}\right) f_{(u:V)}\left(\gamma_u\right) d_{\gamma_u}\right), \tag{3.30}$$

$$\bar{C}_e^{t} = -\frac{1}{\theta_t n}\ln\left(\int_0^\infty \left(\epsilon + (1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} e^{\beta_t}\right) f_{(t:V)}\left(\gamma_t\right) d_{\gamma_t}\right). \tag{3.31}$$

The above integrals can be solved by using similar steps as those in Appendix A and Appendix B. After solving the above integrals, the final closed-form expres-

sions for the achievable ECs of the strong and weak users at high SNR can be approximated as follows:

$$\bar{C}_e^u \approx -\frac{1}{\theta_u n}\ln\left(\epsilon+(1-\epsilon)\left(\frac{\xi_u}{\rho_l\alpha_u}e^{\beta_u}\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i\mathrm{H}\left(1,2+2\zeta_u,\eta_u\right)\right)\right), \tag{3.32}$$

$$\begin{aligned}\bar{C}_e^t \approx -\frac{1}{\theta_t n}\ln\Bigg(&\epsilon+(1-\epsilon)\frac{\alpha_u^{-2\zeta_t}\xi_t}{\rho_l}e^{\beta_t}\Bigg(\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r\frac{1}{\eta_t\alpha_u}+\frac{\theta_t n(\alpha_u-1)}{\alpha_u\ln 2}\\ &\times\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r e^{\eta_t}\mathrm{E_i}(-\eta_t)+\sum_{s=2}^{\infty}\binom{2\zeta_t}{s}\left(\frac{\alpha_u-1}{\alpha_u}\right)^s\sum_{r=0}^{t-1}\binom{t-1}{r}(-1)^r\\ &\times\left(\frac{\sum_{r=1}^{s-1}\frac{(r-1)!}{\alpha_u^{-r}}(-\alpha_u\eta_t)^{s-r-1}}{(s-1)!}-\frac{(-\alpha_u\eta_t)^{s-1}}{(s-1)!}e^{\eta_t}\mathrm{E_i}(-\eta_t)\right)\Bigg)\Bigg).\end{aligned} \tag{3.33}$$

3.5.1 Effective Capacity of Downlink Two-User NOMA with Finite Blocklength at Extremely High Transmit SNR ($\rho \to \infty$)

The impact of the extremely high transmit SNR on the achievable EC of two-user NOMA network is also investigated. In this regard, the achievable EC of the NOMA strong and weak users at extremely high transmit SNR can be derived by inserting $\rho \to \infty$ in the EC formulation. Using the (3.11) and inserting $\rho \to \infty$, the EC of the strong user at extremely high SNR is expressed as

$$\begin{aligned}&\lim_{\rho\to\infty} -\frac{1}{\theta_u n}\ln\left(\mathbb{E}\left[\epsilon+(1-\epsilon)\,e^{-\theta_u n\left(\log_2(1+\alpha_u\gamma_u)-\sqrt{\frac{1-(1+\alpha_u\gamma_u)^{-2}}{n}}Q^{-1}(\epsilon)\right)}\right]\right)\\ &=-\frac{1}{\theta_u n}\log(\epsilon).\end{aligned} \tag{3.34}$$

Likewise, using the (3.18) and inserting $\rho \to \infty$, the EC of the NOMA weak user at extremely high transmit SNR is found as

$$\begin{aligned} &\lim_{\rho\to\infty} -\frac{1}{\theta_t n}\ln\left(\mathbb{E}\left[\epsilon+(1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} e^{\beta_t\sqrt{1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}}}\right]\right) \\ &= -\frac{1}{\theta_t n}\ln\left(\epsilon+(1-\epsilon)\,\alpha_u^{-2\zeta_t} e^{\beta_t\sqrt{1-\alpha_u^2}}\right). \end{aligned} \tag{3.35}$$

It is clear that when the transmit SNR ρ is extremely high, then the EC is limited (upper bounded) by a fixed value which is not a function of ρ. The achievable EC of the strong user is limited by the delay exponent, the transmission error penalty, and the blocklength. In the case of the weak user, it is clear from (3.35) that the achievable EC is limited by $-\frac{1}{\theta_t n}\ln\left(\epsilon+(1-\epsilon)\,\alpha_u^{-2\zeta_t} e^{\beta_t\sqrt{1-\alpha_u^2}}\right)$. When ρ is extremely high, the achievable EC of the weak user is limited by the transmission penalty due to short-packet communications, by the transmission error probability, and by the power co-efficient.

3.6 Numerical Results

Using numerical simulations, the performance of the proposed NOMA network with finite blocklength is now evaluated. The accuracy of the proposed closed-form expressions for the two-user NOMA under Rayleigh fading is also confirmed. Simulation results pertaining to the achievable EC of the strong and weak NOMA users when communicating over Nakagami-m fading channels are also discussed in detail.

In the simulation set-up, the total number of users is taken as $V = 10$, and the 2^{nd} and 8^{th} users are those paired together. These are users with the 2^{nd} and 8^{th} weakest channels, respectively, such that $t = 2$ and $u = 8$. The BS's power coefficients pertaining to these users are taken as $\alpha_t = 0.8$ and $\alpha_u = 0.2$, and blocklength is set $n = 400$, unless otherwise specified. A practical path-loss model with the path-loss exponent $\alpha_L = 2$, and $d_i = 10$ meters, is also adopted. In this work, V has a significant impact on the overall network performance. As the V increases it requires more transmission energy. However, NOMA has the inherent ability to adopt to the transmission policy, so increase in the users will result into

the increased spectrum efficiency (as more users are served) and more energy efficiency. However, sometimes, increased in the number of users may increase the interference, which could degrade the system performance. Therefore, it requires a fair balance when changing the V and adjusting the parameters.

The accuracy of the closed-form expressions for the EC of the NOMA users with finite blocklength are investigated in Fig. 3.2. The figure shows the plots of C_e^u (strong-user) and C_e^t (weak-user) in b/s/Hz versus the transmit SNR (ρ) in dB, where $\epsilon = 10^{-5}$, and the delay exponent $\theta = 0.01$. The results for these curves have been obtained using the proposed extended and simple closed-form expressions, i.e, (3.16), (3.32), (3.23), and (3.33), and Monte-Carlo simulations. Monte-Carlo simulation is the model that is used to determine the output of the uncertain events. These simulations are usually run for the multiple times to determine the probabilistic outcome of the uncertain events and then are matched for their correctness with the closed-form expression. A closed-form expression is the solved solution for a given mathematical problem. In this work, the Monte-carlo simulations of the derived achievable EC problem is verified using the closed-form expression of this problem. The match is perfect for this problem formulation. The accuracy of the closed-form expression for the strong and weak users can be confirmed. The very small mismatch between the results of the analysis and the simulations is due to the approximation $e^{\beta_i \delta_i} \approx 1 + \beta_i \delta_i + \frac{(\beta_i \delta_i)^2}{2}$ (for deriving the extended closed-form expression given in (3.16), and (3.23)), and using $\delta_i \approx 1$ at high SNR (for deriving the simple closed-form expressions shown in (3.32) and (3.33)). The achievable EC of the strong and weak users are upper bounded when the transmit SNR becomes high. However, the achievable ECs saturate at different values of ρ.

Figures 3.3 and 3.4 show the impact of the transmit SNR, ρ, on the achievable EC of the strong and weak NOMA users when considering the Nakagami-m fading model, while $\theta = 0.01$ and $n = 400$. It is clear from Fig. 3.3 that the strong-user's performance does not decrease significantly even under the worst fading conditions. The performance gap under different values for the Nakagami shaping parameter is wider at the central region of ρ, i.e., 25dB to 30dB. This performance gap is true for both, the strong user and the weak user. 3.4 shows the achievable EC of the NOMA weak-user for different fading conditions. It is clear that the

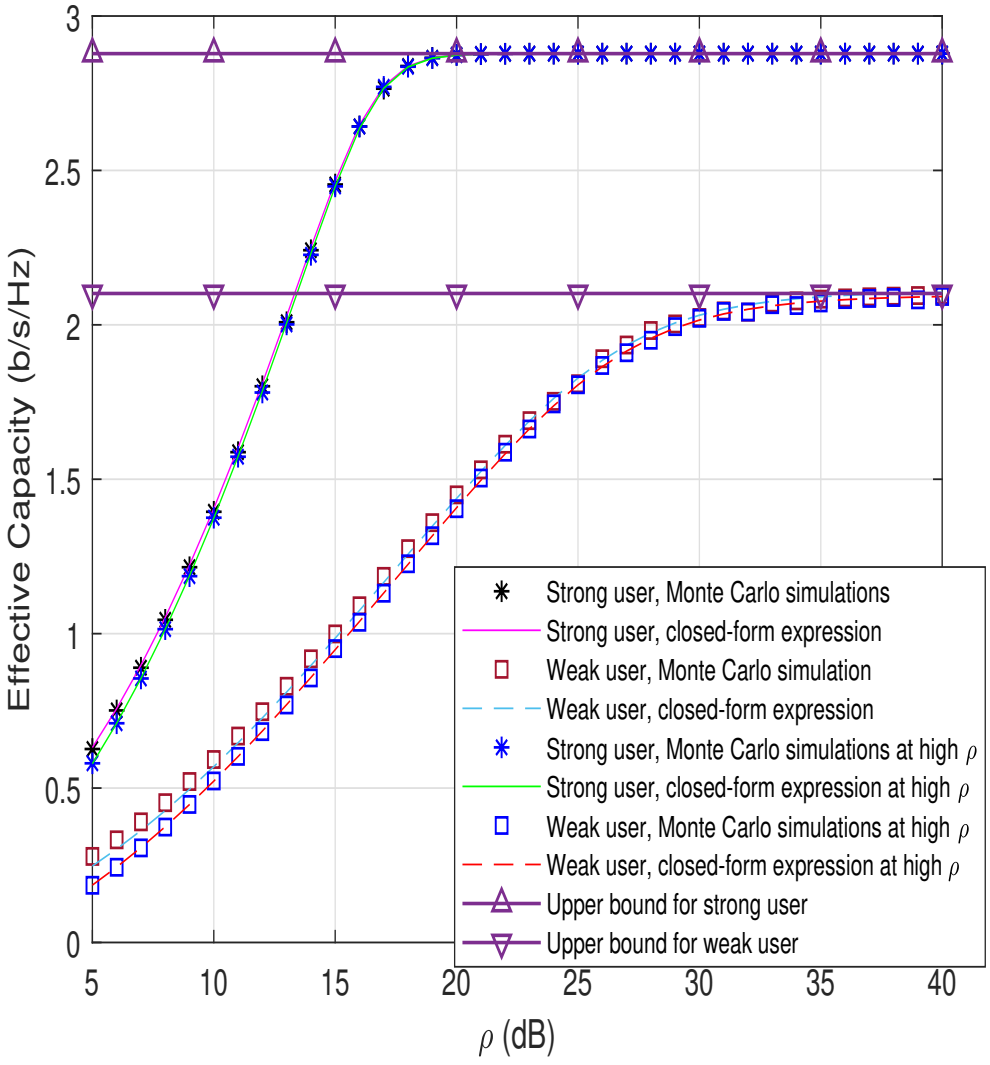

Figure 3.2: Effective Capacity of NOMA weak-user and strong-user versus transmit SNR, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$.

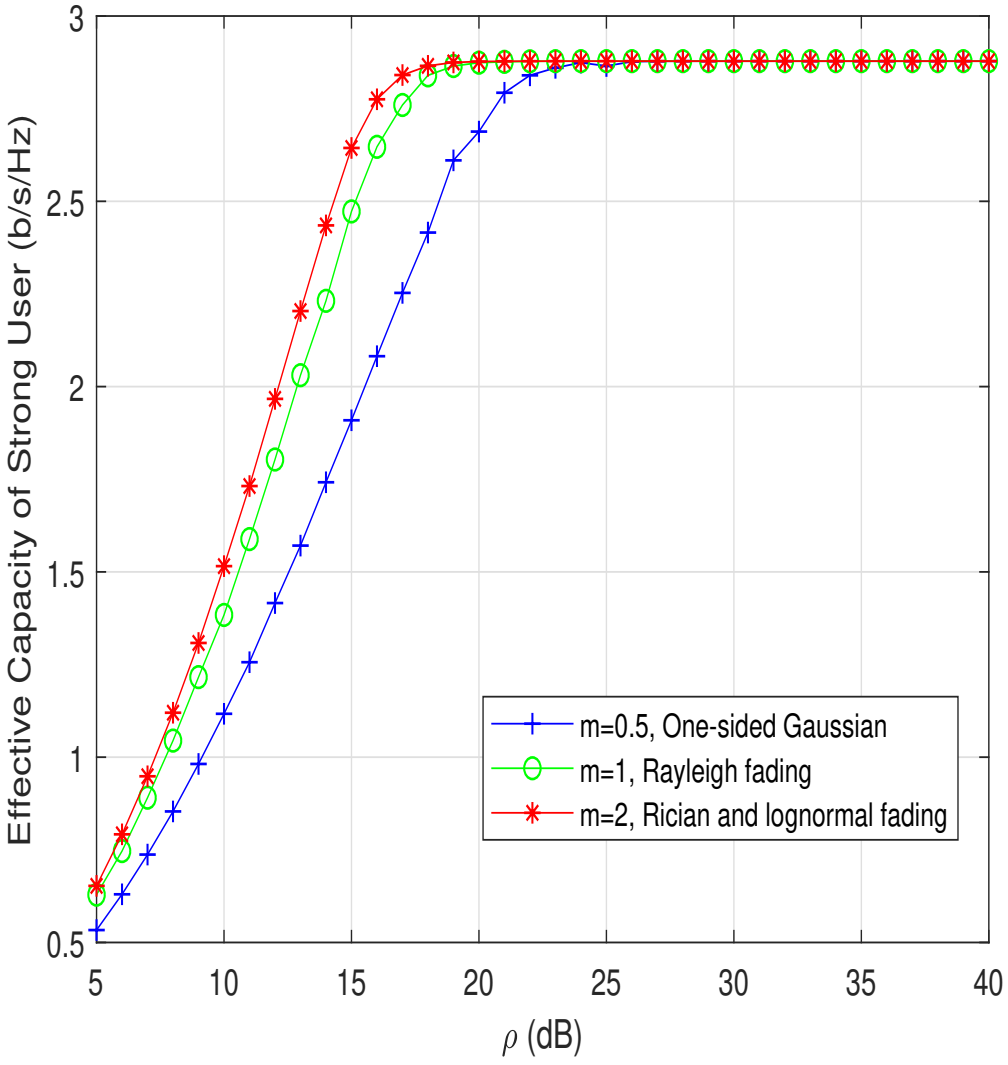

Figure 3.3: Effective Capacity of NOMA strong-user versus transmit SNR under Nakagami-m fading, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$.

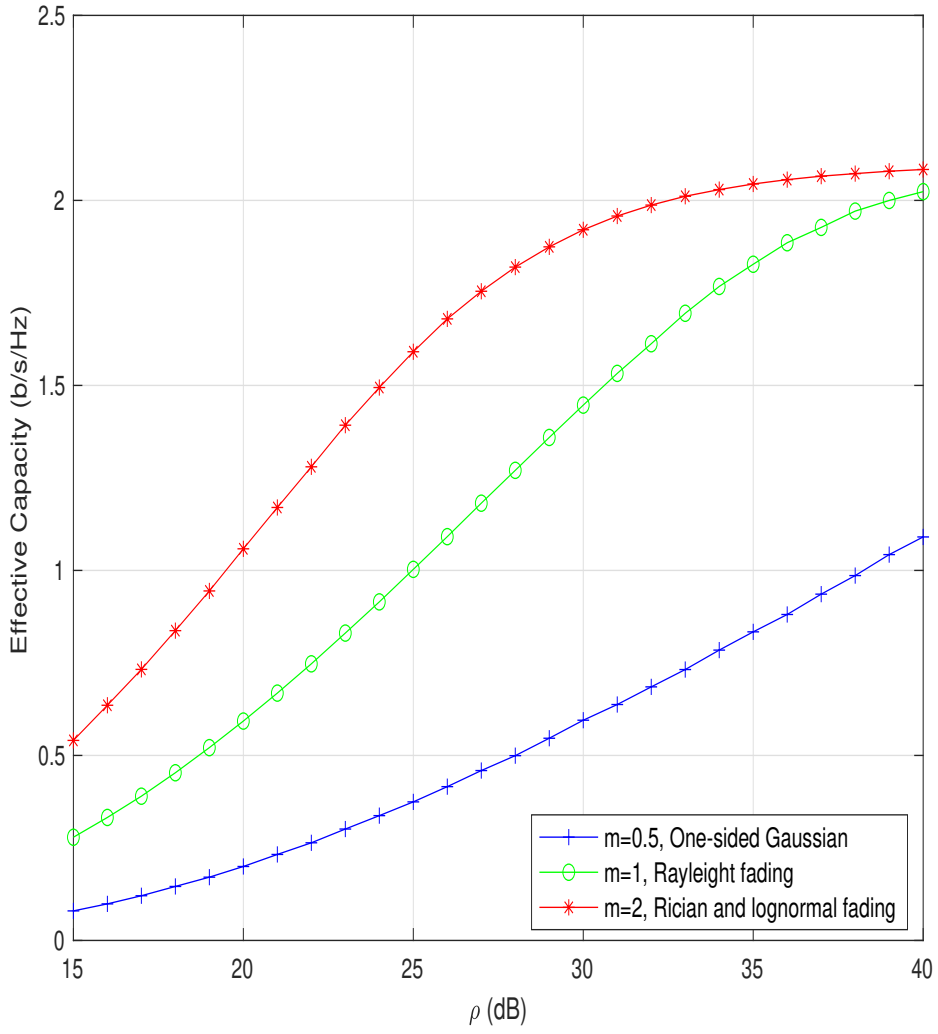

Figure 3.4: Effective Capacity of NOMA weak-user versus transmit SNR under Nakagami-m fading, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$.

performance of the weak user degrades significantly when the fading gets worse. Interestingly, the performance pertaining to each user saturates at a certain high SNR irrespective of the fading conditions. This saturation occurs earlier in the case of the strong user as compared to the weak user who requires even higher values of ρ to defy the worst fading conditions.

Figure 3.5 shows the simulation results for the achievable EC of a two-user NOMA network versus the transmit SNR, with different values of the power coefficient sets (α_u, α_t). Three sets of power coefficients, namely, $(\alpha_u = 0.2, \alpha_t = 0.8)$, $(\alpha_u = 0.3, \alpha_t = 0.7)$, and $(\alpha_u = 0.4, \alpha_t = 0.6)$, are used to better evaluate the performance of the proposed two-user NOMA system. The simulations reveal that increasing the power coefficients of the strong user increases its achievable EC, whereas increasing α_t for the weak user does not have a significant impact on this user's achievable EC.

Figure 3.6 illustrates the variation of T_{ec} versus the transmit SNR for different user pairing set ϕ of multiple NOMA pairs network with finite blocklength. The delay exponent for all users is $\theta = 0.01$, $n = 400$, $\epsilon = 10^{-5}$, and $V = 6$, $\forall m = \mathbb{M}$. Various set of users, depending on their channel conditions, have been paired together. The impact of the transmit SNR on T_{ec} of different pairing sets is illustrated in Fig. 3.6 which shows that the paring set $\phi = \{(1, 6), (2, 5), (3, 4)\}$ provides the higher T_{ec} as compared to the other pairing sets. This shows that when the users with distinct channel conditions are paired together, they achieve higher T_{ec} as compared to the pairing of users with less distinct channel conditions.

Figure 3.7 presents the plots of the achievable EC of the strong user and weak user versus the delay exponent θ, when $\rho = [15\text{dB}, 20\text{dB}]$, $n = 400$, and $\epsilon = 10^{-6}$. It is clear that the achievable EC of both users decreases when the delay exponent becomes stringent. More specifically, the gain in EC of the strong user at the loose delay requirements (low values of θ) is more significant (with large gap) at the same values of ρ as compared to the weak user. However, as the delay exponent becomes stringent, the EC of the weak user seems to be more stable as compared to that of the strong user, i.e., the weak user can tolerate more stringent delay.

One of the finite blocklength features, i.e., the transmission error probability ϵ for the achievable EC of the two-user NOMA has been analyzed in Figs. 3.8 and

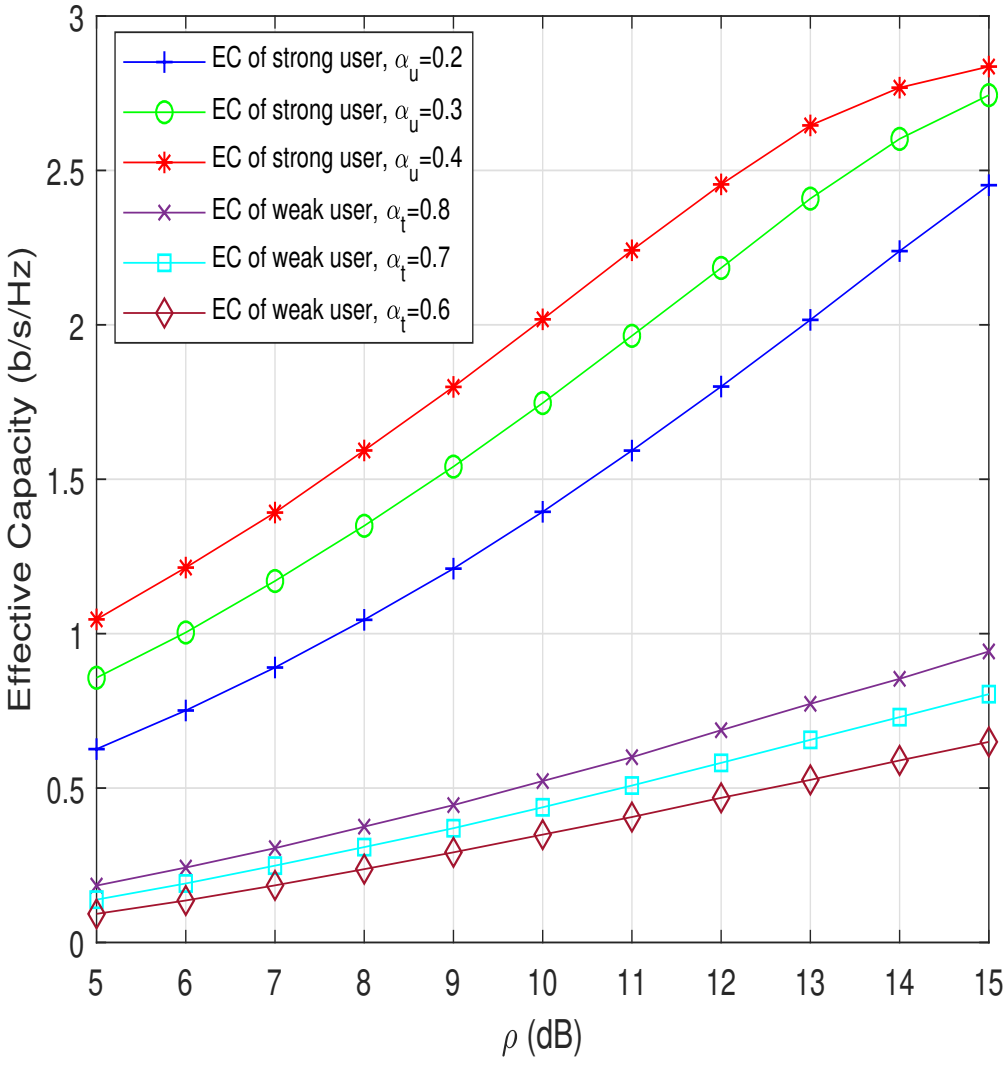

Figure 3.5: Effective capacity of the weak and strong users versus ρ, with $\theta = 0.01$, $n = 400$, and $\epsilon = 10^{-5}$.

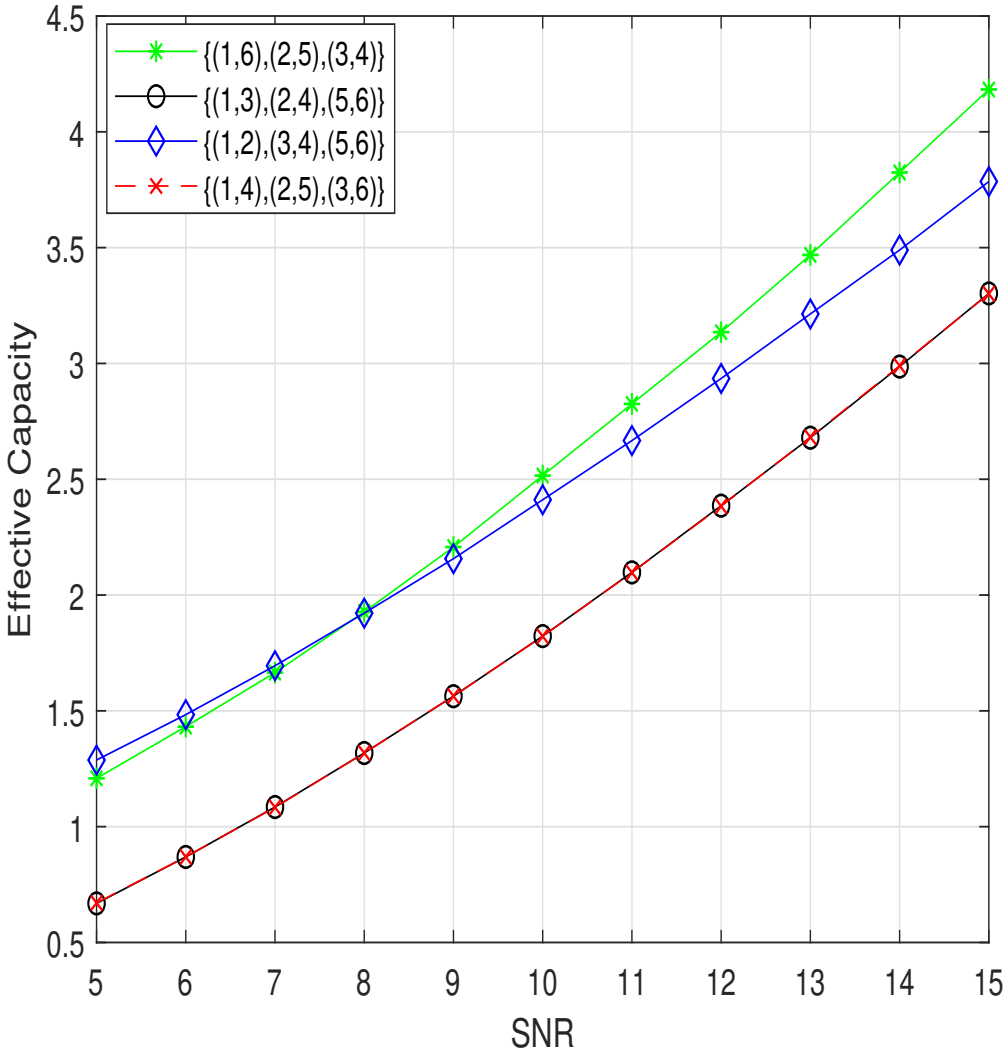

Figure 3.6: Total effective rate of multiple NOMA pairs versus transmit SNR, with $\theta = 0.01$, $\epsilon = 10^{-5}$, and $V = 6$.

