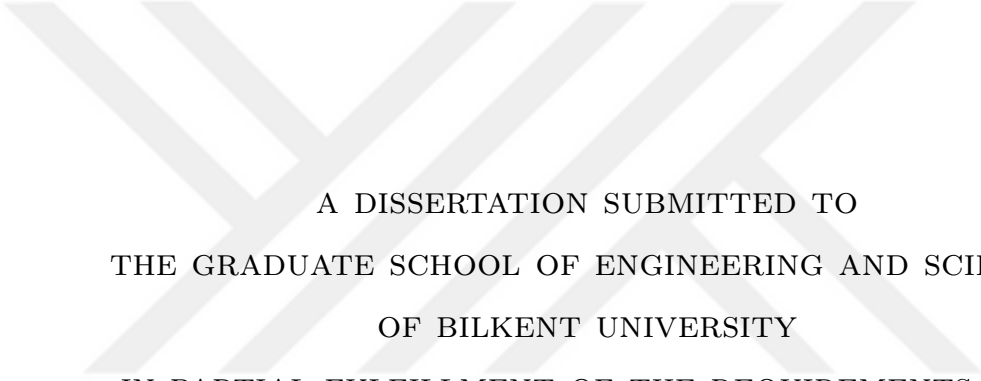


TIMELY THROUGHPUT MAXIMIZATION USING MULTIPLE ACCESS CHANNEL



A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Onur Berkay Gamgam
September 2023

Timely Throughput Maximization using Multiple Access Channel

By Onur Berkay Gamgam

September 2023

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



Ezhan Karařan(Advisor)

Nail Akar

İbrahim Körpeođlu

Cenk Toker

Mehmet Akif Yazıcı

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

TIMELY THROUGHPUT MAXIMIZATION USING MULTIPLE ACCESS CHANNEL

Onur Berkay Gangam

Ph.D. in Electrical and Electronics Engineering

Advisor: Ezhan Karasan

September 2023

Latency and reliability capabilities of currently available fourth-generation (4G) wireless networks paved the path towards massively connected devices requiring much lower latency and much higher reliability. In the fifth-generation (5G) wireless networks, the concept of ultra-reliable and low-latency communications (URLLC) is introduced to fulfill these demands. URLLC aims to deliver short packets with 1 ms latency with a reliability rate of 99.999%. The cellular Internet of Things (IoT) is a framework for conceptualizing such massive connectivity while addressing fundamental challenges such as the ever-increasing number of interconnected devices, latency constraints, and high-throughput demands. One of the challenging tasks for cellular IoT applications is the delivery of deadline-constrained information to densely deployed IoT devices. Increasing demand for delivering timing-critical information in cellular IoT networks poses a URLLC-oriented challenge for both academia and industry. With this motivation, this thesis aims to develop techniques for reliably transferring short packets to densely deployed devices within a given deadline.

In this thesis, we address the problem of latency-constrained communications with strict deadlines under average power constraint using Multiple Access (MA) schemes. The first MA scheme considered in the thesis is Hybrid MA, which consists of both Orthogonal MA (OMA) and power domain Non-Orthogonal MA (NOMA) as transmission scheme options. The second MA scheme studied in the thesis is Rate-Splitting Multiple Access (RSMA), which generalizes OMA, NOMA and Space-Division MA (SDMA) schemes. We maximize the timely throughput, which represents the average number of successfully transmitted packets before deadline expiration, where expired packets are dropped from the buffer. We use Lyapunov stochastic optimization methods to develop a dynamic power assignment algorithm for minimizing the packet drop rate while satisfying time average power constraints. Moreover, we propose a flexible packet dropping mechanism called Early Packet Dropping (EPD) to detect likely to become expired packets

and drop them proactively. Finally, we propose a simple heuristic to reduce the computational load of the proposed algorithm. Numerical results show that Hybrid MA improves the timely throughput compared to conventional OMA by up to 46% and on average by more than 21%. With EPD, these timely throughput gains improve to 53% and 24.5%, respectively. Utilization of RSMA with EPD further improves timely throughput by up to 5.95% and on the average by about 3.12% compared to Hybrid MA with EPD. Simulation results indicate that the proposed heuristic significantly reduces the computational load at the cost of a small loss in the timely throughput performance.

Keywords: Non-Orthogonal Multiple Access (NOMA), Rate-Splitting Multiple Access (RSMA), Deadline-constrained communications, Dynamic decision making, Lyapunov optimization, Power efficient algorithms.