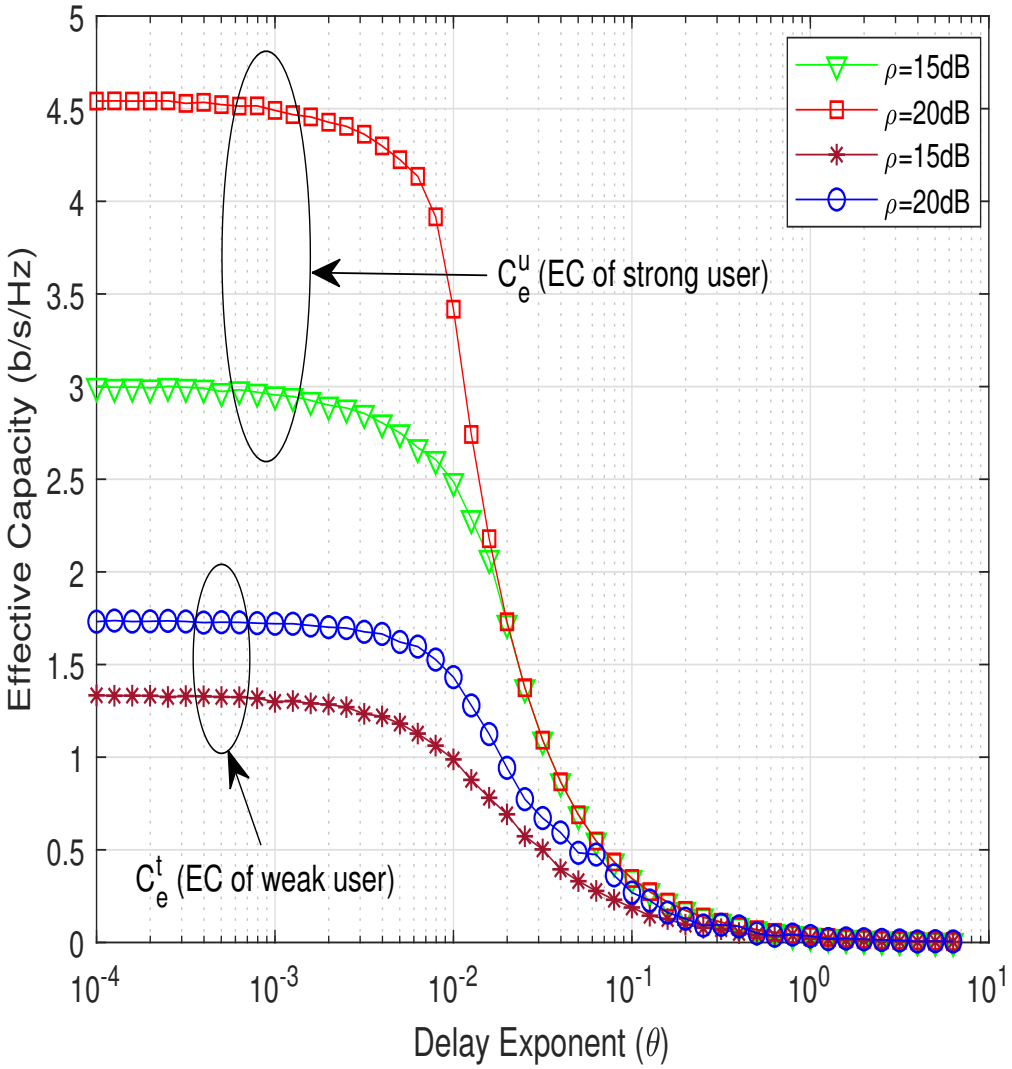

Figure 3.7: Effective capacity of NOMA strong and weak users versus delay exponent θ, with $\epsilon = 10^{-6}$.

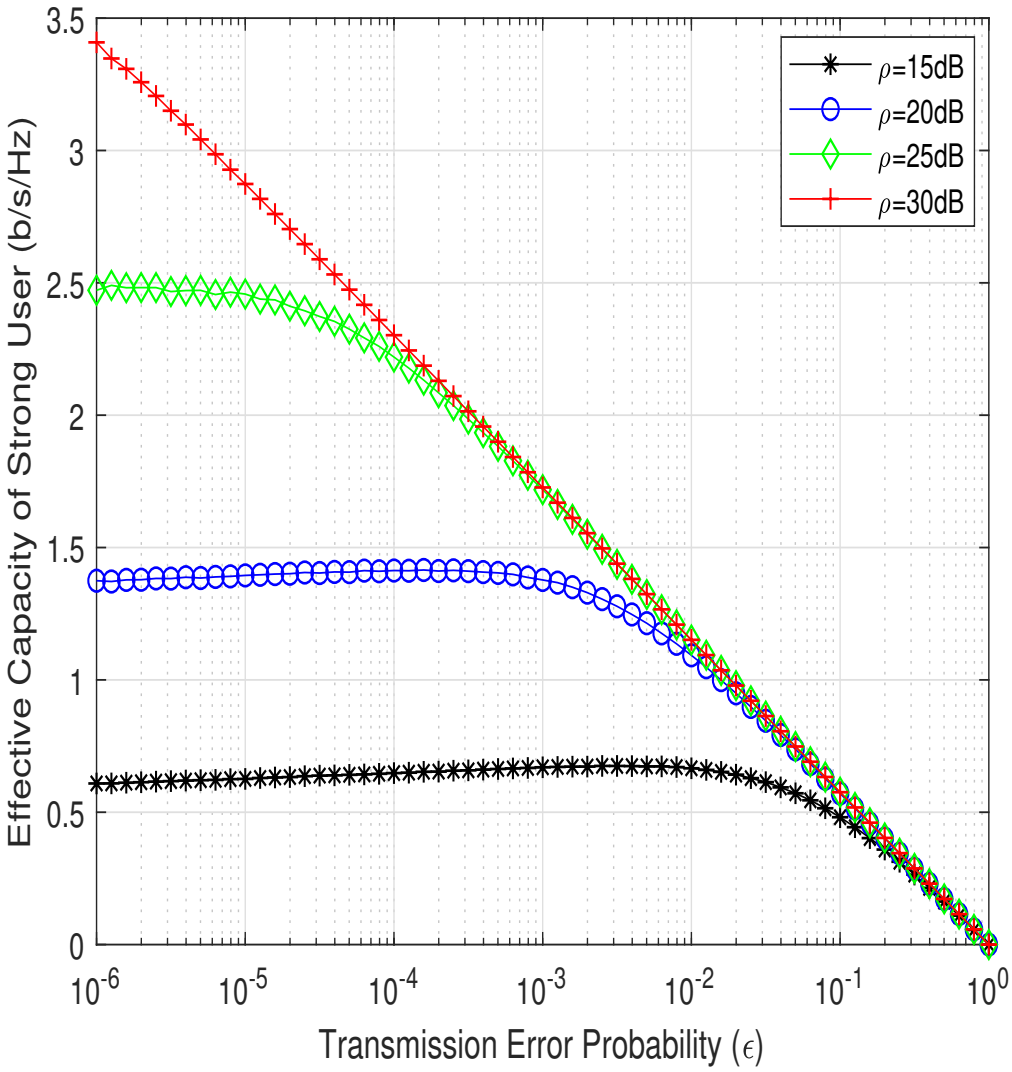

Figure 3.8: Effective capacity of NOMA strong user versus transmission error probability (ϵ), with $n = 400$, and $\theta = 0.01$.

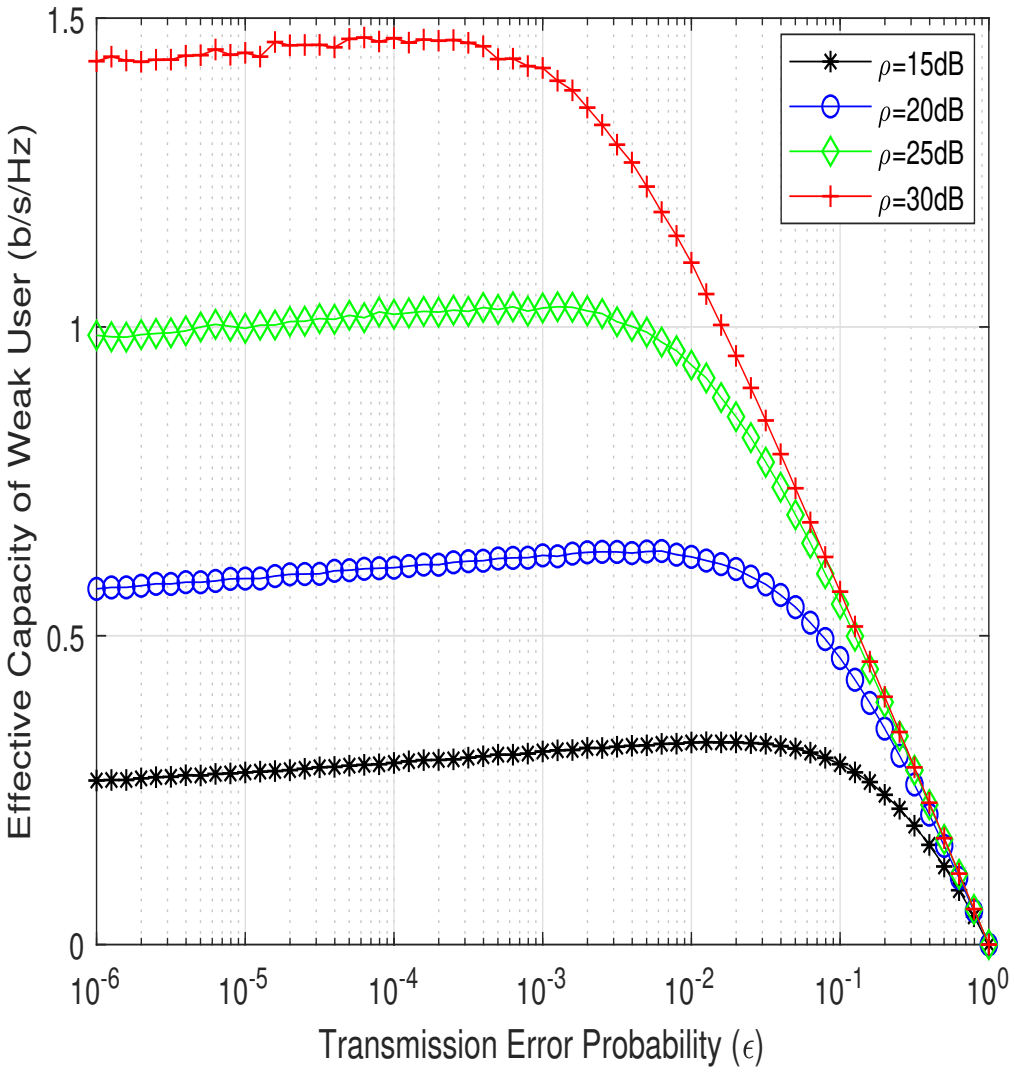

Figure 3.9: Effective capacity of NOMA weak user versus transmission error probability (ϵ), with $n = 400$ and $\theta = 0.01$.

3.9. In Fig. 3.8, the EC of the strong user is plotted versus ϵ for various values of ρ, while the blocklength is set to $n = 400$ and $\theta = 0.01$. The readers should refer to Eq. (3.11) for further clarification on the behavior of the plots in this figure. It is clear that when ρ is very small, i.e., 15 dB, the term $\left((1-\epsilon)\,(1+\alpha_u\gamma_u)^{2\zeta_u}e^{\beta_u\delta_u}\right)$ from the EC formulation in (3.11) is big as compared to ϵ. This results into sudden decrease in the achievable EC at low ρ and higher values of transmission error probability. However, as the value of ρ increases, the ϵ factor becomes dominant, which results into the minimum EC gain yet more stable as compared to the case with low ρ.

Figure 3.9 plots the EC of the NOMA weak user versus ϵ for various values of ρ. As compared to the Fig. 3.8, this figure shows a considerable decrease in the EC due to the weak channel conditions of user v_t. The readers should refer to Eq. (3.18) for further clarification on the behavior of the plots in this figure. As compared to the strong user, the ϵ factor (due to short-packet communication) in the weak-user's achievable EC remains more dominant as compared to the $\left((1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}e^{\beta_t\delta_t}\right)$ factor, even at the higher value of ρ (30dB). When the value of ϵ increases, the steady trend of the EC diminishes and, at very high values of ϵ, the EC of the weak user becomes zero.

Figure 3.10 shows the variation of the queuing delay violation probability versus the delay exponent, θ. The delay threshold is set to $D_{\max} = 400$, $n = 400$, and $\epsilon = 10^{-6}$. The trend of the plots in this figure can well be understood by following the EC formulation of the strong user from (3.11). It is clear that, as θ becomes more stringent, the queueing delay violation probability cannot be improved further below a certain value. This is due to the dominance of ϵ as compared to the term $\left((1-\epsilon)\,(1+\alpha_u\gamma_u)^{2\zeta_u}e^{\beta_u\delta_u}\right)$ in the EC formulation. For high value of θ, $\left((1-\epsilon)\,(1+\alpha_u\gamma_u)^{2\zeta_u}e^{\beta_u\delta_u}\right)$ is very small and, hence, ϵ becomes the dominant factor. However, the strong user shows a considerably high improvement in the queuing delay violation probability as compared to the weak user, which is due to the better channel conditions of the former.

Figure 3.11 illustrates the queuing delay violation probability versus the delay exponent for various values of ρ for the weak user. As the SINR and weak-user channel conditions are in focus here, these result into the queuing delay violation

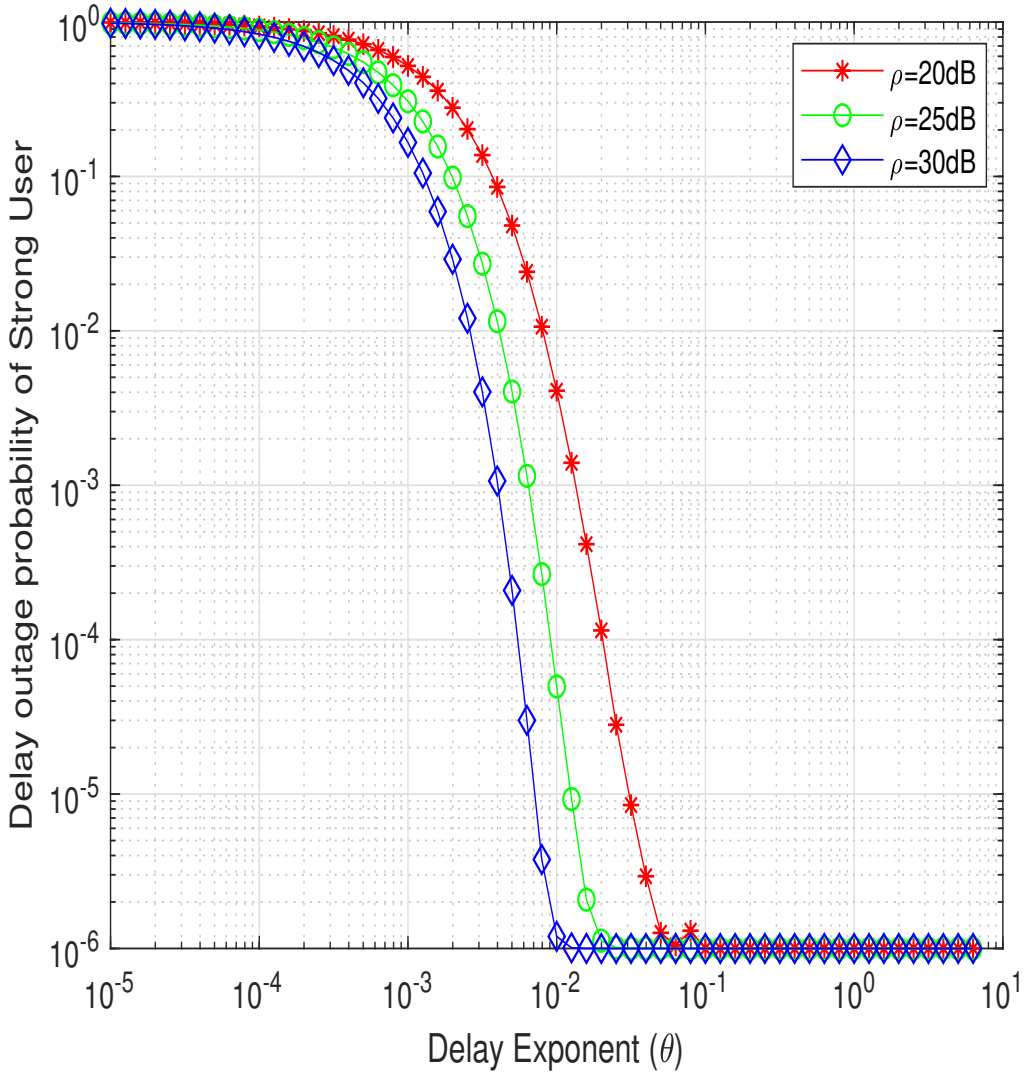

Figure 3.10: Queuing delay violation probability versus QoS exponent (θ) for the strong user, with $D_{\max} = 400$, $\epsilon = 10^{-6}$, and $n = 400$.

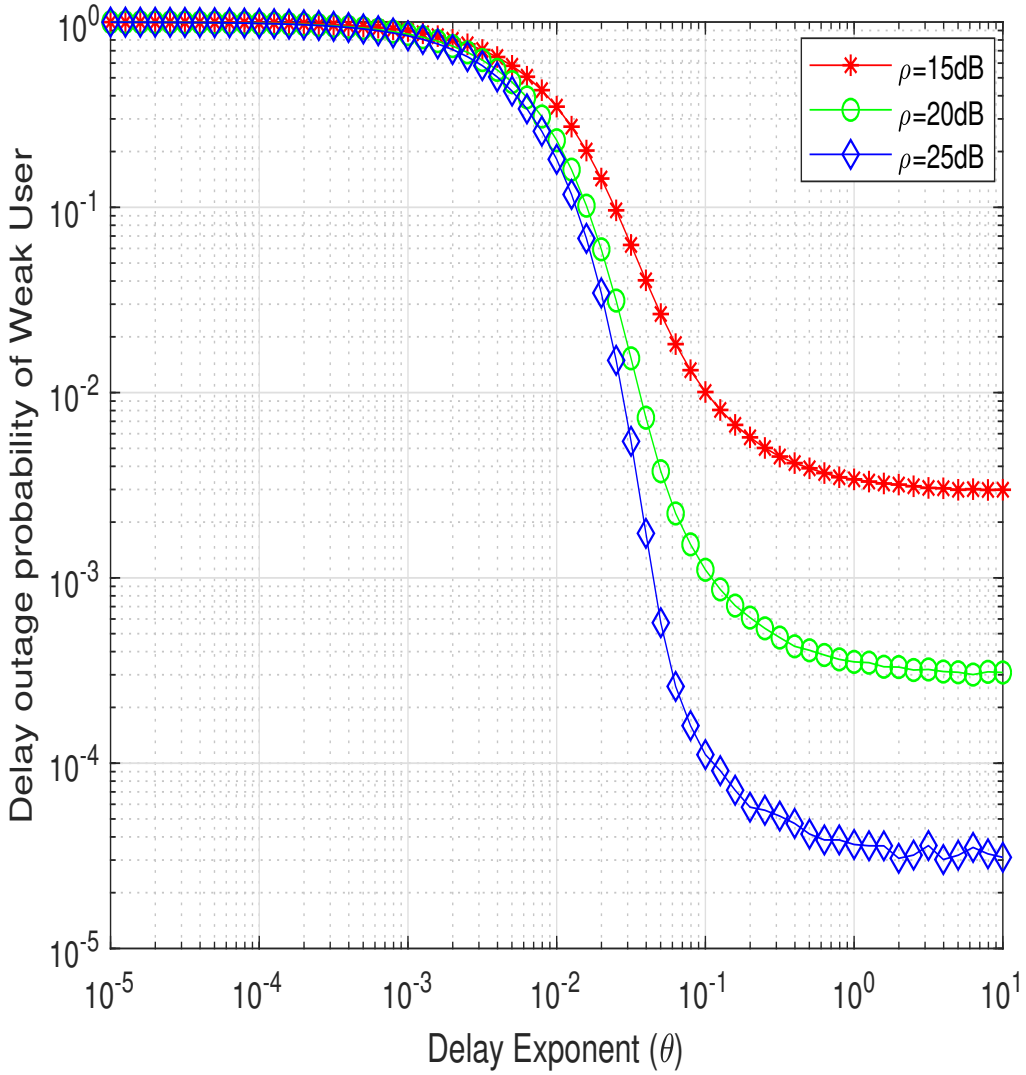

Figure 3.11: Queuing delay violation probability versus QoS exponent (θ) for the weak user, with $D_{\max} = 400$, $\epsilon = 10^{-6}$, and $n = 400$.

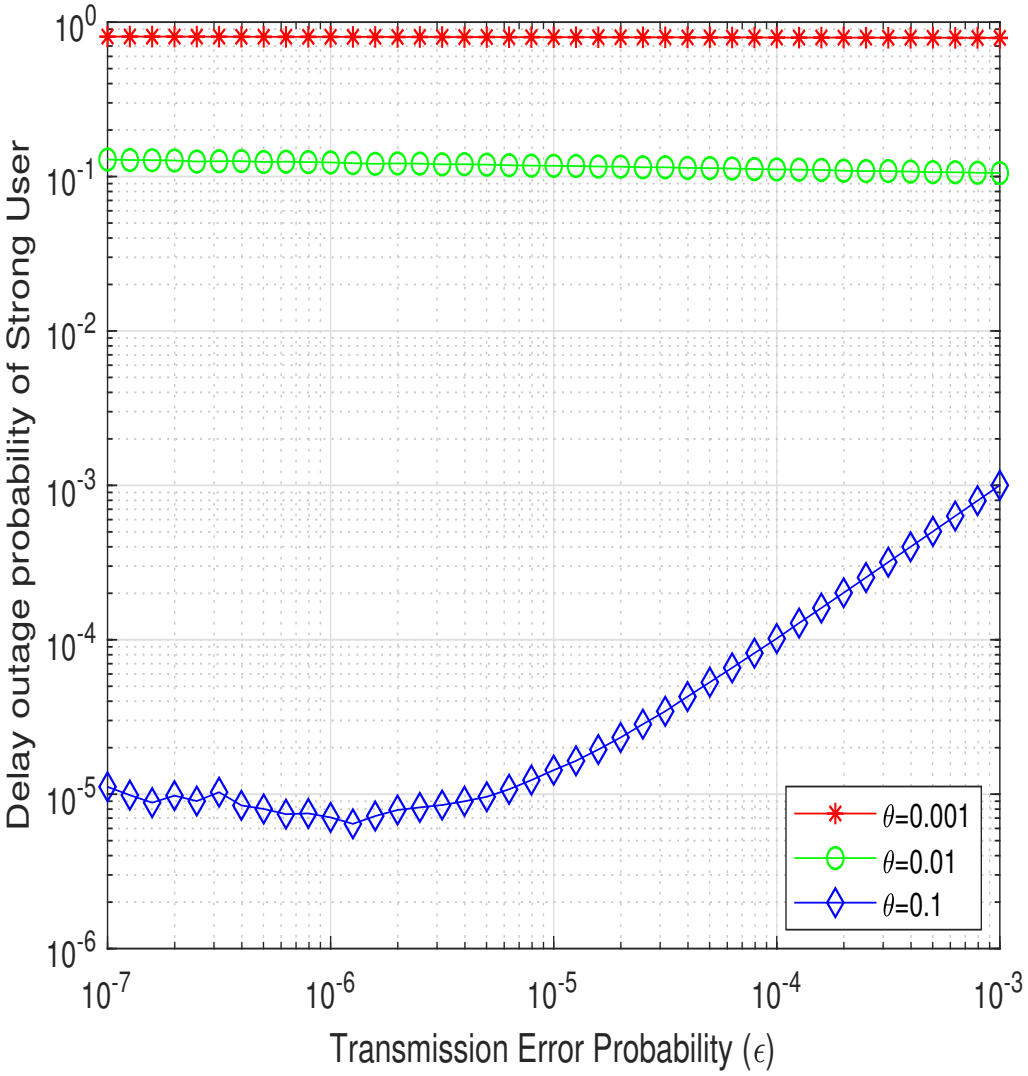

Figure 3.12: Queuing delay violation probability versus transmission error probability (ϵ) for the strong user, with $D_{\max} = 100$, $n = 100$, and $\rho = 20$dB.

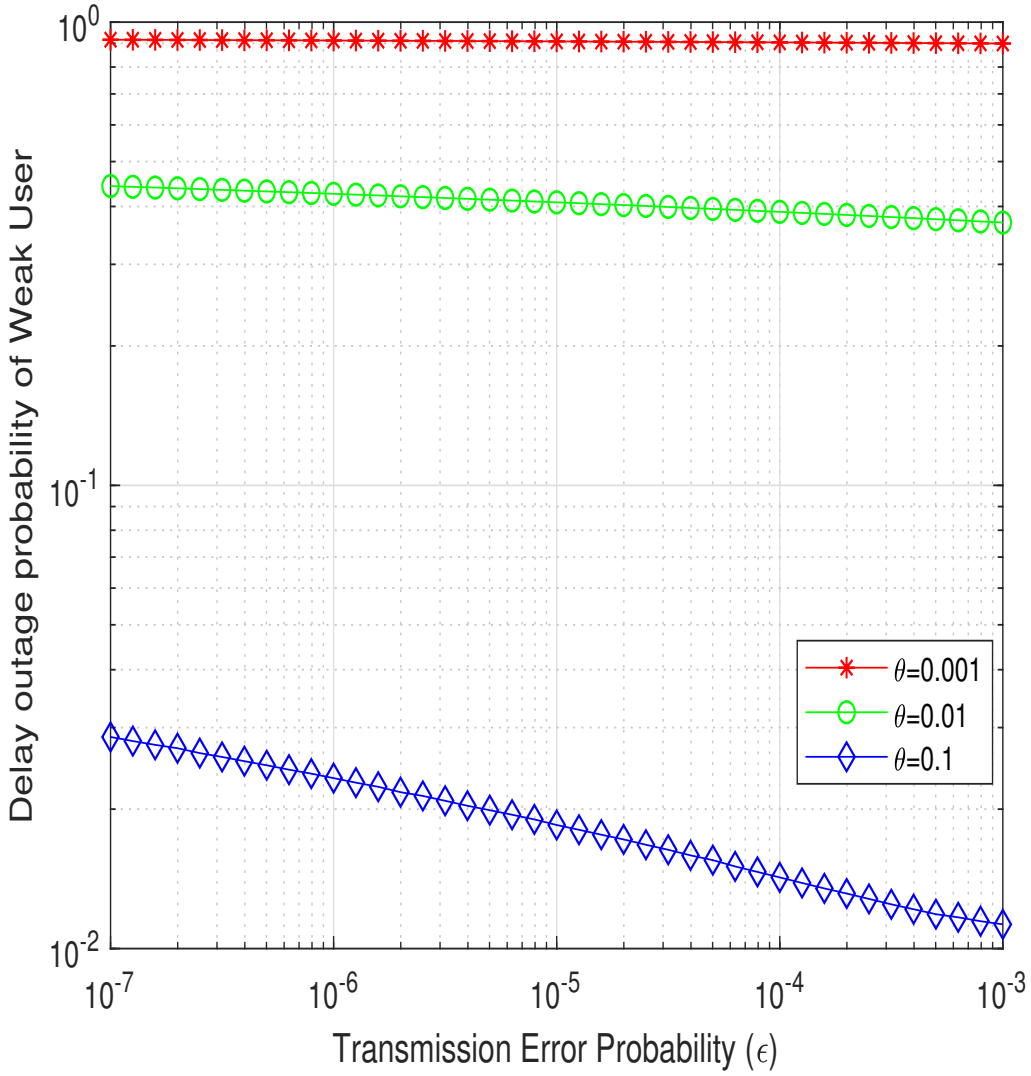

Figure 3.13: Queuing delay violation probability versus transmission error probability (ϵ) for the weak user, with $D_{\max} = 100$, $n = 100$, and $\rho = 20$dB.

probability to be restricted at different values versus θ. As θ becomes stringent, the queuing delay violation probability does not improve below a certain limit due to the characteristics of short-packet communication. When compared to the strong user, the weak user does not show a considerable improvement in queueing delay violation probability.

In Fig. 3.12 and Fig. 3.13, interesting trends of the queuing delay violation probability and ϵ for various values of the delay exponent θ have been analyzed for the strong and weak users, when $D_{\max} = 100$, $\rho = 20$, and the blocklength $n = 100$. As observed, when the delay requirements are loose, ϵ does not have any significant impact on the queuing delay violation probability. However, when the delay exponent becomes more stringent, ϵ has a significant impact on the queuing delay violation probability. This trend further confirms that, when the delay exponent becomes stringent, the queuing delay violation probability does not improve below a certain value due to the dominance factor of ϵ.

Figure 3.13 shows the queuing delay violation probability versus ϵ for different values of the delay exponent for the user with weak channel conditions. It is evident that the impact of the queuing delay violation probability on ϵ is not very significant when the delay exponent is loose. ON the other hand, when the delay exponent is stringent, i.e., $\theta \to [0.1]$, then ϵ has a significant impact on the queuing delay violation probability. As compared to the strong user, the weak user does not show much improvement in queuing delay violation probability. This result also confirms the impact of short-packet communication on the queuing delay violation probability of the NOMA weak user.

3.7 Summary

Effective capacity (EC)-based performance analysis of a two-user (out of V users) non-orthogonal multiple access (NOMA) network in finite blocklength regime was investigated in detail in this chapter. Overall reliability requirements were analyzed by taking into consideration the queuing delay violation probability and the transmission error probability. The closed-form expressions for the individual EC of strong and weak users is derived, and their accuracy was confirmed using Monte-Carlo simulations. The achievable EC of multiple NOMA pairs in finite

blocklength regime was also investigated, which showed that users with distinct channel conditions achieve more EC as compared to the users with less distinct channel conditions. It was found that for two-user NOMA, loose delay requirements do not have a significant impact on the queuing delay violation probability under transmission error probability constraint. However, when the delay exponent becomes more stringent, the queueing delay violation probability cannot be improved below a certain value under constraints on the transmission error probability.

Chapter 4

NOMA versus OMA in Finite Blocklength Regime: Link-Layer Rate Performance

4.1 Introduction

Transition from ultra-low latency to massive ultra reliable and low latency communications (mURLLC) for beyond 5G (B5G) applications demands the researchers from both industry and academia to revisit the enabling technologies. Future smart cities, autonomous robotics, holographic communications, blockchain, and massive sensing are few examples to name that require the mURLLC service class of B5G [3]. However, achieving mURLLC for the target future applications is a challenging task. While the non-orthogonal multiple access (NOMA) in conjunction with finite blocklength (short packet) communications is considered as an enabler for low-latency communications [109], further research is required to quantify the end-to-end latency in these systems. Also, the scalability of this technology is yet to be investigated.

NOMA with finite blocklength has the potential to allow ultra-low latency, massive connectivity, and higher throughput under favorable conditions [105]. The principle of NOMA in finite blocklength regime follows the conventional concept of NOMA, with superposition coding at the transmitter and successive

interference cancellation (SIC) at the receiver [28]. However, when operating with finite blocklength packets, the Shannon formula is not a good approximate for the achievable rate of NOMA, and alternative solutions are needed. In this vein, the authors in [8] provided a framework to approximate the achievable rate of a point-to-point communication link in finite blocklength regime.

Can the latency requirements of mURLLC for B5G services be satisfied with NOMA in finite blocklength regime? This question needs a detailed delay performance analysis of NOMA in finite blocklength regime. In this regard, the authors in [13] investigated the performance of NOMA with short-packet communications subject to reliability constraint. More specifically, the mentioned work showed the reduction in physical-layer transmission latency while using NOMA in conjunction with short-packet communications. The latency performance of NOMA with finite blocklength was further investigated in [106], which confirmed the improved performance of NOMA in terms of reducing latency and improving throughput, in comparison to orthogonal multiple access (OMA). There are also some work that investigated only the conventional NOMA, OMA, and hybrid scheme with NOMA/OMA scheme with the different dimensions such power allocation. Hybrid NOMA/OMA scheme is also employed in many networks and case studies. For example in machine type communications for ultra dense network, the hybrid NOMA and OMA scheme improves the network performance. This hybrid technique provides the flexibility to assign the resources (bandwidth) for a single user and multiple users per resource block depending on the requirements. For example, using the OMA scheme, one user per resource block is assigned the dedicated different normalized bandwidth, while the other cluster of users sharing the same resource block (NOMA) use the other resource block. The optimal power allocation scheme for NOMA, OMA, and hybrid NOMA/OMA scheme was provided in [117]. In this work, a power allocation scheme name the stationary power allocation scheme was proposed to allocate the power to the users. However, this work did not consider the short-packet communications and the link-layer tool to estimate the delay performance of NOMA or OMA. Short-packet communications based NOMA was also compared with OMA while considering the changing blocklength scenario in [118]. The performance of the two-user NOMA and OMA in finite blocklength was studied using the hybrid automatic repeat request with

chase combining (HARQ-CC). A closed-form expression was derived for the two-user NOMA and OMA considering the block error rate (BLER). A performance comparison between the two transmission schemes shows that NOMA outperformed the OMA with the best user fairness.

Power optimization for the NOMA network while residing within the short-packet communication is of significant importance. Employing the short-packet communications means that the blocklength and transmission error penalty should also be considered while designing any power allocation scheme. Joint power and blocklength optimization for the NOMA short-packet communications was studied in [38]. In this work, the performance of the NOMA was also compared with the OMA counterpart while considering the short-packet communications and joint power and blocklength optimization. This work was of significant importance as through the joint power and blocklength optimization the decoding probability was reduced under constraint of energy. However, still the detailed NOMA and OMA short-packet communications analysis based on the latency performance was not done in this work. NOMA short-packet communications is the enabler of the uRLLC. Authors in [119], studied the outage probability of NOMA short-packet communications with the advance feature of wireless power transfer for achieving the uRLLC. A closed-form expression for the outage probability of NOMA users was derived to define the bounds for the battery capacity. A relationship between the latency and reliability was also investigated and a performance comparison was also provided between the NOMA and OMA users using the short-packet communications. NOMA transmission using the short-packet communications with MIMO systems was also investigated in [120]. The downlink MIMO NOMA using the short-packet communications regime was studied using the Nakagami-m fading channels. As compared to the conventional approach that used the ergodic capacity and the outage capacity as the performance metric, in this work BLER was considered. The focus was to minimize the blocklength for its used for the futuristic mission-critical applications of IoT that will require the small data payload for their operation.

Also, a comparative view of the achievable effective capacity (EC) of uplink two-user NOMA and OMA was conducted in [121], but not for transmissions in finite blocklength regime. Later in [104], the achievable EC for systems with fi-

nite blocklength was analyzed, and it was shown that the proposed system within short blocklength and reliability constraint can reduce latency, hence establishing the importance of short-packet communications for achieving low latency. Focusing on the importance of short packet communications, the achievable EC for finite blocklength machine-type communications (MTC) under delay constraint was derived in [122]. In that work, the optimum error probability was characterized under the effect of signal-to-noise ratio (SNR) variations to maximize the achievable EC, and it was confirmed that under strict delay constraints, the SINR variations have less effect on the achievable EC of MTC.

In this chapter, the latency performance of NOMA and OMA with short-packet communications is investigated. The major contributions of this chapter can be summarized as follows:

- The achievable EC of two-user NOMA and OMA in finite blocklength regime is derived.[1] Specifically, the achievable EC (link-layer rate) of NOMA users is investigated in finite blocklength regime under heterogeneous delay quality-of-service (QoS) requirements, in comparison with the OMA counterpart.

- Closed-form expressions for the individual users' EC in the two-user NOMA and OMA networks over Rayleigh fading channels is derived, and its accuracy is confirmed using Monte-Carlo simulations.

- Under Rayleigh fading and with strict delay, the OMA user with better channel conditions outperform both NOMA users at low SNRs,

- The simulation results of the achievable EC of two-users NOMA and OMA in short-packet communications under different fading channels and with heterogeneous delay requirements are also investigated. It is shown that the achievable EC of two-user OMA under strict delay constraint with severe fading is better as compared to the two-users NOMA networks. However, under severe fading with loose delay, NOMA strong user outperforms the OMA counterpart.

[1]Two-user NOMA has been included as a building block in third generation partnership project long-term evolution advanced (3GPP-LTE-A) networks [13].

- Total link-layer rate of two-users NOMA and OMA in short-packet communications under different fading channels and with heterogeneous delay requirements are also compared. As compared to the individual EC of NOMA users, total link-layer rate of NOMA outperforms its counter part OMA under certain conditions.

4.2 Transmission Framework and Fundamentals

Consider a downlink two-user NOMA network with finite blocklength. The users, denoted by v_i, $i = \{1, 2\}$, are equipped with single antennas and communicate with a single base station (BS). The channel coefficient between the BS and v_i at time τ is referred to by $h_i(\tau)$. The two users are classified based on their channel conditions as strong and weak users and, without loss of generality, it is assumed that $|h_1(\tau)|^2 \geq |h_2(\tau)|^2$.

Following the NOMA principle, the BS broadcasts a combined message $\sum_{i=1}^{2} \sqrt{\alpha_i P} u_i(\tau)$ to its users, where u_i is the message corresponding to user v_i, P is the BS's total transmit power, and α_i is the power coefficient for user v_i. With fixed power allocation policy at the BS, the power coefficients for the two users are such that $\alpha_1 \leq \alpha_2$. The received signal at user v_i can now be formulated as[1]

$$y_i = h_i \sum_{i=1}^{2} \sqrt{\alpha_i P} u_i + m_i \tag{4.1}$$

where m_i is the additive white Gaussian noise (AWGN) at v_i, $i \in \{1, 2\}$.

At the receiving side, the strong user (v_1) first performs SIC to remove interference (u_2) from its received signal (y_1), and then decodes its own message. Therefore, for user v_1, the received SNR, denoted by SNR_1^{N},[2] can be found as

$$\text{SNR}_1^{\text{N}} = \alpha_1 \rho \, |h_1|^2 , \tag{4.2}$$

where ρ is the transmit SNR, namely $\rho = \frac{P}{N_o B}$, in which $N_o B$ denotes the noise

[1] As the channel coefficients are assumed stationary and ergodic random processes, the time index τ is omitted hereafter for simplicity of presentation.

[2] Superscript N indicates NOMA. Later, notation O will be used to indicate the OMA operation.

power.

On the other hand, the weak user (v_2) treats u_1 as interference and decodes its own message directly. Hence, its resulting signal-to-interference-plus-noise ratio (SINR) can be derived as

$$\mathrm{SINR}_2^{\mathrm{N}} = \frac{\alpha_2 \rho \left|h_2\right|^2}{\alpha_1 \rho \left|h_2\right|^2 + 1}. \tag{4.3}$$

Channel gains of both users are modeled as Rayleigh distributions with unit variance. Following the NOMA operation, the users v_1 and v_2 are sorted based on their ordered channel gains. Therefore, the probability density function (PDF) of the ordered channel power gains can be obtained using the order statistics [111]. In this regard, using $\rho \left|h_i\right|^2 = \gamma_i$ and denoting its PDF as $f\left(\gamma_i\right)$, the order statistics is applied to get

$$f_{\gamma_{1:2}}\left(\gamma_1\right) = \xi_1 f\left(\gamma_1\right) F\left(\gamma_1\right), \tag{4.4}$$

$$f_{\gamma_{2:2}}\left(\gamma_2\right) = \xi_2 f\left(\gamma_2\right) \left(1 - F\left(\gamma_2\right)\right), \tag{4.5}$$

where $f_{\gamma_{i:2}}$ is the PDF of the ordered γ_i out of two users, $\xi_i = \frac{1}{B(i,2-i+1)}$, in which $B(a,b)$ is the Beta function [116], and $i \in \{1,2\}$.

For the case with OMA operation, both users have access to the same spectrum bandwidth as in the NOMA case but each user can only occupy half of the transmission time slot. Using the results of [8] as starting point, the users' achievable rates with finite blocklength in the NOMA and OMA cases under study can be formulated, in b/s/Hz, as

$$r_1^{\mathrm{N}} = \log_2\left(1 + \alpha_1 \gamma_1\right) - \sqrt{\frac{V_1^{\mathrm{N}}}{n}} Q^{-1}(\epsilon), \tag{4.6}$$

$$r_2^{\mathrm{N}} = \log_2\left(1 + \frac{\alpha_2 \gamma_2}{\alpha_1 \gamma_2 + 1}\right) - \sqrt{\frac{V_2^{\mathrm{N}}}{n}} Q^{-1}(\epsilon), \tag{4.7}$$

$$r_i^{\mathrm{O}} = \frac{1}{2}\left(\log_2\left(1+\gamma_i\right) - \sqrt{\frac{V_i^{\mathrm{O}}}{n}}Q^{-1}(\epsilon)\right),\ i \in \{1,2\}, \tag{4.8}$$

where r_1^{N}, r_2^{N} and r_i^{O} are the achievable rates of the NOMA strong user, NOMA weak user, and OMA users, respectively, n is the blocklength, ϵ is the transmission error probability, and $Q^{-1}(.)$ is the inverse of Gaussian Q-function [123] with

$$Q\left(x\right) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw. \tag{4.9}$$

In this transmission the channel dispersions of the NOMA strong user, NOMA weak user, and OMA users can be derived as

$$V_1^{\mathrm{N}} = 1 - \left(1+\alpha_1\gamma_1\right)^{-2} \tag{4.10}$$

$$V_2^{\mathrm{N}} = 1 - \left(1 + \frac{\alpha_2\gamma_2}{\alpha_1\gamma_2+1}\right)^{-2} \tag{4.11}$$

$$V_i^{\mathrm{O}} = 1 - \left(1+\gamma_i\right)^{-2} \tag{4.12}$$

The above channel dispersion are for the two-users NOMA and OMA networks when short-packet communications is employed. Compared to the Chapter 2, channel dispersion that takes into consideration the case of the transmit SNR and extremely high transit. In this transmission, the two-users NOMA and OMA channel dispersion is also considered as this transmission is at the extremely high SNR.

4.3 Effective Capacity of NOMA and OMA in Finite Blocklength Regime

In this section, the achievable EC of the two-user NOMA and OMA networks in finite blocklength communication regime is derived. Then a closed-form expressions for the EC is provided.

By following [104] and [103], the achievable EC of the two-user NOMA and

the OMA counterpart in finite blocklength regime can be formulated as

$$C_i^{\mathrm{N}} = -\frac{1}{\theta_i n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\, e^{-\theta_i n r_i^{\mathrm{N}}}\right]\right), \tag{4.13}$$

$$C_i^{\mathrm{O}} = -\frac{1}{\theta_i n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\, e^{-\theta_i n r_i^{\mathrm{O}}}\right]\right), \tag{4.14}$$

where C_i^{N} and C_i^{O} represent the EC of user v_i in finite blocklength regime, for NOMA and OMA, respectively, and $\mathbb{E}[.]$ is the expectation operator.

By considering the service rate r_i^{N} for users v_i in finite blocklength regime from (4.6) and (4.7), the achievable EC of the NOMA strong user and the NOMA weak user can be approximated as

$$C_1^{\mathrm{N}} = -\frac{1}{\theta_1 n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\,(1+\alpha_1\gamma_1)^{2\Upsilon_1}\, e^{\psi_1\sqrt{V_1^N}}\right]\right), \tag{4.15}$$

and

$$C_2^{\mathrm{N}} = -\frac{1}{\theta_2 n} \times \ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\left(1+\frac{\alpha_2\gamma_2}{\alpha_1\gamma_2+1}\right)^{2\Upsilon_2} e^{\psi_2\sqrt{V_2^N}}\right]\right), \tag{4.16}$$

respectively, where $\Upsilon_i = -\frac{\theta_i n}{2\ln 2}$, and $\psi_i = \theta_i\sqrt{n}Q^{-1}(\epsilon)$.

As specified, users v_1 and v_2 can also operate according to OMA, by transmitting their messages using time division multiple access (TDMA). For the OMA case, using (4.8) the achievable EC of the two users can be approximated as

$$C_i^{\mathrm{O}} = -\frac{1}{\theta_i n}\ln\left(\mathbb{E}\left[\epsilon + (1-\epsilon)\,(1+\gamma_i)^{\Upsilon_i}\, e^{\frac{\psi_i\sqrt{V_i^O}}{2}}\right]\right). \tag{4.17}$$

The above derived individual EC expressions of the two users with NOMA or OMA in finite blocklength regime can be used to analyze and compare the delay performance in both operation scenarios.

To further simplify the above expressions, a closed-form expressions for the individual EC of the strong and weak NOMA and OMA users in finite blocklength regime is derived. Specifically, using the order statistics from (4.4) and (4.5) the

achievable EC of a two-users NOMA and OMA can be expanded as

$$C_1^{\mathrm{N}} = -\frac{1}{\theta_1 n}\ln\Bigg(\int_0^{\infty}\Big(\epsilon + (1-\epsilon)(1+\alpha_1\gamma_1)^{2\Upsilon_1}e^{\psi_1 V_1^N}\Big) \times f_{\gamma_{1:2}}(\gamma_1)d\gamma_1\Bigg), \tag{4.18}$$

$$C_2^{\mathrm{N}} = -\frac{1}{\theta_2 n}\ln\Bigg(\int_0^{\infty}\Bigg(\epsilon+(1-\epsilon)\Bigg(\frac{\gamma_2+1}{\alpha_1\gamma_2+1}\Bigg)^{2\Upsilon_2} \times e^{\psi_2 V_2^N}\Bigg)f_{\gamma_{2:2}}(\gamma_2)d\gamma_2\Bigg), \tag{4.19}$$

$$C_1^{\mathrm{O}} = -\frac{1}{\theta_1 n}\ln\Bigg(\int_0^{\infty}\Big(\epsilon + (1-\epsilon)(1+\gamma_1)^{\Upsilon_1}e^{\frac{\psi_1 V_1^O}{2}}\Big) \times f_{\gamma_{1:2}}(\gamma_1)d\gamma_1\Bigg), \tag{4.20}$$

$$C_2^{\mathrm{O}} = -\frac{1}{\theta_2 n}\ln\Bigg(\int_0^{\infty}\Big(\epsilon + (1-\epsilon)(1+\gamma_2)^{\Upsilon_2}e^{\frac{\psi_2 V_2^O}{2}}\Big) \times f_{\gamma_{2:2}}(\gamma_2)d\gamma_2\Bigg), \tag{4.21}$$

where $f(\gamma_i) = \frac{1}{\rho}e^{-\frac{\gamma_i}{\rho}}$, $F(\gamma_i) = 1 - e^{-\frac{\gamma_i}{\rho}}$, and it is assumed at high SNR $V_i^N \approx 1, V_i^O \approx 1$ [8]. The final closed-form expressions for the two users, in NOMA and OMA, can be obtained by solving the above integrals.

To obtain the closed-form expressions for C_1^{N}, C_1^{O}, and C_2^{O}, first the simple case of C_1^{O} is considered and its closed-form expression is derived. In this regard,

C_1^{O} (from (4.20)) can further be expanded as

$$C_1^{\mathrm{O}} = -\frac{1}{\theta_1 n}\ln\Bigg(\epsilon + (1-\epsilon)\,\frac{2}{\rho}e^{\frac{\psi_1}{2}}\Big(\underbrace{\int_0^{\infty}(1+\gamma_1)^{\Upsilon_1}\,e^{-\frac{\gamma_1}{\rho}}\,d\gamma_s}_{I_1} - \underbrace{\int_0^{\infty}(1+\gamma_1)^{\Upsilon_1}\,e^{-\frac{2\gamma_1}{\rho}}\,d\gamma_1}_{I_2}\Big)\Bigg), \tag{4.22}$$

where let's recall $\xi_1 = \frac{1}{B(1,2-1+1)}$ and $\Upsilon_1 = -\frac{\theta_1 n}{2\ln 2}$. Introduce the equality from [eq (13.2.5) [116]]

$$\mathrm{H}\,(a,b,z) = \frac{1}{\Gamma\,(a)}\int_0^{\infty} e^{-zt}t^{a-1}\,(1+t)^{b-a-1}\,dt \quad \text{for Re(a), Re(z)} > 0, \tag{4.23}$$

where H(.,.,.) is the confluent hypergeometric function of the second kind [116]. By using (4.23), the integrals I_1 and I_2 can be solved as

$$I_1 = \mathrm{H}\left(1, 2+\Upsilon_1, \frac{1}{\rho}\right), \tag{4.24}$$

$$I_2 = \mathrm{H}\left(1, 2+\Upsilon_1, \frac{2}{\rho}\right). \tag{4.25}$$

Inserting (4.24) and (4.25) into (4.22), the closed-form expression for C_1^{O} can finally be derived and is given in (4.28)

Similarly, following the above steps, the closed-form expressions for C_1^{N} and C_2^{O} (given in (4.26) and (4.29)) can also be obtained. The closed-form expression for the weak NOMA user can also be found. And the steps followed for finding the closed-form expression for the NOMA weak user are given in Appendix C, and the derived final closed form expression is given as (4.27).

$$C_1^{\mathrm{N}} = -\frac{1}{\theta_1 n}\ln\left(\epsilon + (1-\epsilon)\,\frac{2}{\alpha_1\rho}e^{\psi_1}\left(\mathrm{H}\left(1, 2+2\Upsilon_1, \frac{1}{\alpha_1\rho}\right) - \mathrm{H}\left(1, 2+2\Upsilon_1, \frac{2}{\alpha_1\rho}\right)\right)\right). \tag{4.26}$$

$$C_2^{\mathrm{N}} = -\frac{1}{\theta_2 n}\ln\left(\epsilon + (1-\epsilon)\,\frac{2\alpha_1^{-2\Upsilon_2}}{\rho}e^{\psi_2}\left(\mathrm{H}\left(1, 2, \frac{2}{\rho}\right) + \frac{n\theta_2\,(\alpha_1-1)}{\alpha_1 \ln 2}e^{\frac{2}{\alpha_1\rho}}\mathrm{E_i}\left(-\frac{2}{\alpha_1\rho}\right)\right.\right.$$
$$\left.\left. + \sum_{k=2}^{\infty}\binom{2\Upsilon_2}{k}\left(\frac{\alpha_1-1}{\alpha_1}\right)^k\left(\frac{\sum_{j=1}^{k-1}\frac{(j-1)!}{\alpha_1^{-j}}\left(-\frac{2}{\rho}\right)^{k-j-1} - \left(-\frac{2}{\rho}\right)^{k-1}}{(k-1)!}e^{\frac{2}{\alpha_1\rho}}\mathrm{E_i}\left(-\frac{2}{\alpha_1\rho}\right)\right)\right)\right). \tag{4.27}$$

$$C_1^{\mathrm{O}} = -\frac{1}{\theta_1 n}\ln\left(\epsilon + (1-\epsilon)\,\frac{2}{\rho}e^{\frac{\psi_1}{2}}\left(\mathrm{H}\left(1, 2+\Upsilon_1, \frac{1}{\rho}\right) - \mathrm{H}\left(1, 2+\Upsilon_1, \frac{2}{\rho}\right)\right)\right). \tag{4.28}$$

$$C_2^{\mathrm{O}} = -\frac{1}{\theta_2 n}\ln\left(\epsilon + (1-\epsilon)\,\frac{2}{\rho}e^{\frac{\psi_2}{2}}\mathrm{H}\left(1, 2+\Upsilon_2, \frac{1}{\rho}\right)\right). \tag{4.29}$$

4.4 Numerical Results

In this section, extensive simulations to compare the performance of the two-user NOMA and two-user OMA in finite blocklength regime are performed. Numerical results are discussed and compared, considering the users' power coefficients $\alpha_2 = 0.7$ and $\alpha_1 = 0.3$, the blocklength $n = 400$, and a transmission error probability $\epsilon = 10^{-6}$, unless otherwise specified.