ÖZET

ÇOKLU ERİŞİM KANALI KULLANARAK ZAMANINDA VERİM MAKSİMİZASYONU

Onur Berkay Gamgam

Elektrik ve Elektronik Mühendisliği, Doktora

Tez Danışmanı: Ezhan Karaşan

Eylül 2023

Mevcut dördüncü nesil (4G) kablosuz ağların gecikme süresi ve güvenilirlik özellikleri, çok daha düşük gecikme süresi ve çok daha yüksek güvenilirlik gerektiren kitlesel bağlantılı cihazlara giden yolu açmıştır. Beşinci nesil (5G) kablosuz ağlarda, bu talepleri karşılamak için ultra güvenilir ve düşük gecikmeli iletişim (URLLC) kavramı ortaya atılmıştır. URLLC, 99.999% güvenilirlik oranı ile 1 ms gecikme süresine sahip kısa paketler sunmayı amaçlamaktadır. Hücresel Nesnelerin İnterneti (IoT), sürekli artan birbirine bağlı cihaz sayısı, gecikme kısıtlamaları ve yüksek verim talepleri gibi temel zorlukları ele alırken, bu tür büyük bağlantıları kavramsallaştırmak için bir çerçevedir. Hücresel IoT uygulamaları için zorlu görevlerden biri, yoğun olarak konuşlandırılmış IoT cihazlarına son tarih kısıtlı bilgilerin iletilmesidir. Hücresel IoT ağlarında zamanlama açısından kritik bilgilerin iletilmesine yönelik artan talep, hem akademi hem de endüstri için URLLC odaklı bir zorluk teşkil etmektedir. Bu motivasyonla, bu tez, kısa paketlerin yoğun olarak konuşlandırılmış cihazlara belirli bir son tarih içinde güvenilir bir şekilde aktarılması için teknikler geliştirmeyi amaçlamaktadır.

Bu tezde, Çoklu Erişim (MA) şemaları kullanarak ortalama güç kısıtı altında sıkı son teslim tarihlerine sahip gecikme kısıtlı iletişim problemini ele alıyoruz. Tezde ele alınan ilk MA şeması, iletim şeması seçenekleri olarak hem Ortogonal MA (OMA) hem de güç alanı Ortogonal Olmayan MA (NOMA) içeren Hibrit MA'dır. Tezde incelenen ikinci MA şeması ise OMA, NOMA ve Uzam Bölüşümlü MA (SDMA) şemalarını genelleştiren Hız Bölmeli Çoklu Erişim (RSMA) şemasıdır. Süresi dolan paketlerin tampondan düştüğü son tarih sona ermeden önce başarıyla iletilen paketlerin ortalama sayısını temsil eden zamanında verimi en üst düzeye çıkarıyoruz. Lyapunov stokastik optimizasyon yöntemlerini kullanarak, zaman ortalama güç kısıtlamalarını karşılarken paket düşme oranını en aza indiren dinamik bir güç atama algoritması geliştiriyoruz. Ayrıca, süresi dolması muhtemel paketleri tespit etmek ve proaktif olarak düşürmek için

Erken Paket Düşürme (EPD) adı verilen esnek bir paket düşürme mekanizması öneriyoruz. Son olarak, önerilen algoritmanın hesaplama yükünü azaltmak için basit bir sezgisel yöntem öneriyoruz. Sayısal sonuçlar, Hibrit MA'nın geleneksel OMA'ya kıyasla zamanında elde edilen verimi 46%'a kadar ve ortalama olarak 21%'dan fazla artırdığını göstermektedir. EPD ile, bu zamanında verim kazanımları sırasıyla 53% ve 24,5%'a yükselmektedir. EPD ile RSMA'nın kullanılması, EPD'li Hibrit MA'ya kıyasla zamanında verimi 5,95%'a kadar ve ortalama olarak yaklaşık 3,12% artırır. Simülasyon sonuçları, önerilen sezgisel yöntemin zamanında verim performansında küçük bir kayıp pahasına hesaplama yükünü önemli ölçüde azalttığını göstermektedir.

Anahtar sözcükler: Ortogonal Olmayan Çoklu Erişim (NOMA), Hız Bölmeli Çoklu Erişim (RSMA), Son tarih kısıtlı iletişim, Dinamik karar verme, Lyapunov optimizasyonu, Güç verimli algoritmalar.

Acknowledgement

I would first like to thank my PhD thesis advisor Prof. Dr. Ezhan Karařan. The door to Prof. Karařan's office was always open whenever I ran into a trouble or had a question about my PhD studies. He constantly allowed this PhD thesis to be my own work, but steered me in the right direction whenever he thought I needed it. I will always be grateful for his guidance through my PhD thesis study and my career.

I would like to express my deep gratitude for my PhD thesis monitoring committee members Prof. Dr. Nail Akar and Prof. Dr. İbrahim K rpeođlu for their passionate participation and input through my PhD thesis studies. I would also like to thank my dissertation committee members Prof. Dr. Cenk Toker and Asst. Prof. Dr. Mehmet Akif Yazıcı for reviewing my PhD thesis and providing helpful feedback.

My friends Noyan Cem Sev ktekin and Mustafa Uđur Dalođlu deserve a big thank you for always supporting me with their invaluable ideas. Of all things, Noyan's deep knowledge in mathematics and engineering has been a source of pride and inspiration for me.

I want to convey my deepest appreciation to my parents, Zehra and Hamza, as well as my brother Ege, for their unwavering support and continuous encouragement during my educational journey and throughout the research and writing of this thesis. Without them, this achievement would not have been possible. I would also like to thank our beloved dogs Naz, whom we lost 5 years ago, and Heri for the joy and happiness they brought to my life during this journey.