Fig. 4.1 shows the plots of the achievable EC of two-user NOMA and two-user OMA in finite blocklength regime as a function of the transmit SNR (ρ) in dB. For this evaluation, delay exponent is set $\theta = 0.01$. The accuracy of the derived closed-form expressions is confirmed. The figure also shows that, at very low transmit SNRs, the OMA strong user outperforms both NOMA users. However, as ρ increases, the achievable EC of NOMA and OMA does not increase further and saturates at very high values of the SNR. At low SNRs, the achievable EC of the weak user is approximately the same in both NOMA and OMA, whereas at high SNRs the weak user OMA dominates with a big gap.

Fig. 4.2 shows the plot of the achievable EC of two-user NOMA and OMA in finite blocklength regime versus the transmit SNR (ρ) under severe fading (one-sided Gaussian) under the constraint of the stringent delay exponent, $\theta = 0.01$. The two-user NOMA under the strict constraint of fading and delay underperforms compared to the two-user OMA short-packet communications. More specifically, the performance gap is very wide between the NOMA and OMA pairs, which shows that NOMA under fading and delay constraint for the two-user scenario is a poor choice. However, in case of the multi-user NOMA (that will be discussed later in this Section), its performance is much better compared to the OMA counterpart. This feature of NOMA short-packet communications show that NOMA is more suitable when massive connectivity scenarios are considered under stringent delay and fading constraint.

Fig. 4.3 is adds more insights in the previous plot. In this graph, achievable EC of two-user NOMA and OMA short-packet communications is plotted versus transmit SNR (ρ) under severe fading (one-sided Gaussian) under the constraint of the less stringent delay exponent, $\theta = 0.001$. In this case, the requirements of the delay are relaxed. Under loose delay, strong-user NOMA outperforms the

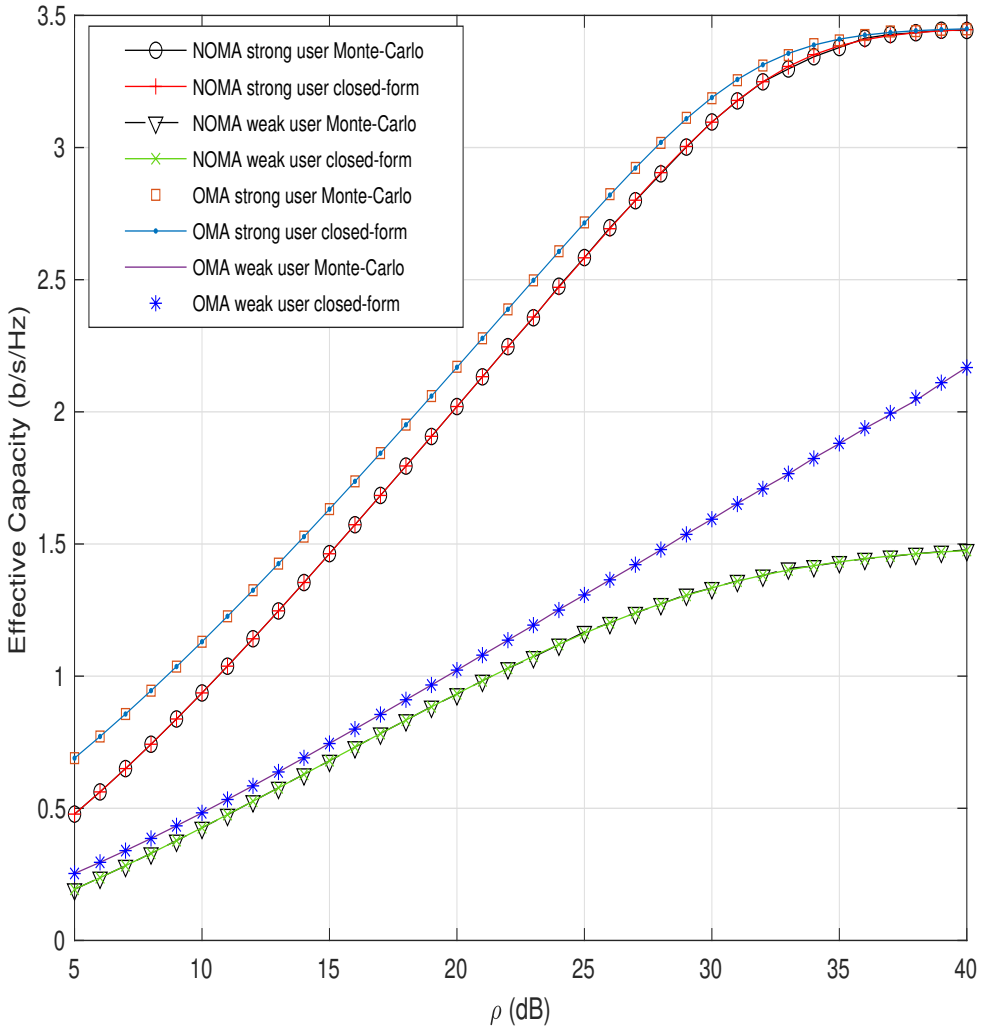

Figure 4.1: Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR with Rayleigh fading channel.

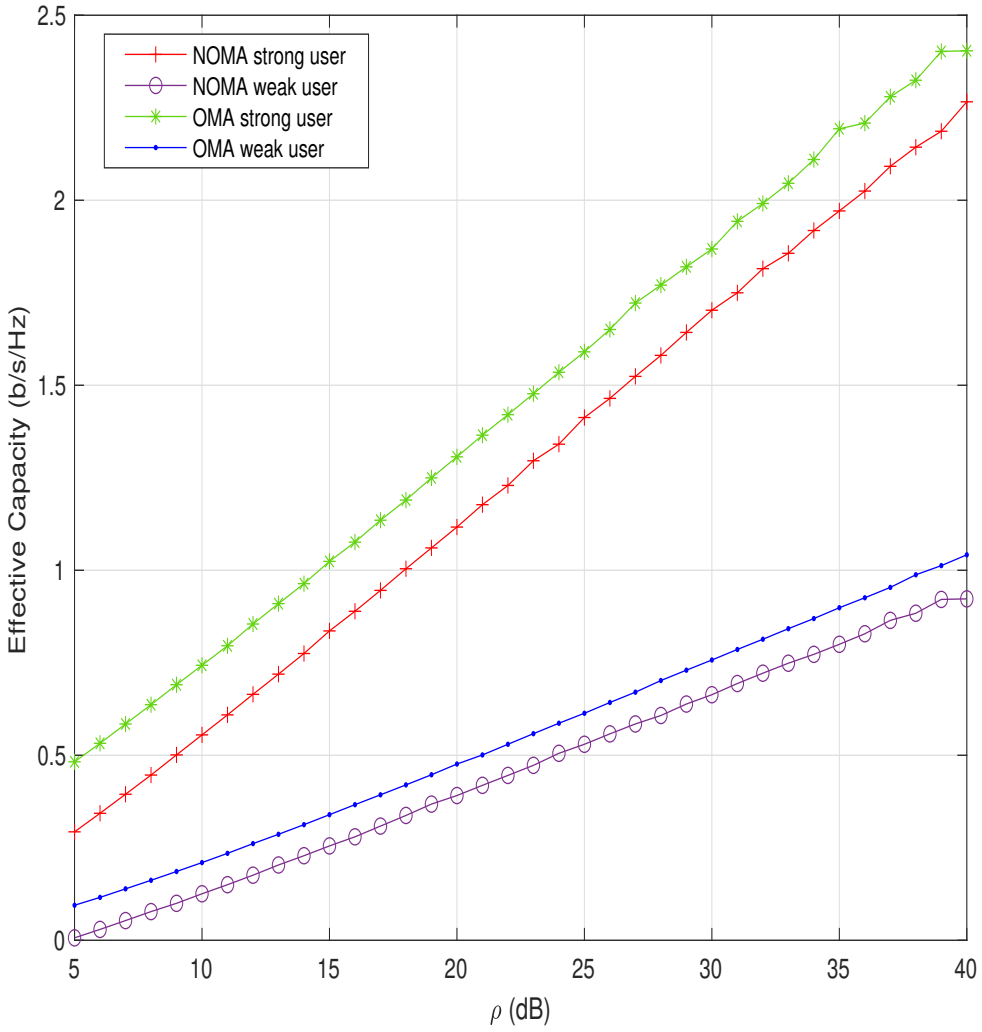

Figure 4.2: Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under severe fading (one-sided Gaussian) with stringent delay, i.e., $\theta = 0.01$.

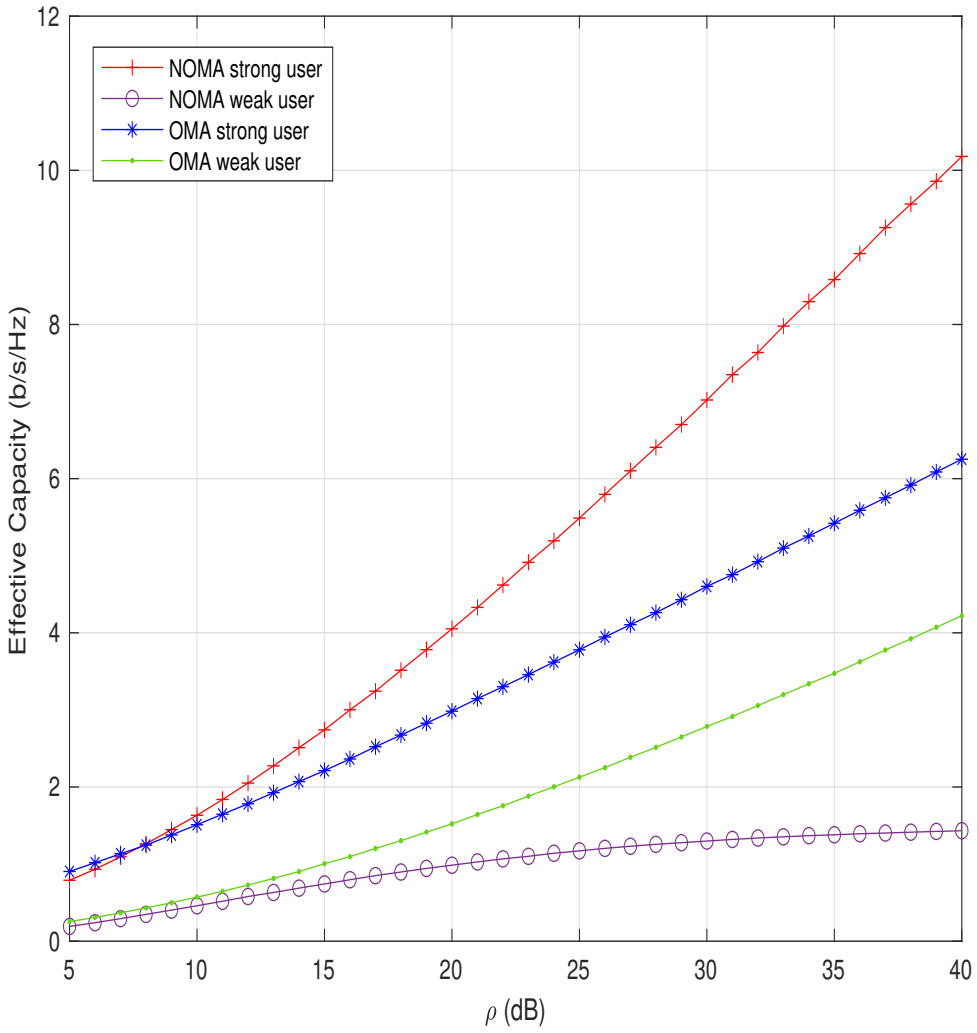

Figure 4.3: Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under severe fading (one-sided Gaussian) with less stringent delay, i.e., $\theta = 0.001$.

OMA with a more wider gap. However, still the weak-user OMA outperforms much better than the weak-user NOMA at high SNR. More interestingly, the performance gap between the strong-user NOMA and OMA is much wider as compared to the previous findings under strict delay constraints, where OMA outperformed the NOMA. At smaller value of the transmit SNR (from 5dB to 10dB), the performance of the strong and weak user OMA is better than the NOMA pairs.

Fig. 4.4 shows the simulation results of the achievable EC of two-user NOMA and two-user OMA against ρ under rician and lognormal fading model with strict delay constraint. At low transmit SNR (ρ), the performance of the OMA strong user is good, however at high SNR, the NOMA strong user performs way better as compared to the strong user OMA even under the strict delay constraint. However, the weak-user OMA shows a very interesting trend. Contrary to its partner (strong user OMA), whose performance degraded at the high SNR, the OMA weak-user achieves much better performance (many fold) at the high transmit SNR (ρ).

Fig. 4.5 shows the graph of the total achievable link-layer rate of two-user NOMA and OMA in finite blocklength regime with generalized fading conditions with loose delay constraint, $\theta = 0.001$. As compared to the individual EC of each NOMA and OMA users, the total EC of two-users NOMA and OMA shows different trends under different fading conditions with loose delay. Total link-layer rate of two-user NOMA is outperforming the two-user OMA under different fading conditions. This is entirely different from Fig. 4.3, where the individual EC of the weak-user OMA under the same conditions of fading and delay was better than the weak-user NOMA. However, in case of the total gain or total Link-layer rate, it is the two-user NOMA which has the more gain as compared to the total link-layer rate of OMA, when the delay is loose.

When the delay exponent becomes more stringent, then the total link-layer rate of OMA is better compared to the two-user NOMA network. Fig. 4.6 shows the plot of the total link-layer rate of the two-user NOMA and OMA network with short-packet communications with $\theta = 0.01$. Under different fading models, two-user OMA shows better performance compared to the two-user NOMA network. More specifically, the performance gap between the two-users NOMA

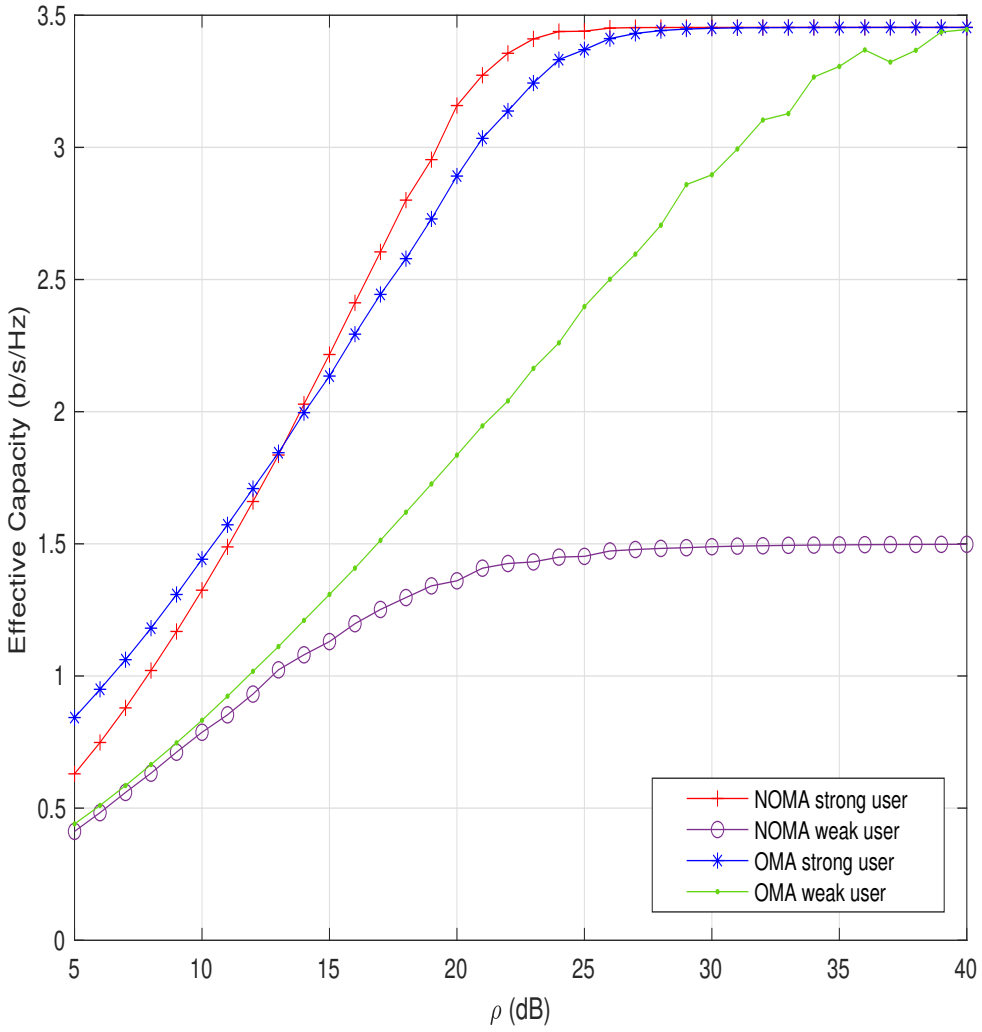

Figure 4.4: Achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under Rician and lognormal fading with stringent delay, i.e., $\theta = 0.01$.

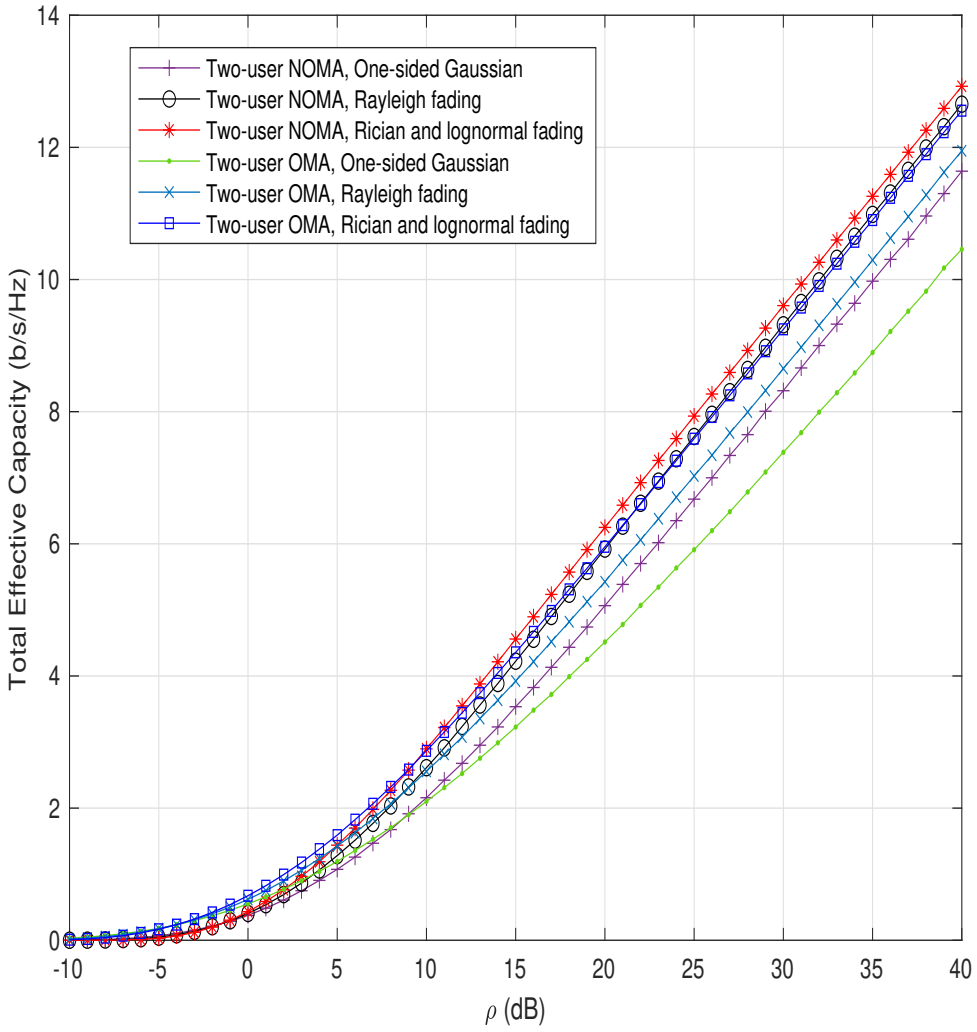

Figure 4.5: Total achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under different fading conditions with less stringent delay, i.e., $\theta = 0.001$.

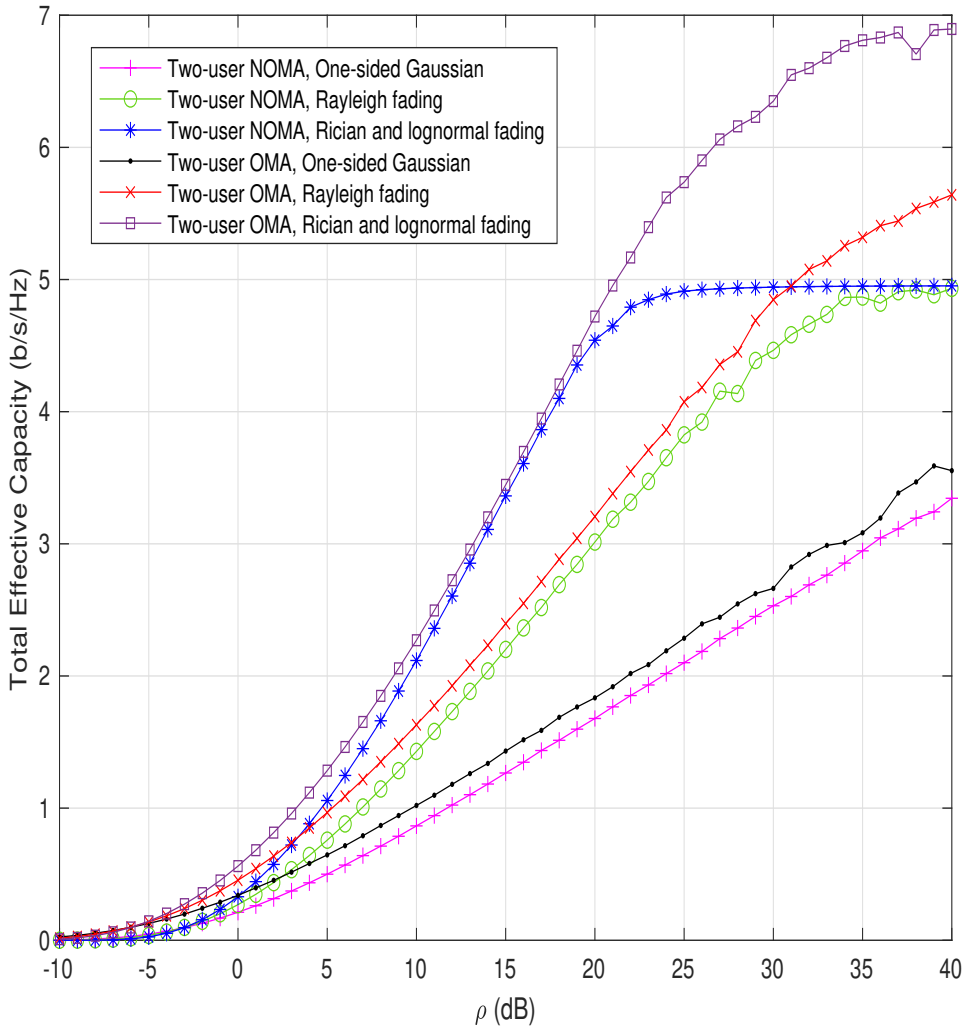

Figure 4.6: Total achievable EC of two-user NOMA and two-user OMA versus the transmit SNR under different fading conditions with stringent delay, i.e., $\theta = 0.01$.

and OMA total link-layer rate is small at the small transmit SNR, however as the transmit SNR increases, the performance gap becomes widen, and two-user OMA overcomes the stringent delay constraint.

Fig. 4.7 shows the plots of the total achievable EC versus the transmit SNR (ρ). The results reveal that the total achievable rate of NOMA outperforms the one for OMA at high SNRs when $\theta = 0.001$. On the other hand, when the delay exponent becomes stringent, i.e., changes from $\theta \to 0.001$ to $\theta \to 0.01$, the total link-layer rate of OMA outperforms the one of NOMA at high SNRs. However, at the low SNRs, the total link-layer rate of NOMA and OMA are approximately the same irrespective of the delay constraints.

Fig. 4.8 provides a comparative view of the total achievable EC of multiple NOMA and OMA pairs when service is provided to 6 users out of 12 possible users, and $\theta = [0.001, 0.01]$. Within a pair, NOMA scheme has been implemented, while the inter-pair multiple access has been achieved using TDMA. It is clear from the simulation results that the multiple-user NOMA network outperforms the OMA one under different delay requirements. The figure also reveals that multiple-user NOMA and OMA perform better than the two-user access cases when the delay exponent becomes stringent. It is also important to note that, these findings are based on the Monte-carlo simulations. These Monte-carlo simulations have already been verified with the closed-form expressions.

Fig. 4.9 plots the simulation results of individual user's achievable EC of two-user NOMA and OMA versus the delay exponent θ when the transmit SNR $\rho =$ 20dB. This figure shows that the NOMA users outperform the OMA users when the delay exponent is very loose. However, when the delay exponent becomes stringent, the NOMA users show a considerable loss in EC as compared to the OMA case.

Finally, Fig. 4.10 shows the curves of the achievable EC of two-user NOMA versus the transmit SNR (ρ) with different values of the power coefficients (α_1,α_2), when $\theta = 0.01$. Compared to the flexible power allocation scheme, this figure shows how the different sets of fixed power coefficients impact the performance of the two-user NOMA network. With the increase in the power coefficients, the achievable EC of both the strong user and the weak user with NOMA also increases. This also confirms that the changing power coefficients has also a

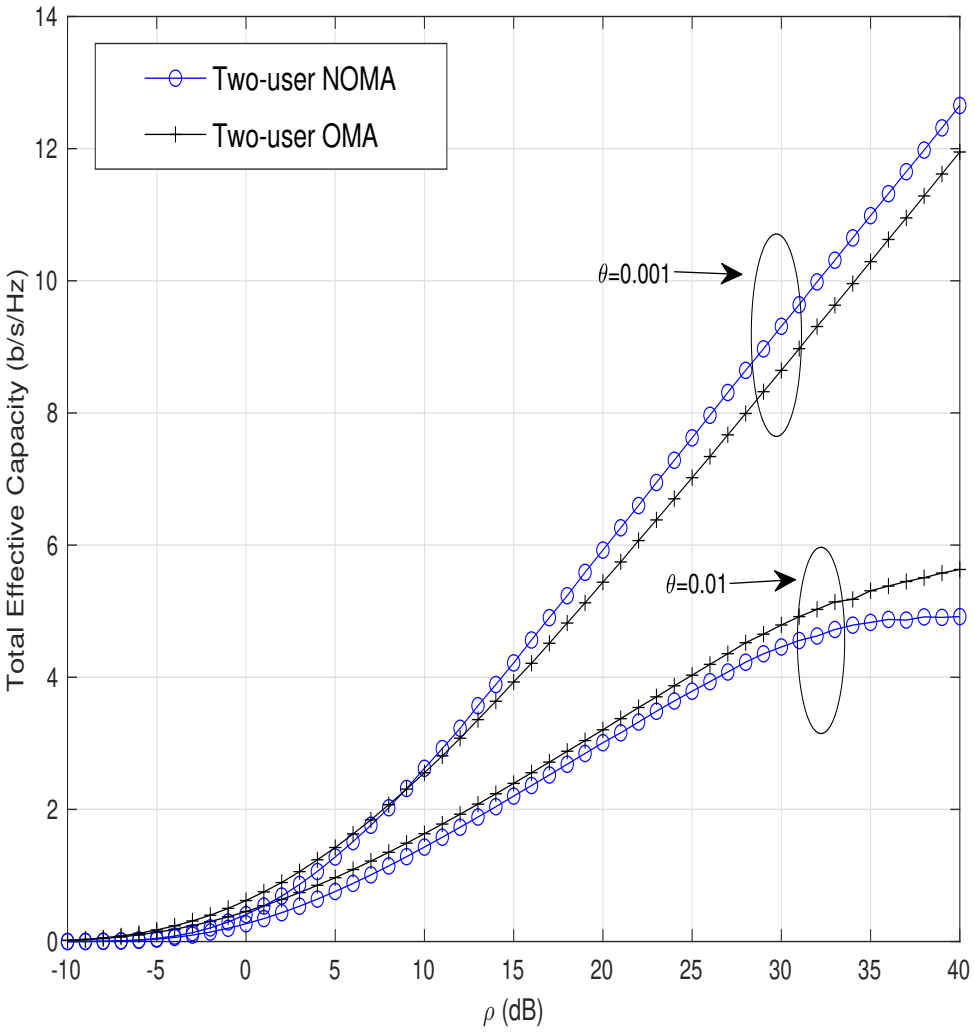

Figure 4.7: Total achievable EC of two-user NOMA and two-user OMA versus the transmit SNR.

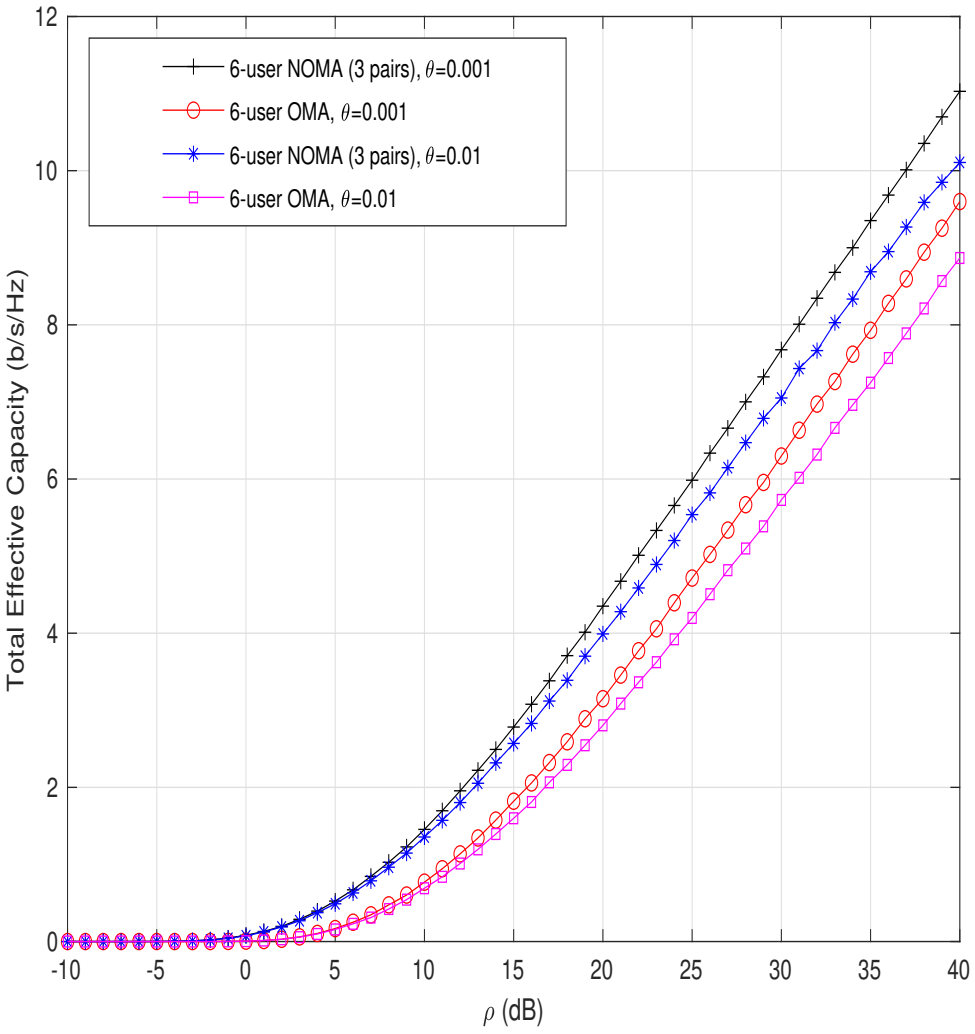

Figure 4.8: Total achievable EC of multiple NOMA pairs and multiple OMA users versus the transmit SNR with 6 users out of 12 users.

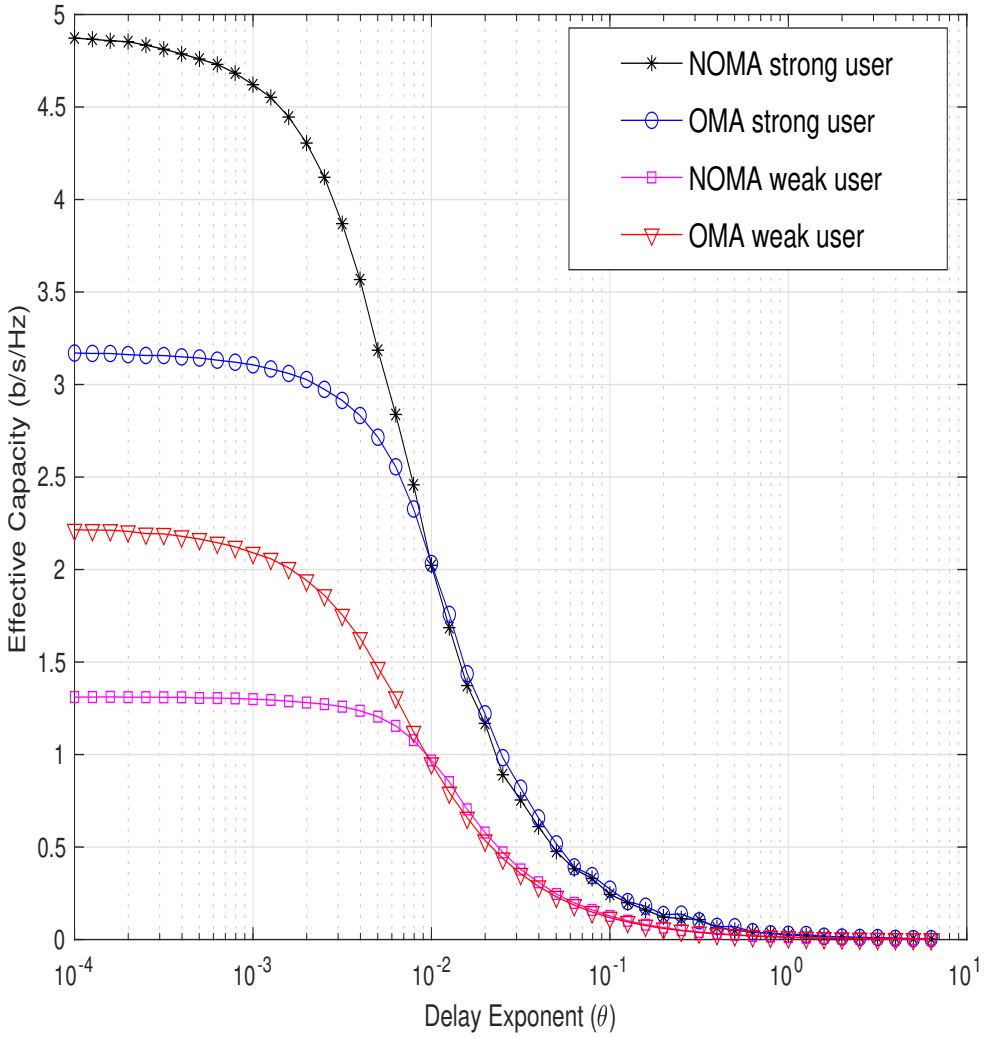

Figure 4.9: Achievable EC of two-user NOMA and two-user OMA versus delay exponent (θ).

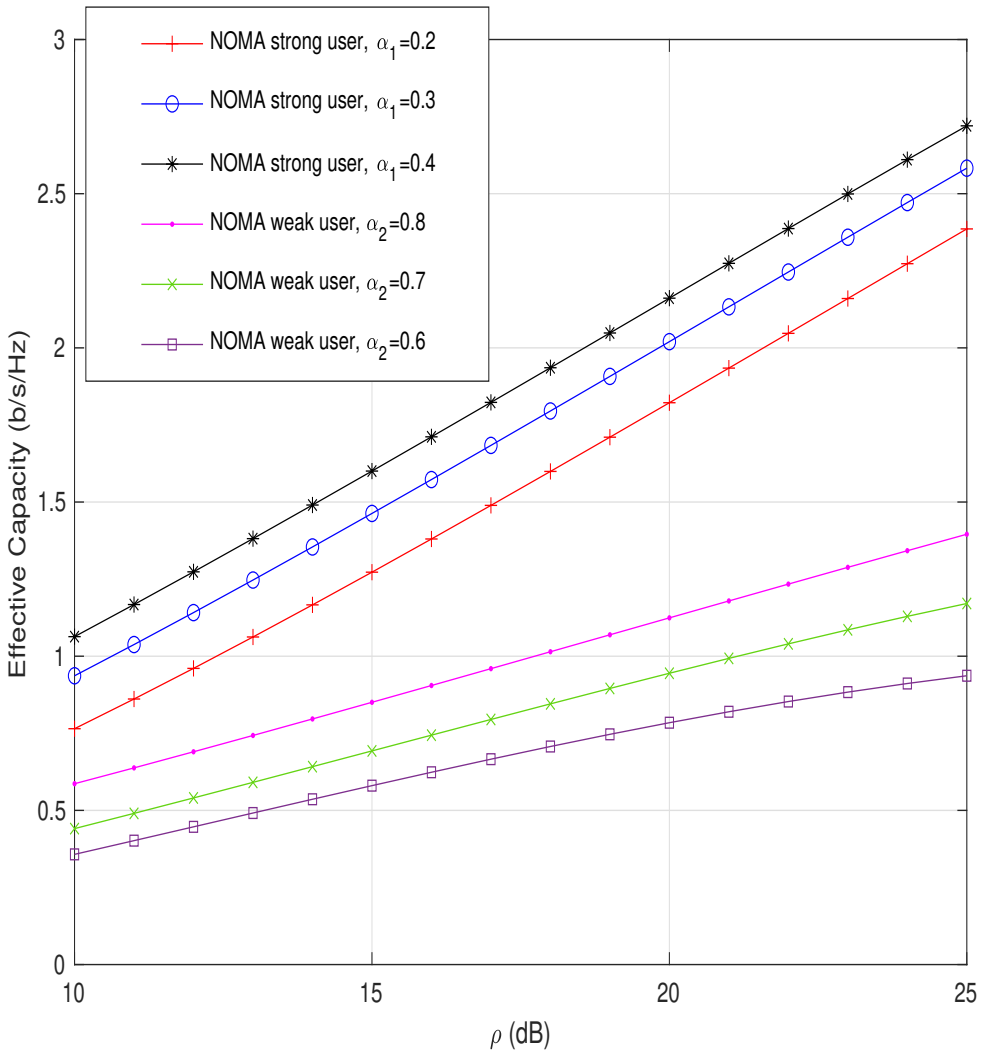

Figure 4.10: Achievable EC of two-user NOMA and two-user OMA versus delay exponent (θ).

significant impact on the two-users NOMA netowrk under delay constraint.

4.5 Summary

In this chapter, considering NOMA and OMA with two users, the individual user's achievable EC in finite blocklength regime was formulated. A closed-form expressions for the individual EC of both users, in NOMA and OMA separately under Rayleigh fading was dervied, and their accuracy was confirmed using Monte-Carlo simulations. Performance of NOMA in comparison with OMA under heterogeneous delay QoS constraints was also investigated. The performance comparison showed that at low SNRs the strong user OMA outperformed both NOMA users with Rayleigh fading. The achievable EC of the two-user NOMA and OMA under different fading models was also investigated in detail. This analysis established that two-user OMA (with individual EC of two-user OMA) under severe fading and stringent delay outperformed the two-user NOMA. However, when the total EC of two-user NOMA and OMA was compared, then it was the NOMA that performed better compared to its OMA counterpart.

Chapter 5

Short-Packet Assisted Non-orthogonal Multiple Access Based Random Access

5.1 Introduction

NOMA in conjunction with the short-packet communications is regarded as the enabler of the uRLLC. In power-domain NOMA, multiple users share the same resource block, while exploiting the power difference of their signals. Mainly, NOMA is used for the downlink transmission and this transmission is termed as the coordinated transmission. This coordination is achieved through BS. However, to further reduce the signalling overhead, NOMA can also be used for non-coordinated transmission random access for accessing the channel. NOMA short-packet communications with the conventional random access scheme, i.e, multichannel ALOHA can be employed to achieve the massive connectivity, spectrum efficiency, and low latency communications. NOMA-random access (RA) in conjunction with the short-packet communications is the non-coordinated transmission, that will play a critical role in overcoming the spectrum scarcity [20].

To provide the seamless connectivity for billions of machine-to-machine (M2M) connections, advance multiple access techniques with less access delay are required [35, 124]. These M2M connections also operate with the short data payload. For

this purpose, the NOMA-RA with short-packet communications have been envisioned as an enabler technology for the challenging use cases of the 5G, mMTC, and uRLLC.

With the help of NOMA-RA, the users will randomly select one of the channel and one predefined power level and start transmission. Then the receiver performs the SIC, and decode the received message, while considering the received signal strength. Considering the importance of the NOMA-RA for improving the spectrum efficiently, researchers from industry and academia has explored its different dimensions. NOMA-RA in uRLLC energy harvesting networks was investigated in [125]. In this work, the goodput (goodput is the actual useful information transmitted over the network and received by the receiver. This is considered as the application level throughput of the network), reliability, and average packet latency were studied for the NOMA-RA. The analytical results were compared with the OMA-RA and showed the improved performance of the NOMA-RA in terms of goodput, reliability, and average packet latency. Employing the random access technique with the NOMA is not only restricted to the power domain NOMA. Authors in [126], employed the code domain NOMA (CD-NOMA) (there are two basic types of NOMA, namely, code domain NOMA and the power domain NOMA. In code domain NOMA, the users are superimposed on the resource block using the Gaussian codes or spreading non-orthogonal codes. While in power domain NOMA, the power levels are used by different users to exploit the NOMA operation.) with the random access technique and showed that the throughput of the proposed CD-NOMA with random access feature was improved compared to the conventional random access scenarios.

NOMA-RA has a direct impact on the industry, due to this factor, researchers from academia are investigating the NOMA-RA from all dimensions and for all application scenarios. One of the real-world application is the unmanned aerial vehicle (UAV). For this application, NOMA-RA was investigated in [36]. In this work, the focus was to achieve the stable throughput for UAVs by employing the NOMA-RA and the optimal power level selection. Further, the stable throughput was achieved for the UAVs while considering the altitude and beam width of the UAVs.

Due to the practical applications of the NOMA-RA, NOMA-RA was regarded

as the enabler of the uRLLC use case of 5G and also for the 6G. In [35], NOMA-RA was investigated as the multiple access technique of choice for the 6G IoT networks. In this work, NOMA-RA which was based on the slotted ALOHA technique was modified using the spreading non-orthogonal code (the code domain NOMA used the Gaussian codes or spreading non-orthogonal codes to make the distinction among different users while using the one resource block) , and was termed as the spreading slotted ALOHA based NOMA-RA. A completely different approach compared to the conventional approach for reducing collision in NOMA-RA was adopted based on the spreading non-orthogonal code. This proposed approach increased the throughput and hence deemed suitable for the challenging scenarios of 6G IoT networks. NOMA-RA is regarded as of capital importance and as a core technology for the emerging IoT networks. NOMA-RA used the combined operation of the NOMA and random access techniques such as ALOHA and mult-channel ALOHA. As in the future, millions of the IoT devices will require a short transmission delay and higher reliability for their operation. Therefore, the NOMA-RA due to the short access delay is regarded as the key multiple access technology for the future IoT devices. Also, the transition from the existing multiple access techniques (ALOHA) to the NOMA-RA is less complex. The NOMA-RA based on the slotted ALOHA for the emerging IoT networks was discussed in [20]. To minimize the interference two detection technique was employed SIC with optimal decoding and the joint decoding. Then the outage probability of the SIC and the joint decoding was derived and investigated. The combined approach enhanced the throughput of the NOMA-RA.

NOMA-RA due to the provision of short access delay, massive connectivity, and higher throughput, has attracted a lot of attention in todays research domain. For example, in [127], to avoid the collision in case when multiple packets are transmitted simultaneously using the same assigned slot, NOMA-RA scheme was proposed with collision avoidance mechanism. Usually the collision resulted into the reduction of the throughput and overall reduction in the performance of the NOMA-RA. This work [127] considered the collision resolution period during which the collided packets are transmitted. In this way, the NOMA-RA achieved the maximum throughput under the controlled conditions. NOMA-RA was also investigated with its data transmission scheme under the umbrella of NOMA-RA

data transmission ((NOMA-RA DT) in [128]. In this work, NOMA-RA DT users considered the advance design of preamble selection based on the collision. BS in advance could detect the preamble with possible collision and only assign the physical uplink shared channel (PUSCH) to the NOMA-RA DT users without collision.

The most recent work of the NOMA random access that considers the access probability and conditional throughput is very limited. The throughput of the NOMA-RA for users randomly accessing the channel was investigated in [33]. This NOMA-RA was based on the multi-channel ALOHA scheme and showed the proposed NOMA-RA improved the spectrum efficiency compared to the conventional random access technique. However, this work did not consider the short-packet communications. How the optimal power levels should be selected in NOMA-RA, to increase the throughput? This question was answered in [129], where the proposed users in NOMA-RA scheme optimally select the power levels from the pre-determined power levels and achieved the higher throughput. This work not only considered the optimal power level selection, but also focused on the collision avoidance, so that different users did not select the same power level. Another work [130] also considered the NOMA-RA as the core technology for the IoT networks and investigated the core network performance metrics such latency and the GoodPut. The impact of the packet replicas was studied on the network metrics, where the different replicas were accommodated in different time slots. The proposed scheme showed the improved throughput in terms of the maximum GoodPut.

The throughput of the NOMA-RA was also studied in [131]. In this work, the throughput of the NOMA-RA was compared with the multichannel ALOHA with and without capture effect. Also, the sensitivity of the given load on the throughput of NOMA-RA was calculated. Then an adaptive load distribution scheme was developed for this proposed NOMA-RA scheme via user barring algorithm. Still this detailed analysis is without taking into consideration the short-packet communication. However, this chapter not only considers the NOMA-RA but also the short-packet communications, and shows that how the BS assigns the resources that will facilitate the NOMA-RA short-packet communications.

The main contributions of this chapter are summarized as follows:

- NOMA-RA with short-packet communications based on the multichannel ALOHA has been proposed and its performance in terms of conditional throughput is investigated with respect to different number of power levels and subchannels.
- Proposed NOMA-RA with short-packet communications outperformed the conventional random access techniques such as multichannel ALOHA.
- Through analytical results, it has been shown that the conditional throughput of NOMA-RA with short-packet communications improved with adding more power levels, however it saturates at a very high value. Which establish that adding more power levels adds complexity and further addition in power levels will not translate into throughput improvement.
- It is also shown that the short access error probability (due to short-packet communications) has a impact on the throughput of the NOMA-RA. When this error probability becomes stringent, the throughput of the NOMA-RA reduces.
- Users are randomly accessing the channels based on some access probability. Increasing the access probability should increase the throughput. However, this is not the case for NOMA-RA and multi-channel ALOHA. There is the optimum value for this access probability. Through analytical results, it is clear that with further increase in the access probability the throughput of the proposed NOMA-RA short-packet communications increases first and then starts decreasing afterwards. The optimum value of the access probability has also been highlighted through simulation results.

Based on the multi-channel ALOHA, the NOMA-RA with short-packet communications is designed in this chapter. The rest of the chapter is organised as follows: The Section 5.2 provides the system model and the transmission framework for multi-channel ALOHA and the NOMA-RA, based on which the conditional throughput of the NOMA-RA is derived. The performance of proposed NOMA-RA with short-packet communications is then evaluated in Section 5.3, then the whole chapter is summarised in 5.4.

5.2 System Model and Transmission Framework

This system model considers the uplink transmission with K orthogonal subchannels, one BS, and D multiple users. To complete the uplink transmission, first the random access procedure for the grant of uplink resources is performed between the BS and the user equipment (UE). This four step handshake procedure for NOMA-RA short-packet communication is based on message passing between the UE and BS. Figure 5.1 shows the four stages of handshake procedure for the NOMA-RA for accessing the uplink resources. The first stage is the *preamble transmission* which involves the reception of system information from the BS on physical broadcast channel (PBCH) and then selection and transmission of preamble to BS via physical random access channel (PRACH). Second step is the *preamble detection and RAR transmission*, through which UE receives the random access response via physical downlink control channel (PDCCH) and physical downlink shared channel (PDSCH) from the BS. In this step, UE receives the uplink grant and synchronization information. In the third step, UEs with the recognized preambles will start *layer 3 message transmission* for the connection request using the physical uplink shared channel for short-packet transmission (PUSCH-SPT). As compared to the NOMA-RA, where the uplink channel used for layer 3 message is the PUSCH, the BS assigns the PUSCH-SPT for the NOMA-RA short-packet communications. This enables the devices to transmit using short data payload with less access delay. The fourth step for handshake procedure is the *contention resolution*. If multiple users select the same preamble, then the collision occurs. The collided users will then go into the fourth step bypassing the layer 3 message transmission step.

As the NOMA-RA is based on the multichannel (slotted) ALOHA, therefore a very brief discussion about the multichannel ALOHA is given here. At any given time, if the A_j set of multiple users actively and randomly selects the jth subchannel for transmission, then the received signal y_j at the BS can be approximated as $y_j = \sum_{a \in A_j} h_{j,a}\sqrt{P_{j,a}}s_{j,a} + k_j$, where $h_{j,a}$, $P_{j,a}$, and $s_{j,a}$ are the channel coefficients, transmit power and the signal from user a with k_j as the spectral noise. There is possibility of the collision, when multiple users access the same subchannel. However, in this system model, the collision probability is not

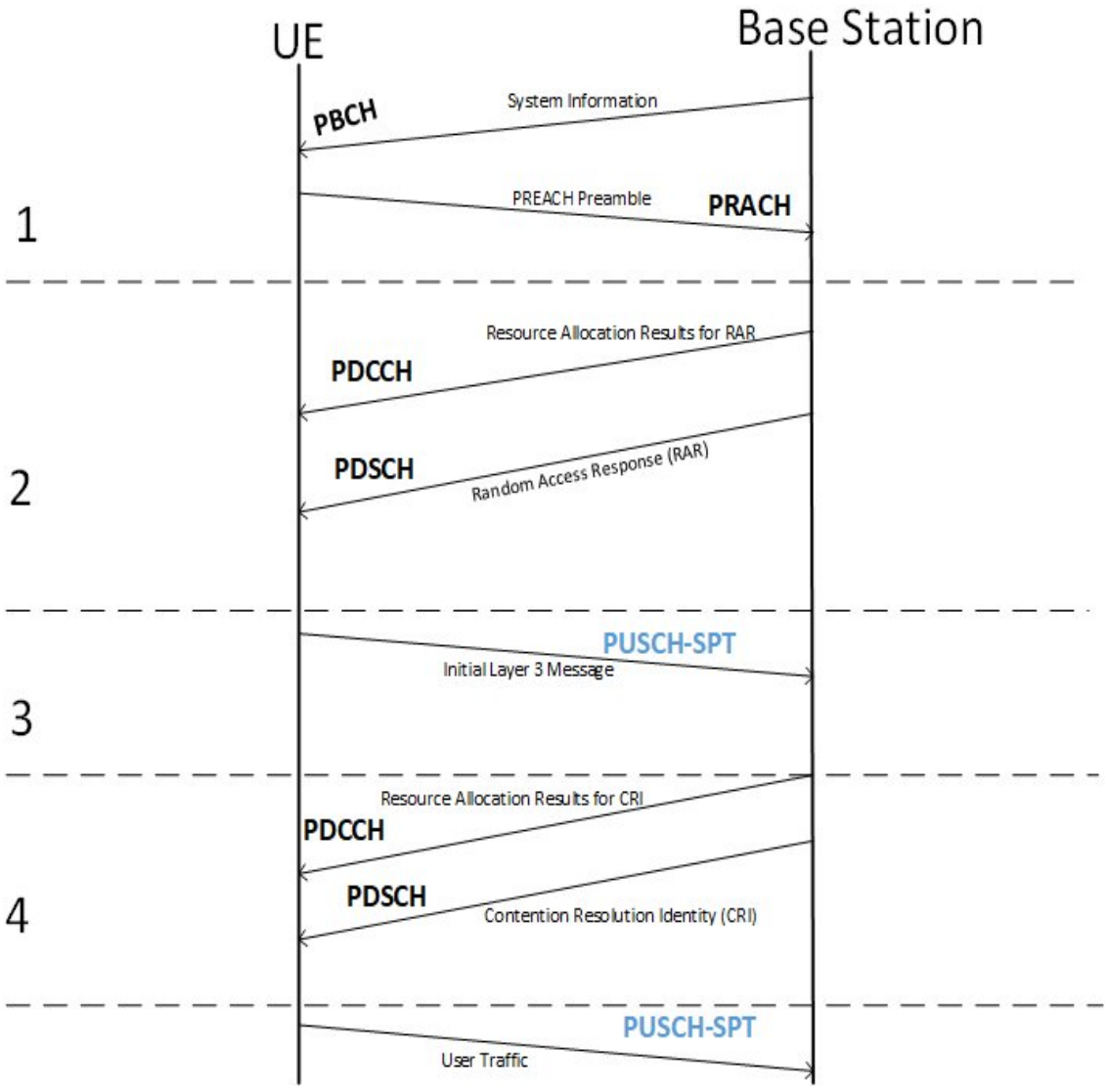

Figure 5.1: NOMA-RA Basic operation with the short-packet communications assisted PUSCH-SPT from BS.

taken into account and a very simple collision model is adopted. Following the [131], the throughput of the multichannel ALOHA is approximated as,

$$P_{AL}(D;K) = D\Big(1-\frac{1}{K}\Big)^{D-1}, \tag{5.1}$$

where active users D randomly select the K subchannels. By finding the distribution of the D, the average throughput of the multi-channel ALOHA can be derived. If there are S users and each user becomes active with some access probability, a_p. Denoting the number of active users that select the subchannel with M, the approximation of the D users can be, $\mathbb{E}[D] = Sa_p$, and $\mathbb{E}[M] = Sa_p$. Considering the M as the Poisson random variable with parameter λ, and $p_\lambda(m)$ as the probability mass function (pmf), then the Poisson random variable M can be written as,

$$M \sim p_\lambda(m) = \frac{e^{-\lambda}\lambda^m}{m!}. \tag{5.2}$$

For a very large number of users S, the average throughput of the multichannel ALOHA can be derived as,

$$\begin{aligned} U_{AL}(K) &= \mathbb{E}[P_{AL}(D;K)]. \\ &= K\lambda e^{-\lambda} \end{aligned} \tag{5.3}$$

During the orthogonal random access grant procedure, if the assigned uplink resource from the BS is PUSCH-SPT as is explained above in Fig. (5.1), then the average throughput of the multichannel ALOHA for short-packet communications links can be approximated as

$$\begin{aligned} U_{AL-S}(K) &= \mathbb{E}[P_{AL-S}(D;K)](1-S_\epsilon) \\ &= K\lambda e^{-\lambda}(1-S_\epsilon), \end{aligned} \tag{5.4}$$

where S_ϵ is the short access error probability due to the grant of the PUSCH-SPT resource while employing the short-packet communications. The derived throughput of multi-channel ALOAH with short-packet communications as is

given in (5.4) can be used to investigate the multi-channel ALOHA short-packet communications performance in conventional networks. This expression shows that the throughput of the multi-channel ALOHA is K times higher than that of the single channel random access scheme. As, inserting the $K = 1$, the (5.4) reduces to the single channel ALOHA scheme with short-packet communications. The above mentioned throughput however can be maximised by adjusting the intensity of the λ in (5.4). After using the $\lambda = 1$ in (5.4), the maximum throughput of the multi-channel ALOHA with short-packet communications can be achieved.

Based on the above random access scheme, the NOMA uplink random access scheme for the short-packet communications regime is considered. This NOMA random access scheme is entirely different from the conventional NOMA approach where the whole scheme is coordinated by the BS. Here, the pilot signal transmitted by the BS to the users is used to synchronize the whole communications. This help the users to estimate the channel state information (CSI). Taking into consideration the transmitter design simplicity, the CSI estimation is taken as perfect. If there are P_L predetermined power levels, then the active users have the flexibility to select one of the power level. The predetermined power levels are $c_1 > ...c_{P_L} > 0$. To perform the random access procedure, the user v_i selects one of the power level, c_l . After selecting the power level c_l , the transmission power can be approximated as

$$\rho_i = \frac{c_l}{\alpha_i}, \tag{5.5}$$

where α_i is the channel gain for user v_i, ρ_i is the transmitted power, and ultimately c_l becomes the received signal power at the BS. So, the c_l can be the SNR or SINR at the BS. Depending on the target SINR γ, the power level c_l can be decided via $c_l = \gamma(C_l + 1)$, where $C_l = \sum_{m=l+l}^{P_L} c_m$. Desired transmission rate or channel conditions are taken into consideration while deciding the target SINR. Therefore, the target SINR for one of the active user selecting the c_l power level can be derived as

$$\gamma = \frac{c_l}{C_l + 1}. \tag{5.6}$$

The signal from the user with power level c_l is then decoded and removed at the BS using the SIC. If multiple users are involved in the transmission using the P_L power levels, then the total of P_L signals are decoded using the SIC. There is

a possibility that, some users can choose the same power levels. In this case, the power collision occurs, which result in the decoding error and the BS do not able to decode some of the signals. Considering the power domain NOMA random access procedure, if the D active users randomly selects the power levels such that $D \leq P_L$, then the probability of selecting the different power levels by the active users is

$$D = \prod_{d=1}^{m-1} \left(1 - \frac{d}{P_L}\right). \tag{5.7}$$

By employing the above probability and keeping the subchannels constant, the average conditional throughput of the NOMA random access procedure using the PUSCH-SPT resource can be found using the (5.2), (5.3), and (5.4) as

$$U_{NR}(P_L, K) = K \sum_{m=1}^{P_L} m \left(\prod_{d=1}^{m-1} \left(1 - \frac{d}{P_L}\right) \right) \frac{e^{-\lambda} \lambda^m}{m!} (1 - S\epsilon), \tag{5.8}$$

where the K are the number of the subchannels and have been approximated using the Poisson approximation. It is clear that as the number of the power levels increase the throughput of the NOMA random access increases, without addition of the subchannels. This confirms the superior performance of the NOMA random access as compared to the multichannel ALOHA. However, taking the $P_L = 1$ in the above expression reduces the conditional throughput of the NOMA random access to multi channel ALOHA. While taking the $P_L = 2$ results into the higher throughput (approximately 1.5 times) than the conventional multi-channel ALOHA scheme.

However, the performance of the NOMA-RA with extensively larger number of the power levels will not translate into the maximum throughput. This is due to the fact that increasing number of power levels will also introduce complexity, and the BS will not perform a perfect SIC and the resulting decoding errors will significantly reduce the overall gain of the proposed scheme.

In addition to the above limitation of NOMA-RA with extensively large number of the power levels, there is also a concern that when the multiple users selects

the same power levels as is mentioned earlier. If the multiple users selects the same power levels then the power collision occurs and the BS fails to decode the rest of the power levels. As compared to the collision that occurs in the conventional multi-channel random access scheme, the power collision in the NOMA-RA scheme is not an independent event. It means, if in one transmission, power collision occurs, then the BS will not be able to decode the rest of the power levels and hence messages of the other uses will also not be decoded. In the event of the power collision, it is still possible to decode some of the signal, however all the rest of the signals from the power level facing the power collision will be dropped. This is serious challenge in designing the NOMA-RA schemes, that can significantly reduce the performance of the NOMA-RA. Hence, the researchers from the academia and industry are focusing on the power collision aspect of the NOMA-RA. Also, the above derived throughput is for the uplink NOMA-RA with short-packet communications where the users takes their own decision regarding the selection of the power levels. Therefore, this non-coordinated transmission as compared to the NOMA transmission also adds complexity in more complex and diverse network environment.

The final conditional throughput of NOMA-RA derived in (5.8) can be used to investigate the NOMA-RA for the potential users accessing the users through selection of the different power levels from the pre-determined power levels. This expression is also used to investigate the impact of the access probability, short-access error probability, and power levels on the performance of the proposed NOMA-RA and multi-channel ALOHA scheme. This work mainly focus on the theoretical conditional throughput of the NOMA-RA that is entirely based on the probabilistic analysis.

5.3 Performance Evaluation

In this section, the performance of the NOMA-RA with short-packet communication is extensively evaluated. The simulation parameters are set as, $S_\epsilon = 10^{-5}$, K=250, and access probability is set to 0.06, unless otherwise specified.

Figure 5.2 shows the throughput of the multiple access techniques (MS-ALOHA and NOMA-RA) with short-packet communication versus the number of sub-

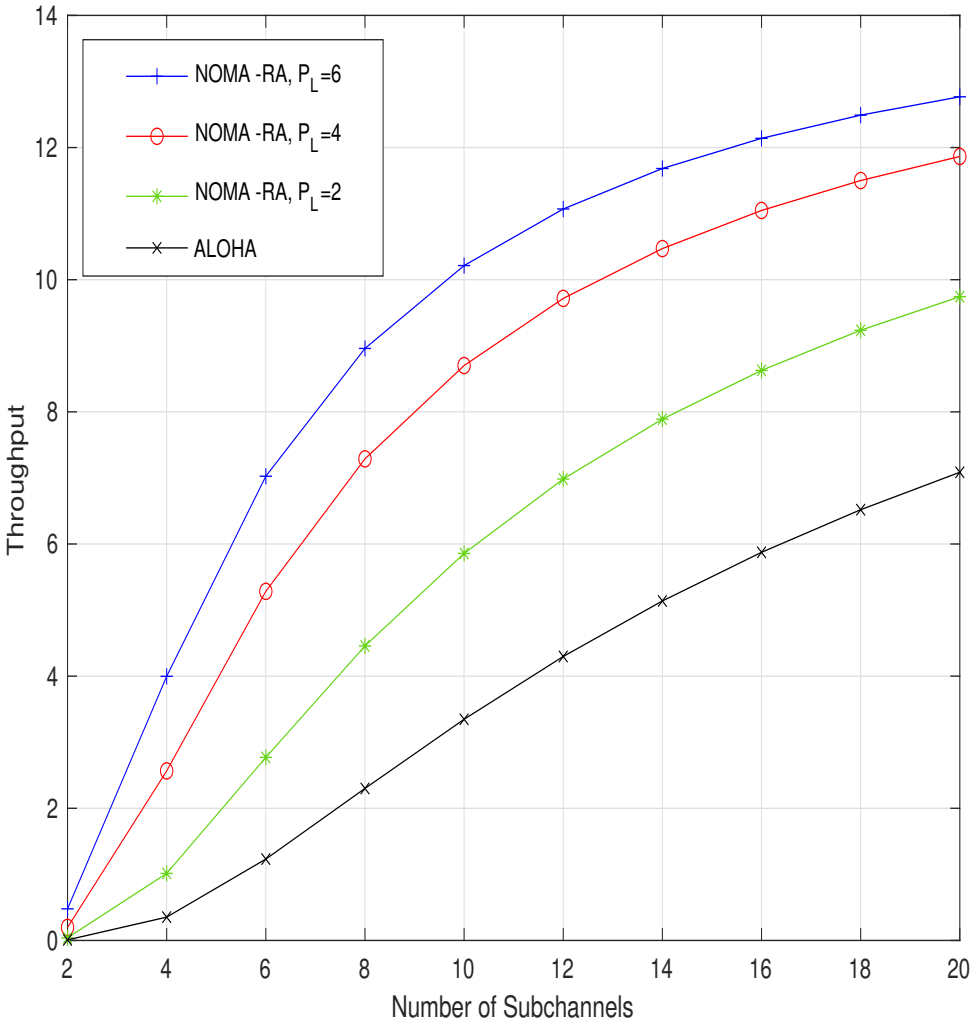

Figure 5.2: Conditional throughput of the NOMA-RA and multi-channel ALOHA short-packet communications versus subchannels.

channel, while the $S_\epsilon = 10^{-5}$, K=250, and access probability is set to 0.06. It is clear that the throughput of the MS-ALOHA and NOMA-RA with short- packet communications increases with the increase of the number of sub-channels. However, the NOMA-RA outperforms the MS-ALOHA and shows improvement in throughput. The performance gap between the MS-ALOHA and NOMA-RA is wide at the lower number of sub-channels, i.e, when the number of subchannels are limited. However, when the more bandwidth resources are added this performance gap become short. This is due to the addition of more complexity due to the increased number of subchannels.

Figure 5.3 shows the throughput of the NOMA-RA versus number of power levels, while keeping the subchannels constant at 8, and access probability set to 0.06. These results further confirm the enhanced performance of the NOMA-RA with short-packet communications. Without addition of further bandwidth resources (subchannels), NOMA-RA throughput increases with the addition of more power levels. However, this performance decreases when the short access error probability (due to short-packet communication) becomes more stringent. This simulation result also shows that the increase in the performance is not consistent with the increase in the power levels. At some point, further increase in the throughout do not occur. This is due to the power collision and complexity because of the addition of the large number of power levels

User are randomly accessing the subchannels based on some access probability. The increase in the access probability should increase the overall performance of the MS-ALOHA and NORA-RA. However, this is not the case as it requires the further investigation. For this purpose, simulation results of Figure 5.4 shows the throughput of the NOMA-RA short-packet communications versus power levels while considering different values of the access probability with subchannels taken as 8. Proposed NOMA-RA shows the improvement in throughput with the access probability 0.06. However, further increase in the access probability results into decrease of the throughput initially and then is maximum at higher values of power levels. This shows that there exists a optimal value of the access probability to achieve a optimum performance of the NOMA-RA.

To further investigate the impact of the access probability on the performance of the NOMA-RA and MS-ALOHA, Figure 5.5 shows the simulation result of

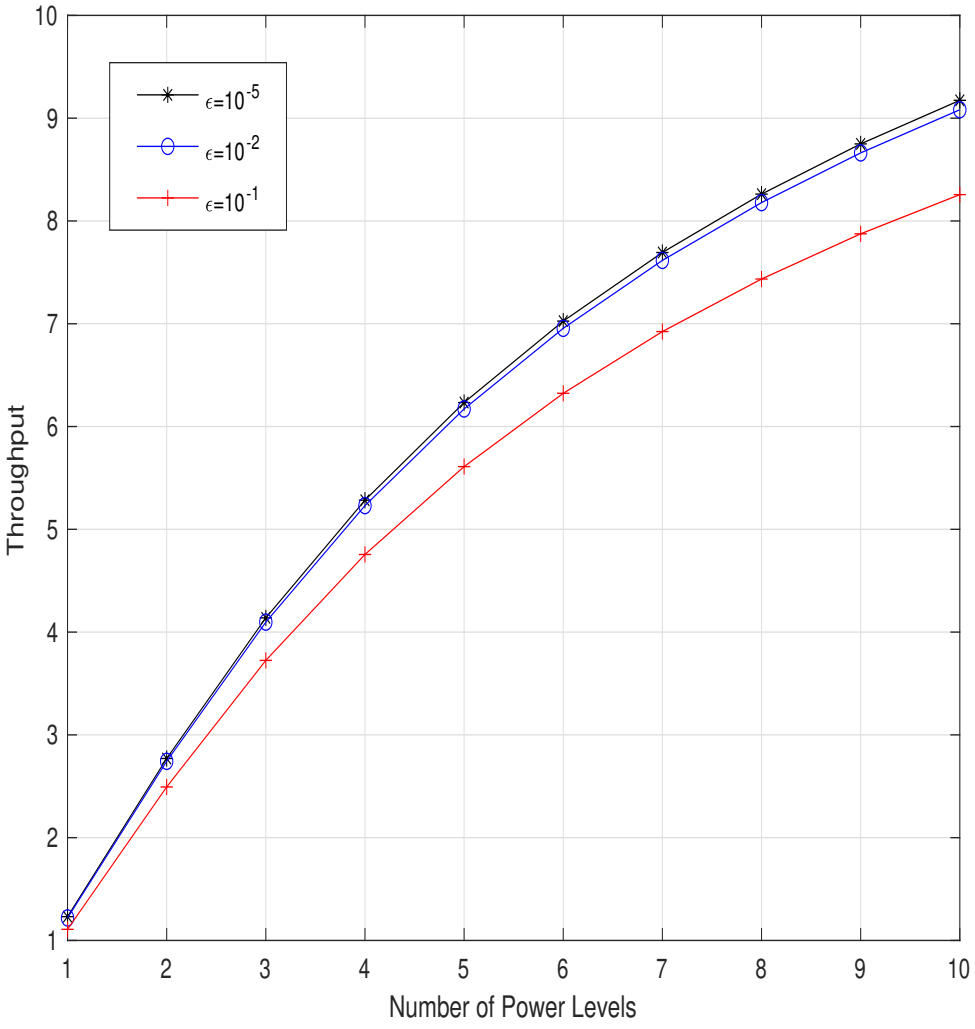

Figure 5.3: Conditional throughput of the NOMA-RA short-packet communications versus different power levels.

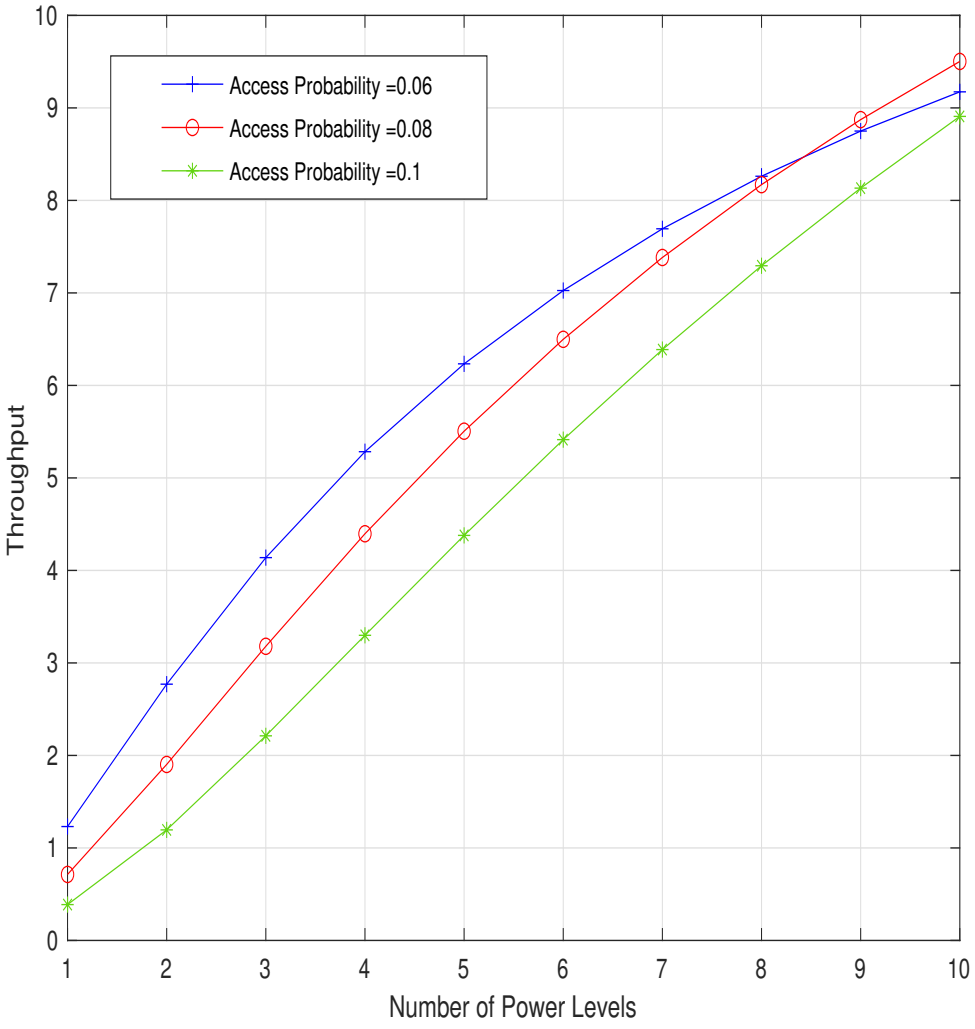

Figure 5.4: Conditional throughput of the NOMA-RA short-packet communications versus power levels, while considering different values of access probability.

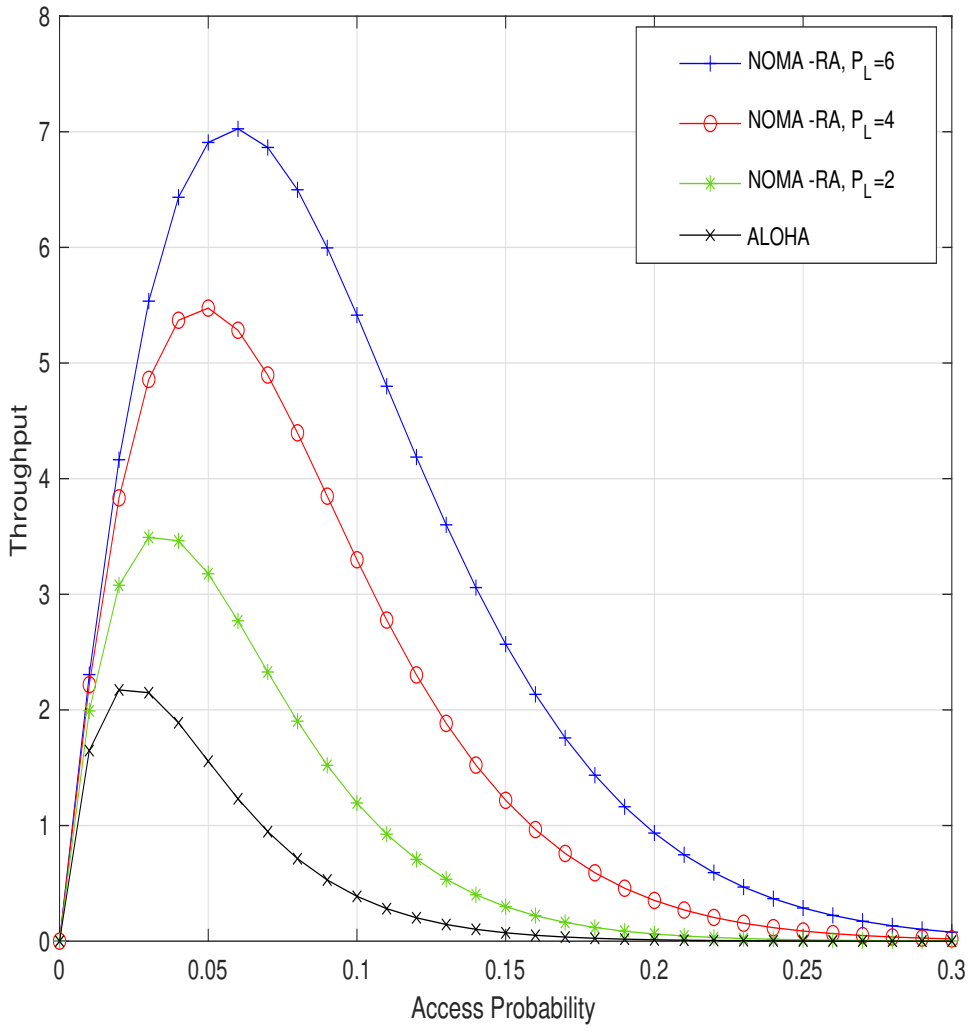

Figure 5.5: Conditional throughput of the NOMA-RA and multi-channel ALOHA short-packet communications versus access probability, with different values of power levels.

the throughput versus the access probability, with short access error probability set to 10^{-5} and number of subchannel taken as 8. It shows that there exists a optimum area of access probability to gain the maximum performance of NOMA-RA and MS-ALOHA. The increase in the number of power levels also increases the throughput, however still it requires the some access control mechanism to achieve a maximum performance. Usually, the increase in the power levels and the access probability should increase the throughput of the NOMA-RA. However, this is not the case in NOMA-RA. The increase in the access probability does not translate into the increase in the throughput. There is the optimum value of the access probability, from 0.05 to 0.1, where the NOMA-RA shows maximum performance in terms of the conditional throughput. Beyond this optimum range (area), further increase in the access probability does not increase the performance of the NOMA-RA. In addition to increase, the performance of the NOMA-RA decreases further.

Figure 5.6 shows the impact of the short-packet communications on the performance of the NOMA-RA technique, while keeping the power levels and subchannels constant at 6 each. This further establish that increase in the access probability does not translate into performance gain, but there exists a optimal performance area with the access probability. However, it also shows that the stringent short access error probability decreases the performance of the NOMA-RA.

Figure 5.7 shows the plots of the throughput of the NOMA-RA versus the short access error probability for different values of the power levels with, subchannels set to 6. Higher power levels shows the significant improvement in throughput as compared to the lower power levels while residing within the S_ϵ. However, when the S_ϵ requirements becomes stringent, the performance of the NOMA-RA decreases significantly irrespective of the power levels and subchannels. This drop in performance is due to the more errors that leads to the collision during the accessing of the bandwidth resources.

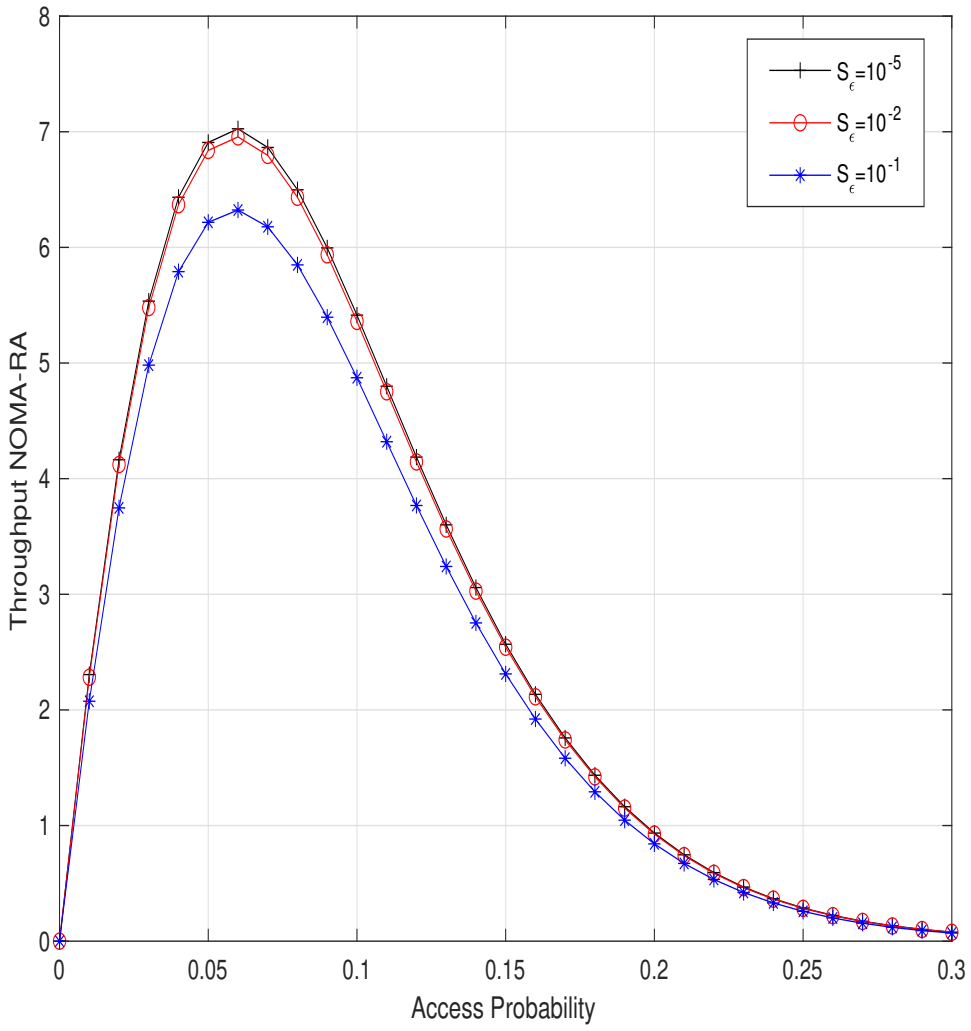

Figure 5.6: Conditional throughput of the NOMA-RA short-packet communications versus access probability, with different values of short access error probability.

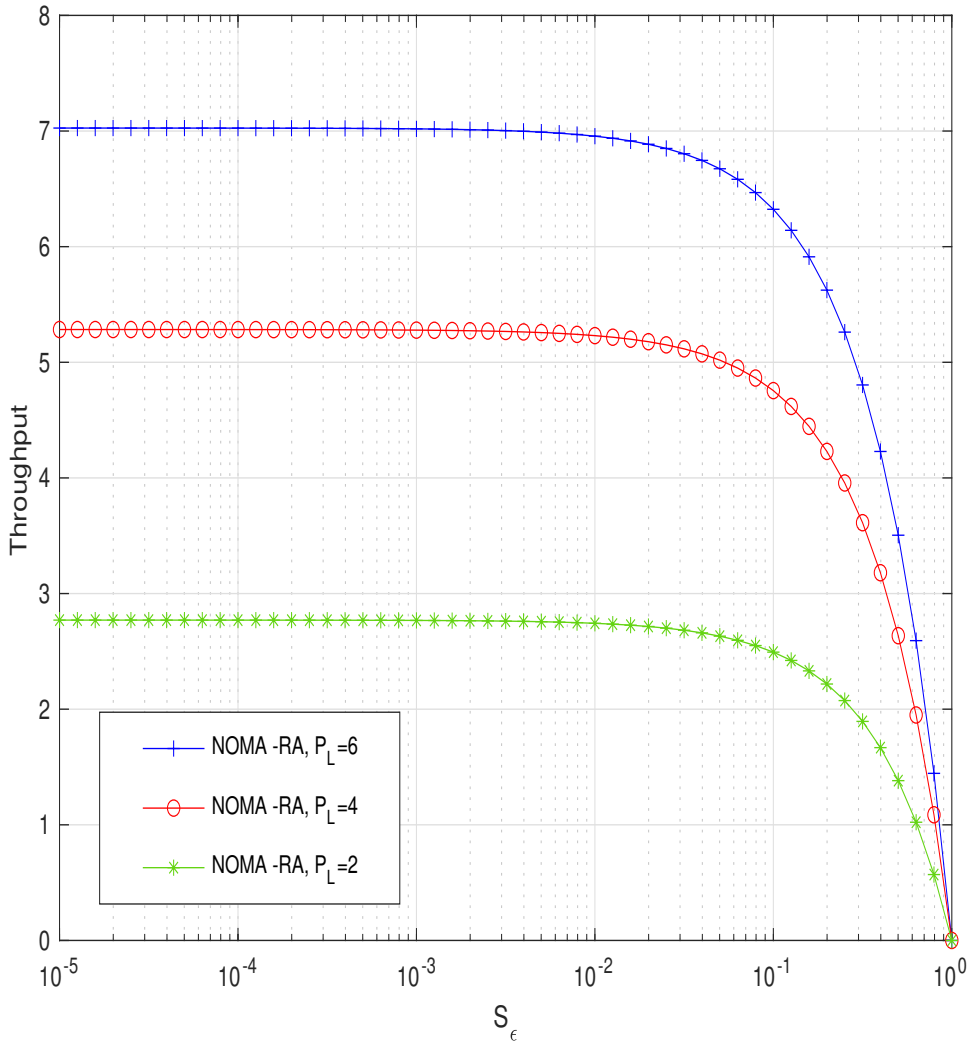

Figure 5.7: Conditional throughput of the NOMA-RA short-packet communications versus short access error probability.

5.4 Summary

In this chapter, one of the core enabler of the uRLLC, i.e., NOMA-RA with short-packet communications was investigated. In the conventional NOMA networks, the coordination is usually performed by the BS, through which BS allocates the power to the users based on their channel conditions. However, as compared to the conventional NOMA, users in NOMA-RA selects the power levels based on the pre-determined power levels. In this work, NOMA-RA was based on the multi-channel ALOHA, with power levels. In NOMA-RA with short-packet communications, the BS assigns the PUSCH-SPT which paved the way for the short-packet communications. It was clear from the simulation results that NOMA-RA with short-packet communications outperformed the conventional random access technique (multi-channel ALOHA). Access probability had a significant impact on the performance of the proposed NOMA-RA technique. The optimal area of the access probability from 0.05 to 0.1 showed the better performance of the NOMA-RA, and further increase in access probability beyond this area did not translate into enhanced performance of the NOMA-RA short-packet communications.

Chapter 6

Conclusions and Future Works

6.1 Conclusions and Discussions

The uRLLC is regarded as the challenging use case of the 5G and B5G wireless networks. Achieving the ultra low latency while residing within the constraint of the reliability requires the state-of-the-art enabling technologies. Researchers from the industry and academia have chalked out multiple enablers of the uRLLC [3]. This thesis mainly focused on the core enablers of the uRLLC, i.e, NOMA, NOMA-RA, and short-packet communications. The multiple access technologies in conjunction with the short-packet communications were investigated in detail in this work. The core point which differentiated this thesis from the existing works was the in-depth investigation of the multiple access technologies (NOMA and OMA) with short-packet communications while employing the frame work of the effective capacity. How the resources are accessed in case of the novel NOMA-RA short-packet communications, also made this thesis significantly different from the existing literature.

NOMA and NOMA-RA with finite blocklength improve the spectrum efficiency, reduce delay, increase the bandwidth. The EC framework was employed to investigate the delay performance of the NOMA and OMA with short-packet communications. EC is the dual concept of the effective bandwidth, which shows the maximum arrival rate that a given service process can support while satisfying the certain delay requirements. Finding the achievable EC of the NOMA short-packet communications is difficult as compared to the conventional NOMA

operations (without short-packet communications). In conventional wireless networks, NOMA operations did not take into consideration the reliability constraint and only considered the Shannon formula. However, in NOMA short-packet communications operations, the capacity penalty made it more complex to investigate with the EC formulation.

Considering the capacity penalty based on the transmission error probability, the achievable EC of the two users and multiple users NOMA network with short-packet communications were derived in this thesis. More specifically, the achievable EC of the two users NOMA (out of the multiple users) with short-packet communications were investigated while considering the transmit SNR, delay exponent, power coefficient, transmission error probability, and delay violation probability. Findings were very interesting such as, NOMA users with distinct channel conditions achieved higher EC as compared to the users with less distinct channel conditions. As mentioned above, investigating the NOMA with short-packet communication is more difficult and complex compared to the conventional NOMA networks. To address this challenge, the closed-form expression of the achievable EC of the NOMA networks with short-packet communications were derived and then verified using the Monte-carlo simulations. The total achievable EC of multiple NOMA pairs were also investigated by taking into consideration the impact of transmit SNR and delay exponent. It was clear that delay violation probability had a limit and could not be improved further under the constraint of transmission error probability and delay exponent. Most of the performance analysis of two users NOMA (out of the multiple users) and multiple users NOMA (multiple subsets) was performed under the Rayleigh fading channel. However, to provide the higher level insights into the impact of the fading on the achievable EC of NOMA users with short packets, Nakagami-m fading channel was also employed. Under this generalized fading model, it was shown that strong user's (user with good channel conditions) performance did not decrease significantly as compared to the weak user (user with weak channel conditions).

Achievable EC of the only one pair NOMA network was also derived and compared with its counterpart OMA with short-packet communications. As compared to the two-users NOMA (out of the multiple users), the single pair NOMA

has different closed-form expression because of the different PDF. This closed-form expression of the achievable EC of a single pair NOMA and OMA with short-packet communications was derived under Rayleigh fading channel and then verified using the Monte-Carlo simulations. The simulation results showed that the OMA short-packet-communications outperformed the NOMA at low SNR (20 dB). However, the analysis of total achievable EC of the two-users showed that the NOMA outperformed the OMA counterpart at high SNR (from 20dB to 40dB). This performance analysis was also performed with the generalized fading model, Nakagami-m for the NOMA and OMA short-packet communications. This analysis also confirmed that the strong user NOMA and OMA performed better under severe fading conditions.

NOMA with short-packet communications is the coordinated transmission where the BS performed the power allocation based on the user's channel conditions. However, to take the advantage of the less access delay of random access techniques, NOMA short-packet communications was also studied in non-coordinated transmission such as random access. For this purpose, in the last Chapter of this thesis, NOMA-RA that was based on the multi-channel ALOHA was proposed for the short-packet communications. Performance analysis showed that the NOMA-RA performed better as compared to the conventional random access technique. The performance of the NOMA-RA enhanced by the addition of the power levels, without adding more sub-channels. Finally, the access probability for accessing the channel played an important role in the whole operation of the NOMA-RA. There existed an optimal area of access probability from 0.05 to 0.1, where the NOMA-RA showed the maximum performance.

The core enablers of the uRLLC, NOMA, NOMA-RA, and short-packet communications were investigated in more detail in this thesis. It is clear that, not only the one technique or transmission scenario was sufficient to achieve the stringent requirements of the uRLLC, but a combination of the approaches was adopted to satisfy the delay and reliability constraints.

6.2 Future Research Directions

Although, a substantial research work was carried out in this thesis regarding the core enablers of the uRLLC. However, based on the existing results in this thesis, following work can be done in the future:

1. Throughout the thesis, a fixed power allocation scheme is used for the NOMA, OMA and NOMA-RA scheme. In this fixed-power allocation scheme users were allocated the power levels from the pre-fixed or pre-determined power levels. However, it is of capital importance to consider the optimal and flexible power allocation scheme to optimize the resources in the transmission. So, the existing work of NOMA, OMA, and NOMA-RA with short packet communications can further be investigated by considering the flexible power allocation [132] scheme based on the channel conditions of the users.

2. The operation of the multiple access techniques in this thesis considered the perfect CSI and SIC. However, due to the fluctuating channel conditions and addition of multiple users, the complexity in the system could increases. Due to these factors, impact of the imperfect CSI and SIC needs more investigation [133]. How the imperfect CSI and SIC affects the achievable EC and overall throughput of the systems, requires more insights in the future.

3. In chapter 3 and 4, the achievable EC of the downlink NOMA and OMA networks with short-packet communications was derived. The delay performance of these technologies was studied by considering the impact of the transmit SNR, path-loss, and queuing delay violation probability. All these investigations were based on the downlink transmission. However, it is very challenging to investigate the performance of the uplink NOMA and OMA networks with short-packet communications. Therefore, this work can further be extended in the future while considering the uplink transmission of the NOMA and OMA with short-packet communications.

4. In NOMA-RA, multiple users selected different power levels from the predetermined power levels. However, different users can select the same power

levels. This situation can lead to the power collision, during which the SIC will not be performed at the BS and the message of the users will not be decoded and removed. This invites the future work, that can resolve the issue of the power collision, so that the different users will not be able to select the same power levels.

5. In chapter 5, the conditional throughput of the NOMA-RA and multi-channel ALOHA was investigated. This showed, on average, how many users are successfully accessing the channels using the NOMA-RA short-packet communication regime. However, this work can further be studied with the EC or EB framework. This can be achieved by considering the users with single or multiple queues and approximating the arrival rate at these queues. This will pave the way for understanding the NOMA-RA with more detail under the statistical delay requirements.

6. The core metrics investigated in this thesis were the latency and reliability. However, the performance of the uRLLC can also be studied with respect to another key metric such as availability. Latency, reliability, and availability will make a good combination to study in more detail the challenging use case of the uRLLC.

7. The service class of the uRLLC supports the emerging applications with the stringent latency and reliability requirements. In this research work, only the NOMA, NOMA-RA, and short-packet communications were discussed as the core enablers of the uRLLC. However, there are also other enablers of the uRLLC such as proactive packet dropping, edge caching and computing, and multicasting. To study all these enablers of uRLLC is beyond the scope of this research work. To investigate all the potential enablers of the uRLLC with the EC tool, is a new research direction for the future.

6.3 The Impact of this Research on Industrial Practice

The use case of the uRLLC can support diverse range of the futuristic applications, such as autonomous vehicles, factory automation, AR and VR, and smart grid. The research work in this thesis, investigated the performance analysis of the uRLLC enabler's (NOMA, NOMA-RA, short-packet communications). Through this work, the technical guidelines can be derived that will support the practical applications of uRLLC in the future, and that will also have a direct impact on the industrial practices. Following can be the industrial and real-life scenarios, on which this research work will have a impact

- Through the investigation of the delay and reliability performance of the uRLLC, the findings can be used for the mission-critical applications such as autonomous vehicles. This will significantly improve the road-safety and driving efficiency.
- By using the short-packet communications in conjunction with the NOMA or NOMA-RA, the critical applications in smart-grid that requires small data payload can be operated with more reliability. This will help in the more reliable fault detection and diagnosis in the intelligent grid.
- The futuristic applications such as AR or VR work with the stringent requirements of latency and reliability. The use case of uRLLC supports these applications and the enablers of uRLLC investigated in this thesis will provide a technical base for minimizing the delay and improving the reliability. So, the insights from this thesis will become a base to improve the user comfort in case of the emerging applications of the AR or VR.

The above mentioned cases of the industry and applications are few to name, that can be improved or further studied through this research work. This research work will also impact other industrial scenarios such as factory automation. This requires the researchers from industry and academia to investigate in more detail the other enablers of the uRLLC in conjunction with the NOMA, NOMA-RA, and

short-packet communications. This study also provides a starting point for the future research on the enablers of the uRLLC, while residing within the stringent latency and reliability constraints.

Appendix A

The achievable EC with order statistics from (3.15) is given by

$$C_e^u = -\frac{1}{\theta_u n}\ln\left(\frac{\xi_u}{\rho_l}\int_0^\infty\left(\epsilon+(1-\epsilon)\left(1+\alpha_u\gamma_u\right)^{2\zeta_u}+\beta_u\left(1+\alpha_u\gamma_u\right)^{2\zeta_u}\sqrt{1-\left(1+\alpha_u\gamma_u\right)^{-2}}\right.\right.$$
$$\left.\left.+\frac{{\beta_u}^2}{2}\left(1+\alpha_u\gamma_u\right)^{2\zeta_u}\left(1-\left(1+\alpha_u\gamma_u\right)^{-2}\right)\right)e^{-\frac{(V-u+1)\gamma_u}{\rho_l}}\left(1-e^{-\frac{\gamma_u}{\rho_l}}\right)^{u-1}d_{\gamma_u}\right). \tag{1}$$

By using the binomial expansion [134], the expression $\left(1-e^{-\frac{\gamma_u}{\rho_l}}\right)^{u-1}$ in the above equation can be expanded as

$$\left(1-e^{-\frac{\gamma_u}{\rho_l}}\right)^{u-1}=\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i e^{-\frac{\gamma_u}{\rho_l}i}. \tag{2}$$

Using the order statistics and binomial expansion as mentioned above, the equation (1) can further be expanded into three integrals as follows,

$$C_e^u = -\frac{1}{\theta_u n}\ln\left(\epsilon + (1-\epsilon)\left(\frac{\xi_u}{\rho_l}\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i \underbrace{\int_0^\infty (1+\alpha_u\gamma_u)^{2\zeta_u} e^{-\frac{(V-u+1+i)}{\rho_l}\gamma_u} d_{\gamma_u}}_{I_a}\right.\right.$$
$$\underbrace{+\beta_u \int_0^\infty (1+\alpha_u\gamma_u)^{2\zeta_u}\sqrt{1-(1+\alpha_u\gamma_u)^{-2}}e^{-\frac{(V-u+1+i)}{\rho_l}\gamma_u} d_{\gamma_u}}_{I_b}$$
$$\left.\left.\underbrace{+\frac{{\beta_u}^2}{2}\int_0^\infty (1+\alpha_u\gamma_u)^{2\zeta_u}\left(1-(1+\alpha_u\gamma_u)^{-2}\right)e^{-\frac{(V-u+1+i)}{\rho_l}\gamma_u} d_{\gamma_u}}_{I_c}\right)\right). \tag{3}$$

Now, by applying the confluent hypergeometric function of the second kind from (3.17) to the integrals I_a, I_b and I_c of (3), we can write

$$I_a = \mathrm{H}\left(1, 2+2\zeta_u, \frac{V-u+1+i}{\rho_l\alpha_u}\right), \tag{4}$$

$$I_b = \beta_u\left(\mathrm{H}\left(1, 2+2\zeta_u, \frac{V-u+1+i}{\rho_l\alpha_u}\right) - \frac{1}{2}\mathrm{H}\left(1, 2\zeta_u, \frac{V-u+1+i}{\rho_l\alpha_u}\right)\right), \tag{5}$$

$$I_c = \frac{{\beta_u}^2}{2}\left(\mathrm{H}\left(1, 2+2\zeta_u, \frac{V-u+1+i}{\rho_l\alpha_u}\right) - \mathrm{H}\left(1, 2\zeta_u, \frac{V-u+1+i}{\rho_l\alpha_u}\right)\right). \tag{6}$$

Then, inserting (4), (5), and (6) into (3) and using $K_u = \frac{\beta_u^2}{2} + \beta_u$ and $\eta_u = \frac{V-u+1+i}{\rho_l \alpha_u}$, we can approximate the closed-form expression for C_e^u as

$$C_e^u = -\frac{1}{\theta_u n} \ln \left(\epsilon + (1-\epsilon) \left(\frac{\xi_u}{\rho_l \alpha_u} \sum_{i=0}^{u-1} \binom{u-1}{i} (-1)^i \, \mathrm{H}\left(1, 2+2\zeta_u, \eta_u\right) \left(\mathrm{K_u}+1\right) \right. \right. \\ \left. \left. -\mathrm{H}\left(1, 2\zeta_u, \eta_u\right) \left(\mathrm{K_u} - \frac{\beta_u}{2} \right) \right) \right). \tag{7}$$

The closed-form expression for the strong-user's achievable EC at high SNR presented in (3.32) can also be derived by following similar steps as in the above.

Appendix B

With order statistics, the achievable EC of the weak user (v_t) is formulated as

$$C_e^t = -\frac{1}{\theta_t n}\ln\Bigg(\frac{\xi_t}{\rho_l}\int_0^{\infty}\Bigg(\epsilon+(1-\epsilon)\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}+\beta_t\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t} \times\sqrt{1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}}+\frac{\beta_t^2}{2}\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}\left(1-\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{-2}\right)\Bigg) \times e^{-\frac{(V-t+1)\gamma_t}{\rho_l}}\left(1-e^{-\frac{\gamma_t}{\rho_l}}\right)^{t-1}d_{\gamma_t}\Bigg). \tag{8}$$

To simplify (8), we use the generalized binomial expansion [134] and expand the terms $\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}$ and $\left(1-e^{-\frac{\gamma_t}{\rho_l}}\right)^{t-1}$, such that

$$\left(\frac{\gamma_t+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}=\left(\frac{1}{\alpha_u}\right)^{2\zeta_t}\left(1+\frac{\alpha_u-1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}. \tag{9}$$

Then $\left(1+\frac{\alpha_u+1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}$ can be expanded as

$$\left(1+\frac{\alpha_u-1}{\alpha_u\gamma_t+1}\right)^{2\zeta_t}=\sum_{s=0}^{\infty}\binom{2\zeta_t}{s}\left(\frac{\alpha_u-1}{\alpha_u\gamma_t+1}\right)^{s}, \tag{10}$$

where, from [134], it is clear that

$$(1+a)^x = \sum_{y=0}^{\infty} \binom{x}{y} a^y \quad \text{for} \quad |\mathrm{a}| < 1. \tag{11}$$

For $y \geq 1$, $\binom{x}{y}$ can be written as

$$\binom{x}{y} = \frac{x(x-1)...(x-y+1)}{y!} = \frac{(x)_y}{y!}, \tag{12}$$

where $\binom{x}{0} = 1$, and $(.)_y$ is the Pochhammer symbol. Using the binomial expansion, the expression $\left(1 - e^{-\frac{\gamma_t}{\rho_l}}\right)^{t-1}$ from (8) can be expanded as

$$\left(1 - e^{-\frac{\gamma_t}{\rho_l}}\right)^{t-1} = \sum_{r=0}^{t-1} \binom{t-1}{r} (-1)^r e^{-\frac{\gamma_t}{\rho_l} r}. \tag{13}$$

By further simplifying Eq. (8) with the above expansions, we achieve

$$\begin{aligned} C_e^t = -\frac{1}{\theta_t n} \ln \Bigg(\epsilon + (1-\epsilon) \Bigg(& \frac{\alpha_u^{-2\zeta_t} \xi_t}{\rho_l} \Bigg(\sum_{r=0}^{t-1} \binom{t-1}{r} (-1)^r \\ & \times \underbrace{\int_0^{\infty} \left(1 + 2\zeta_t \frac{\alpha_u - 1}{\alpha_u \gamma_t + 1} + \sum_{s=2}^{\infty} \binom{2\zeta_t}{s} \left(\frac{\alpha_u - 1}{\alpha_u \gamma_t + 1} \right)^s \right) e^{-\frac{(V-t+1+r)\gamma_t}{\rho_l}} d_{\gamma_t}}_{I_1} + \beta_t \sum_{r=0}^{t-1} \binom{t-1}{r} (-1)^r \\ & \times \underbrace{\int_0^{\infty} \sqrt{1 - \left(\frac{\gamma_t + 1}{\alpha_u \gamma_t + 1} \right)^{-2}} \left(1 + 2\zeta_t \frac{\alpha_u - 1}{\alpha_u \gamma_t + 1} + \sum_{s=2}^{\infty} \binom{2\zeta_t}{s} \left(\frac{\alpha_u - 1}{\alpha_u \gamma_t + 1} \right)^s \right) e^{-\frac{(V-t+1+r)\gamma_t}{\rho_l}} d_{\gamma_t}}_{I_2} \\ & + \frac{{\beta_t}^2}{2} \sum_{r=0}^{t-1} \binom{t-1}{r} (-1)^r \\ & \times \underbrace{\int_0^{\infty} \left(1 - \left(\frac{\gamma_t + 1}{\alpha_u \gamma_t + 1} \right)^{-2} \right) \left(1 + 2\zeta_t \frac{\alpha_u - 1}{\alpha_u \gamma_t + 1} + \sum_{s=2}^{\infty} \binom{2\zeta_t}{s} \left(\frac{\alpha_u - 1}{\alpha_u \gamma_t + 1} \right)^s \right) e^{-\frac{(V-t+1+r)\gamma_t}{\rho_l}} d_{\gamma_t} }_{I_3} \Bigg) \Bigg) \Bigg). \end{aligned} \tag{14}$$

For solving the integrals I_1, I_2 and I_3 of (14), we use equations (3.353.2) and (3.352.4) from [135], i.e.,

$$\int_0^\infty \frac{e^{-zt}}{t+b} d_t = -e^{bz} E_i(-bz), \ [|\arg\ \mathrm{b}| < \pi, \ \mathrm{Re\ z} > 0], \tag{15}$$

$$\int_0^\infty \frac{e^{-zt}}{(t+b)^n} d_t = \frac{1}{(n-1)!} \sum_{j=1}^{n-1} (j-1)!(-z)^{n-j-1}(b^{-j}) - \frac{(-z)^{n-1}}{(n-1)!} e^{bz} E_i(-bz), \\ [n \geq 2, \ |\arg\ \mathrm{b}| < \pi, \ \mathrm{Re\ z} > 0]. \tag{16}$$

Finally, by using (22) and (23) we reach the closed-form expression for C_e^t as shown in (3.23). The closed-form expression for the achievable EC of the weak user at high SNR presented as (3.33) can also be derived by following the above steps.

Appendix C

The achievable EC of the weak NOMA user is given by

$$C_2^{\mathrm{N}} = -\frac{1}{\theta_2 n}\ln\Bigg(\epsilon + (1-\epsilon)\frac{2}{\rho}e^{\psi_2}\int_0^{\infty}\left(\frac{\gamma_2+1}{\alpha_1\gamma_2+1}\right)^{2\Upsilon_2} \times e^{-\frac{2\gamma_2}{\rho}}d\gamma_2\Bigg). \tag{17}$$

Following the generalized binomial expansion, we can write

$$\left(\frac{\gamma_2+1}{\alpha_1\gamma_2+1}\right)^{2\Upsilon_2} = \left(\frac{1}{\alpha_1}\right)^{2\Upsilon_2}\left(1+\frac{\alpha_1-1}{\alpha_1\gamma_2+1}\right)^{2\Upsilon_2}, \tag{18}$$

where the expression $\left(1+\frac{\alpha_1-1}{\alpha_1\gamma_2+1}\right)^{2\Upsilon_2}$ can further be expanded as

$$\left(1+\frac{\alpha_1-1}{\alpha_1\gamma_2+1}\right)^{2\Upsilon_2} = \sum_{k=0}^{\infty}\binom{2\Upsilon_2}{k}\left(\frac{\alpha_1-1}{\alpha_1\gamma_2+1}\right)^{k}. \tag{19}$$

Using the above expansions, (17) can be transformed into

$$C_2^{\mathrm{N}} = -\frac{1}{\theta_2 n}\ln\Bigg(\epsilon + (1-\epsilon)\,\frac{2\alpha_1^{-2\Upsilon_2}}{\rho}e^{\psi_2}\Bigg(\underbrace{\int_0^\infty e^{-\frac{2\gamma_2}{\rho}}d\gamma_2}_{I_a\ (k=0)} + \underbrace{\int_0^\infty 2\Upsilon_2\frac{\alpha_1-1}{\alpha_1\gamma_2+1}e^{-\frac{2\gamma_2}{\rho}}d\gamma_2}_{I_b\ (k=1)} + \underbrace{\int_0^\infty\sum_{k=2}^{\infty}\binom{2\Upsilon_2}{k}\left(\frac{\alpha_1-1}{\alpha_1\gamma_2+1}\right)^k e^{-\frac{2\gamma_2}{\rho}}d\gamma_2}_{I_c\ (k\geq 2)}\Bigg)\Bigg). \tag{20}$$

Applying (4.23) for the integral I_a, we get

$$I_a = \mathrm{H}\left(1, 2, \frac{2}{\rho}\right). \tag{21}$$

For the integrals I_b and I_c, we use (3.353.2) and (3.352.4) from [135] such that,

$$\int_0^\infty \frac{e^{-zt}}{t+b}d_t = -e^{bz}\mathrm{E_i}(-bz), [|\arg\ \mathrm{b}| < \pi,\ \mathrm{Re(z)} > 0], \tag{22}$$

$$\begin{aligned}\int_0^\infty \frac{e^{-zt}}{(t+b)^n}d_t &= \frac{1}{(n-1)!}\sum_{j=1}^{n-1}(j-1)!(-z)^{n-j-1}(b^{-j}) \\ &- \frac{(-z)^{n-1}}{(n-1)!}e^{bz}\mathrm{E_i}(-bz), \quad [n\geq 2,\ |\arg\ \mathrm{b}| < \pi,\ \mathrm{Re\ z} > 0],\end{aligned} \tag{23}$$

where $\mathrm{E_i}(.)$ is the exponential integral [135]. Using (22) and (23), the closed-form expression for C_2^{N} can be derived as given in (4.27).

References

[1] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. Leung, "Enabling massive iot toward 6G: a comprehensive survey," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11 891–11 915, Aug. 2021. 1

[2] C. De Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836–886, 2021. 1

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019. 1, 25, 36, 72, 117

[4] P. P. Ray, N. Kumar, and M. Guizani, "A vision on 6G-enabled NIB: requirements, technologies, deployments and prospects," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 120–127, Aug. 2021. 1, 11

[5] C. She, C. Yang, and T. Q. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, 2017. 2, 25

[6] Ö. F. Gemici, İ. Hökelek, and H. A. Çırpan, "Modeling queuing delay of 5G NR with NOMA under SINR outage constraint," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2389–2403, 2021. 2, 20

[7] 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. 131 796–131 813, 2020. 2, 12

[8] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010. 2, 3, 37, 41, 43, 52, 73, 77, 80

[9] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018. 2, 5, 12, 13, 14, 16, 17, 20, 24, 28, 36, 37

[10] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, *et al.*, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2488–2524, 2019. 2, 13

[11] 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 Network*, vol. 32, no. 2, pp. 8–15, 2018. 2, 13, 14

[12] Z. Xiang, W. Yang, Y. Cai, Z. Ding, Y. Song, and Y. Zou, "NOMA-assisted secure short-packet communications in IoT," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 8–15, 2020. 3, 4

[13] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2018. 3, 4, 37, 38, 73, 75

[14] J. Khan and L. Jacob, "Availability maximization framework for CoMP enabled URLLC with short packets," *IEEE Networking Letters*, vol. 2, no. 1, pp. 1–4, 2020. 3

[15] K. Wang, C. Pan, H. Ren, W. Xu, L. Zhang, and A. Nallanathan, "Packet error probability and effective throughput for ultra-reliable and low-latency UAV communications," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 73–84, 2020. 3

[16] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE communications surveys & tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018. 3, 18, 36

[17] M. Vaezi, G. A. A. Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 900–919, 2019. 3, 21

[18] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2016. 4, 18, 20, 22, 36

[19] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2736–2743, 2017. 4

[20] S. A. Tegos, P. D. Diamantoulakis, A. S. Lioumpas, P. G. Sarigiannidis, and G. K. Karagiannidis, "Slotted ALOHA with NOMA for the next generation IoT," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6289–6301, 2020. 4, 97, 99

[21] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003. 5, 26, 30, 37, 43, 44

[22] D. M. Abdullah and S. Y. Ameen, "Enhanced mobile broadband (EMBB): A review," *Journal of Information Technology and Informatics*, vol. 1, no. 1, pp. 13–19, 2021. 12

[23] S. Henry, A. Alsohaily, and E. S. Sousa, "5G is real: Evaluating the compliance of the 3GPP 5G new radio system with the ITU IMT-2020 requirements," *IEEE Access*, vol. 8, pp. 42 828–42 840, 2020. 12, 13, 14

[24] S. A. Busari, S. Mumtaz, S. Al-Rubaye, and J. Rodriguez, "5G millimeter-wave mobile broadband: Performance and challenges," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 137–143, 2018. 12

[25] C.-H. Lee, R. Y. Chang, C.-T. Lin, and S.-M. Cheng, "Beamforming and power allocation in dynamic TDD networks supporting machine-type communication," in *IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6. 12

[26] D. Feng, L. Lai, J. Luo, Y. Zhong, C. Zheng, and K. Ying, "Ultra-reliable and low-latency communications: applications, opportunities and challenges," *Science China Information Sciences*, vol. 64, no. 2, pp. 1–12, 2021. 15, 16, 17

[27] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6G: integrating domain knowledge into deep learning," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 204–246, 2021. 15, 16

[28] M. Amjad, L. Musavian, and S. Aïssa, "Link-layer rate of noma with finite blocklength for low-latency communications," in *IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2020, pp. 1–6. 18, 73

[29] Z. Liu and L.-L. Yang, "Sparse or dense: A comparative study of code-domain noma systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 8, pp. 4768–4780, Aug 2021. 19

[30] Z. Xiang, X. Tong, and Y. Cai, "Secure transmission for NOMA systems with imperfect sic," *China Communications*, vol. 17, no. 11, pp. 67–78, 2020. 21

[31] H. M. Gürsu, M. Ç. Moroğlu, M. Vilgelm, F. Clazzer, and W. Kellerer, "System level integration of irregular repetition slotted aloha for industrial IoT in 5G new radio," in *IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2019, pp. 1–7. 23

[32] M. Vilgelm, S. R. Linares, and W. Kellerer, "On the resource consumption of M2M random access: Efficiency and pareto optimality," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 709–712, 2018. 23

[33] J. Choi, "Noma-based random access with multichannel ALOHA," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2736–2743, 2017. 23, 100

[34] S. Moon, H.-S. Lee, and J.-W. Lee, "Sara: Sparse code multiple access-applied random access for iot devices," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3160–3174, 2018. 23, 24

[35] Y. Ma, Z. Yuan, W. Li, and Z. Li, "Novel solutions to NOMA based modern random access for 6G enabled IoT," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15 382–15 395, Oct 2021. 24, 97, 99

[36] J.-B. Seo, S. Pack, and H. Jin, "Uplink NOMA random access for UAV-assisted communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8289–8293, 2019. 24, 98

[37] Y. Gu, H. Chen, Y. Li, and B. Vucetic, "Ultra-reliable short-packet communications: Half-duplex or full-duplex relaying?" *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 348–351, 2017. 25, 37

[38] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint power and blocklength optimization for URLLC in a factory automation scenario," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1786–1801, 2019. 25, 74

[39] Q. Du and X. Zhang, "Base-station selections for QoS provisioning over distributed multi-user MIMO links in wireless networks," in *IEEE INFOCOM*, 2011, pp. 3038–3046. 26

[40] M. Amjad, M. H. Rehmani, and S. Mao, "Wireless multimedia cognitive radio networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 2, pp. 1056–1103, 2018. 26

[41] M. G. Khoshkholgh, K. Navaie, K. G. Shin, and V. CM Leung, "Provisioning statistical QoS for coordinated communications with limited feedback," in *IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6. 26

[42] X. Guo, L. Dong, Y. Li, and L. Wang, "Effective capacity of mimo mrc system with constant and variable power loading," in *13th Canadian Workshop on Information Theory (CWIT)*, 2013, pp. 117–121. 26

[43] K. Angrishi and U. Killat, "Analysis of a real-time network using statistical network calculus with effective bandwidth and effective capacity," in *14th GI/ITG Conference Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB)*, 2008, pp. 1–15. 26

[44] K. Angrishi, "An end-to-end stochastic network calculus with effective bandwidth and effective capacity," *Computer Networks*, vol. 57, no. 1, pp. 78–84, 2013. 26

[45] M. Hammouda, S. Akin, and J. Peissig, "Effective capacity in multiple access channels with arbitrary inputs," in *IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2015, pp. 406–413. 26

[46] C. Wang, B. Urgaonkar, A. Gupta, L. Y. Chen, R. Birke, and G. Kesidis, "Effective capacity modulation as an explicit control knob for public cloud profitability," in *IEEE International Conference on Autonomic Computing (ICAC)*, 2016, pp. 95–104. 28

[47] Z. Feng, G. Wen, and C. W. Chen, "Multiuser effective capacity analysis for queue length based rate maximum wireless scheduling," in *1st IEEE International Conference on Communications in China (ICCC)*, 2012, pp. 438–442. 28

[48] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2018. 28, 30

[49] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *IEEE Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People NFOCOM'95*, vol. 3, 1995, pp. 1001–1009. 28, 29

[50] C.-S. Chang, *Performance guarantees in communication networks*. New York: Springer Science & Business Media, 2012. 28, 29

[51] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Effective capacity region and optimal power control for fading broadcast channels," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2011, pp. 2974–2978. 32

[52] W. Yu, L. Musavian, and Q. Ni, "Statistical delay QoS driven energy efficiency and effective capacity tradeoff for uplink multi-user multi-carrier systems," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3494–3508, Aug 2017. 32

[53] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1849–1858,, April 2017. 32

[54] A. H. Anwar, K. G. Seddik, T. ElBatt, and A. H. Zahran, "Effective capacity of delay-constrained cognitive radio links exploiting primary feedback," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7334–7348, 2016. 32

[55] R. Sassioui, L. Szczecinski, L. Le, and M. Benjillali, "AMC and HARQ: effective capacity analysis," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2016, pp. 1–7. 32

[56] D. Qiao, "Effective capacity of buffer-aided relay systems with selection relaying," in *IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–7. 32

[57] Y. Khan, M. Derakhshani, S. Parsaeefard, and T. Le-Ngoc, "Self-organizing TDMA MAC protocol for effective capacity improvement in IEEE 802.11 WLANs," in *IEEE Globecom Workshops (GC Wkshps)*, 2015, pp. 1–6. 32

[58] D. Qiao, "Effective capacity of buffer-aided full-duplex relay systems with selection relaying," *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 117–129, 2016. 32

[59] Z. Xiao, X. Xie, and S. Zhang, "Effective capacity analysis of link layer based on cognitive multiple access channel," in *5th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 2015, pp. 136–141. 32

[60] W. Yu, L. Musavian, and Q. Ni, "Multi-carrier link-layer energy efficiency and effective capacity tradeoff," in *IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 2763–2768. 32

[61] M. Hammouda, S. Akin, and J. Peissig, "Effective capacity in cognitive radio broadcast channels," in *IEEE Global Communications Conference (GLOBECOM)*, 2014, pp. 1071–1077. 32

[62] M. M. Butt, A. H. Anwar, A. Mohamed, and T. ElBatt, "Effective capacity of cognitive radio links: Accessing primary feedback erroneously," in *11th International Symposium on Wireless Communications Systems (ISWCS)*, 2014, pp. 395–400. 32

[63] R. Li and J. Li, "Effective capacity analysis in underlay cooperative cognitive radio network," in *IEEE XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, 2014, pp. 1–4. 32

[64] X.-Q. Shi and Q.-X. Chu, "Effective capacity of delay quality-of-service guarantees with imperfect channel information in spectrum-sharing environment," in *International Conference on Computer Communication and Informatics (ICCCI)*, 2014, pp. 1–6. 32

[65] S. Efazati and P. Azmi, "Effective capacity maximization in multirelay networks with a novel cross-layer transmission framework and power-allocation scheme," *IEEE transactions on vehicular technology*, vol. 63, no. 4, pp. 1691–1702, 2014. 32

[66] M. Elalem, "Effective capacity with interference constraints in multichannel spectrum sharing system," in *IEEE International Conference on Research and Innovation in Information Systems (ICRIIS)*, 2013, pp. 404–409. 32

[67] Y. Yang, H. Ma, and S. Aissa, "Relay selection from an effective capacity perspective," in *IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013, pp. 1066–1070. 32

[68] W. Cheng, X. Zhang, and H. Zhang, "QoS-aware power allocations for maximizing effective capacity over virtual-MIMO wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 10, pp. 2043–2057, 2013. 32

[69] M. Elalem and L. Zhao, "Effective capacity and interference constraints in multichannel cognitive radio network," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 2993–2998. 32

[70] X. Han, H. Chen, L. Xie, and K. Wang, "Effective capacity region in a wireless multiuser OFDMA network," in *IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 1794–1799. 32

[71] Q. Zhu and X. Zhang, "Effective-capacity based auctions for relay selection over wireless cooperative communications networks," in *IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6. 33

[72] K. P. Peppas, P. T. Mathiopoulos, and J. Yang, "On the effective capacity of amplify-and-forward multihop transmission over arbitrary and correlated

fading channels," *IEEE Wireless Communications Letters*, vol. 5, no. 3, pp. 248–251, 2016. 33

[73] W. Yu, L. Musavian, and Q. Ni, "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3339–3353, 2016. 33

[74] Q. Zhu and X. Zhang, "Effective-capacity based gaming for optimal power and spectrum allocations over big-data virtual wireless networks," in *IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6. 33

[75] L. Musavian and Q. Ni, "Effective capacity maximization with statistical delay and effective energy efficiency requirements," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3824–3835, 2015. 33

[76] W. Zhou, J. Wang, and S. Li, "Effective capacity of a secondary user without channel side informatuon in nakagami fading channels," in *International Conference on Communications, Circuits and Systems (ICCCAS)*, vol. 1, 2013, pp. 17–20. 33, 34

[77] F. Benkhelifa, Z. Rezki, and M.-S. Alouini, "Effective capacity of nakagami-m fading channels with full channel state information in the low power regime," in *IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013, pp. 1883–1887. 33

[78] S. Aksu and G. K. Kurt, "Effective capacity in multihop multi-rate adaptive cooperative networks under nakagami-m fading," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013, pp. 3926–3931. 33, 34

[79] W. Cheng, X. Zhang, and H. Zhang, "Maximizing effective capacity over wireless links under average and peak power constraints," in *IEEE International Conference on Communications (ICC)*, 2012, pp. 5190–5194. 33

[80] M. Pirmoradian and C. Politis, "Power allocation in cognitive spectrum sharing area using effective capacity in imperfect fading channel," in *In-*

ternational Symposium on Computer Networks and Distributed Systems (CNDS), 2011, pp. 116–121. 33

[81] L. Musavian and S. Aïssa, "Effective capacity of delay-constrained cognitive radio in nakagami fading channels," *IEEE Transactions on Wireless Communications*, vol. 9, no. 3, pp. 1054–1062, March 2010. 33, 34

[82] S. Ren and K. B. Letaief, "Maximizing the effective capacity for wireless cooperative relay networks with QoS guarantees," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 2148–2159, July 2009. 33

[83] X. Zhang and J. Wang, "Statistical QoS-driven power allocation for wifi offloading over heterogeneous wireless networks," in *Information Sciences and Systems (CISS), 2017 51st Annual Conference on*, 2017, pp. 1–6. 33

[84] J. Wang and X. Zhang, "Statistical QoS-driven cooperative power allocation game over wireless cognitive radio networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6. 33

[85] X. Zhang and J. Wang, "Statistical QoS-driven power adaptation over Q-OFDMA-based full-duplex D2D 5G mobile wireless networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6. 33

[86] W. Cheng, X. Zhang, and H. Zhang, "Decentralized heterogeneous statistical QoS provisioning for uplinks over 5G wireless networks," in *IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–7. 33

[87] J. Wang and X. Zhang, "Heterogeneous QoS-driven resource adaptation over full-duplex relay networks," in *IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6. 33

[88] W. Cheng, X. Zhang, and H. Zhang, "Statistical-QoS driven energy-efficiency optimization over green 5G mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3092–3107, 2016. 33

[89] X. Zhang and Q. Zhu, "Information-centric network virtualization for QoS provisioning over software defined wireless networks," in *IEEE Military Communications Conference (MILCOM)*, 2016, pp. 1028–1033. 33

[90] Y. Chen and I. Darwazeh, "An accurate approximation of delay with nakagami-m channels and exponential arrivals," in *IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6. 33, 34

[91] P. B. Oni, R. Duan, and M. Elmusrati, "Dual analysis of the capacity of spectrum sharing cognitive radio with MRC under nakagami-m fading," in *Conference Papers in Science*, vol. 2013, 2013. 34

[92] U. Celentano and S. Glisic, "Effective capacity of imperfect adaptive wireless communication systems," in *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 4, 2005, pp. 2191–2195. 34

[93] Y. Chen and I. Darwazeh, "An estimator for delay distributions in packet-based wireless digital communication systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 825–829. 34

[94] M. Matthaiou, G. C. Alexandropoulos, H. Q. Ngo, and E. G. Larsson, "Effective rate analysis of MISO rician fading channels," in *IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2012, pp. 53–56. 34

[95] Y. Chen and I. Darwazeh, "End-to-end delay performance analysis in IEEE 802.16j mobile multi-hop relay (mmr) networks," in *18th International Conference on Telecommunications*, 2011, pp. 488–492. 34

[96] S. Vassaki, A. D. Panagopoulos, and P. Constantinou, "Effective capacity and optimal power allocation for mobile satellite systems and services," *IEEE Communications Letters*, vol. 16, no. 1, pp. 60–63, 2012. 34

[97] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key technologies for 5G wireless systems*. Cambridge university press, 2017. 36

[98] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016. 36, 37

[99] P. Popovski, J. J. Nielsen, C. Stefanovic, E. De Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park, *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, 2018. 37

[100] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 618–629, 2015. 37

[101] J. Choi, "An effective capacity-based approach to multi-channel low-latency wireless communications," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2476–2486, 2019. 37

[102] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3007–3038, 2019. 37

[103] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, pp. 1–13, 2013. 37, 44, 78

[104] M. Shehab, H. Alves, and M. Latva-aho, "Effective capacity and power allocation for machine-type communication," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4098–4102, 2019. 37, 74, 78

[105] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550–4564, 2018. 38, 72

[106] E. Dosti, M. Shehab, H. Alves, and M. Latva-aho, "On the performance of non-orthogonal multiple access in the finite blocklength regime," *Ad Hoc Networks*, vol. 84, pp. 148–157, 2019. 38, 73

[107] Y. Xu, C. Shen, T.-H. Chang, S.-C. Lin, Y. Zhao, and G. Zhu, "Energy-efficient non-orthogonal transmission under reliability and finite blocklength constraints," in *IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2017, pp. 1–6. 38

[108] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of noma under statistical delay qos guarantees," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4907–4922, 2018. 38

[109] M. Amjad and L. Musavian, "Performance analysis of noma for ultra-reliable and low-latency communications," in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–5. 38, 72

[110] M. Amjad, L. Musavian, and S. Aïssa, "Noma versus oma in finite blocklength regime: Link-layer rate performance," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 253–16 257, 2020. 38

[111] H. A. David and H. N. Nagaraja, *Order statistics Encyclopedia of Statistical Sciences*, 2004. 41, 77

[112] J. E. Gentle, *Computational statistics*. Springer, 2009, vol. 308. 41

[113] C. C. Maican, *Integral evaluations using the Gamma and Beta functions and elliptic integrals in engineering: A Self-study Approach*. International Press, 2005. 41

[114] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, no. 5, pp. 913–931, 1994. 43

[115] J. Bucklew, *Introduction to rare event simulation*. Springer Science & Business Media, 2004. 44

[116] M. Abramowitz, I. A. Stegun, and R. H. Romer, "Handbook of mathematical functions with formulas, graphs, and mathematical tables," 1988. 47, 77, 81

[117] C. Guo, S. Wu, Z. Deng, J. Jiao, N. Zhang, and Q. Zhang, "Age-optimal power allocation policies for NOMA and hybrid NOMA/OMA systems," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6. 73

[118] D. Marasinghe, N. Rajatheva, and M. Latva-Aho, "Block error performance of NOMA with HARQ-CC in finite blocklength," in *IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–6. 73

[119] Z. Wang, T. Lv, Z. Lin, J. Zeng, and P. T. Mathiopoulos, "Outage performance of URLLC NOMA systems with wireless power transfer," *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 380–384, 2019. 74

[120] D.-D. Tran, S. K. Sharma, S. Chatzinotas, I. Woungang, and B. Ottersten, "Short-packet communications for MIMO NOMA systems over nakagami-m fading: BLER and minimum blocklength analysis," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3583–3598, 2021. 74

[121] M. Bello, "Asymptotic regime analysis of noma uplink networks under qos delay constraints," *arXiv preprint arXiv:2001.11423*, 2020. 74

[122] M. Shehab, H. Alves, E. A. Jorswieck, E. Dosti, and M. Latva-aho, "Effective energy efficiency of ultra-reliable low latency communication," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 135–11 149, July 2021. 75

[123] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014. 78

[124] M. Qu, J. Liu, J.-B. Seo, and H. Jin, "Distributed fair channel access in NOMA random access systems," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6. 97

[125] M. R. Amini and M. W. Baidas, "Random-access NOMA in URLL energy-harvesting IoT networks with short packet and diversity transmissions," *IEEE Access*, vol. 8, pp. 220 734–220 754, 2020. 98

[126] S. Zhou, Z. Zhang, J. Wang, C. Jiao, and Y. Chen, "NOMA/CA: NOMA-based random access with pattern detection and collision avoidance," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2018, pp. 1–7. 98

[127] G. Zhanyang and H. Jin, "Collision resolution algorithm for multi-user NOMA random access systems," in *International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2020, pp. 1121–1123. 99

[128] Y. Wu, N. Zhang, and K. Rong, "Non-orthogonal random access and data transmission scheme for machine-to-machine communications in cellular networks," *IEEE Access*, vol. 8, pp. 27 687–27 704, 2020. 100

[129] L. Mai, Q. Zhang, and J. Qin, "System throughput maximization of uplink NOMA random access systems," *IEEE Communications Letters*, vol. 25, no. 11, pp. 3654–3658, Nov 2021. 100

[130] M. R. Amini and M. W. Baidas, "Performance analysis of grant-free random-access NOMA in URLL IoT networks," *IEEE Access*, vol. 8, no. 3, pp. 709–712, 2021. 100

[131] W. Yu, C. H. Foh, A. ul Quddus, Y. Liu, and R. Tafazolli, "Throughput analysis and user barring design for uplink NOMA-enabled random access," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6298–6314, Oct 2021. 100, 104

[132] S. Han, X. Xu, Z. Liu, P. Xiao, K. Moessner, X. Tao, and P. Zhang, "Energy-efficient short packet communications for uplink noma-based massive mtc networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 066–12 078, 2019. 120

[133] D.-T. Do, T. A. Le, T. N. Nguyen, X. Li, and K. M. Rabie, "Joint impacts of imperfect CSI and imperfect SIC in cognitive radio-assisted NOMA-V2X communications," *IEEE Access*, vol. 8, pp. 128 629–128 645, 2020. 120

[134] Y. Wang and K. R. Liu, "Statistical delay QoS protection for primary users in cooperative cognitive radio networks," *IEEE Communications Letters*, vol. 19, no. 5, pp. 835–838, 2015. 124, 127, 128

[135] S. Efazati and P. Azmi, "Statistical quality of service provisioning in multi-user centralised networks," *IET Communications*, vol. 9, no. 5, pp. 621–629, 2015. 129, 131