Finally, I would like to express my gratitude from the bottom of my heart to the love of my life, G k en, for her endless and invaluable support throughout this long journey. Thank you for everything you have brought to my life in ways I could have never imagined.

I would like to dedicate my PhD thesis to Atat rk, as his spiritual heir.

Contents

List of Abbreviations	xv
List of Symbols	xvii
1 Introduction	1
1.1 Background and Motivation of the Thesis	1
1.2 Contributions	9
1.3 Outline	13
2 Literature Review	14
2.1 Problem of Latency Constrained Communication using MA Schemes	14
2.2 Studies on Timing in Wireless Communications using Hybrid MA	16
2.2.1 Average Latency as Timing Metric	16
2.2.2 Strict Deadlines as Timing Metric	17
2.2.3 Information Freshness as Timing Metric	19
2.2.4 Packet Blocklength as Timing Metric	22
2.3 Studies on Timing in Wireless Communications using RSMA . . .	23
2.3.1 Average Latency as Timing Metric	23
2.3.2 Information Freshness as Timing Metric	24
2.3.3 Packet Blocklength as Timing Metric	25
3 Timely Throughput Maximization using Hybrid MA	27
3.1 System Model	27
3.2 Early Packet Dropping Mechanism	33
3.3 Optimization Problem Formulation for Power Allocation using Hybrid MA	35

3.4	Optimal Power Assignment for Hybrid MA	40
3.4.1	Optimal Power Assignment for OMA	40
3.4.2	Optimal Power Assignment for NOMA	43
3.5	Computational Load Reduction Heuristic	49
4	Numerical Results and Discussions for Hybrid MA	52
4.1	Dynamic Power Allocation without EPD	54
4.2	Dynamic Power Allocation with EPD	62
4.3	Computational Load Reduction Heuristic	68
5	Timely Throughput Maximization with RSMA	78
5.1	System Model and Problem Formulation	78
5.2	Optimal Power Assignment for two-user layer-1 RSMA	82
6	Numerical Results and Discussions for RSMA	90
6.1	Dynamic Power Allocation with EPD	91
6.2	Complexity Reduction Heuristic	97
7	Conclusions and Future Work	104

List of Figures

1.1	Generalization provided with RSMA	6
1.2	Comparison of OMA, SDMA, NOMA and RSMA schemes in terms of spectrum utilization.	8
3.1	Illustration of an access point equipped with a single antenna and users in the coverage area. Each user is equipped with a single antenna, too.	28
3.2	Illustration of queue dynamics for 3 consecutive time slots.	32
3.3	Flow chart of the proposed algorithm for timely throughput max- imization using Hybrid-MA.	39
4.1	Average drop rate as a function of α with $\pi = 0.1$	55
4.2	Average drop rate as a function of α with $\pi = 0.3$	55
4.3	Average drop rate as a function of α with $\pi = 0.5$	56
4.4	Average power consumption as a function of γ with $\pi = 0.1$	57
4.5	Average power consumption as a function of γ with $\pi = 0.3$	58
4.6	Average power consumption as a function of γ with $\pi = 0.5$	58
4.7	Average incomplete bit-rate as a function of N with $\pi = 0.3$	61
4.8	Average drop rate with $\pi = 0.1$ as a function of S in EPD mechanism.	63
4.9	Average drop rate with $\pi = 0.3$ as a function of S in EPD mechanism.	64
4.10	Average drop rate with $\pi = 0.5$ as a function of S in EPD mechanism.	64
4.11	Average incomplete bit-rate as a function of N with $\pi = 0.3$ and EPD utilization.	67
4.12	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 5$ and $\pi = 0.1$	69

4.13	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 5$ and $\pi = 0.3$.	70
4.14	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 5$ and $\pi = 0.5$.	70
4.15	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 10$ and $\pi = 0.1$.	71
4.16	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 10$ and $\pi = 0.3$.	71
4.17	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 10$ and $\pi = 0.5$.	72
4.18	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 20$ and $\pi = 0.1$.	73
4.19	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 20$ and $\pi = 0.3$.	73
4.20	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 20$ and $\pi = 0.5$.	74
4.21	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 40$ and $\pi = 0.1$.	75
4.22	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 40$ and $\pi = 0.3$.	75
4.23	Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 40$ and $\pi = 0.5$.	76
5.1	Flow chart of the proposed algorithm for timely throughput maximization using RSMA.	81
6.1	Average drop rate $\overline{D}$ as a function of α with $\pi = 0.1$	92
6.2	Average drop rate $\overline{D}$ as a function of α with $\pi = 0.3$	92
6.3	Average drop rate $\overline{D}$ as a function of α with $\pi = 0.5$	93
6.4	Average power $\overline{P}$ as a function of γ with $\pi = 0.1$ and $\alpha = 0.1$. . .	93
6.5	Average power $\overline{P}$ as a function of γ with $\pi = 0.3$ and $\alpha = 0.1$. . .	94
6.6	Average power $\overline{P}$ as a function of γ with $\pi = 0.1$ and $\alpha = 0.5$. . .	94
6.7	Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 5$ and $\pi = 0.1$.	98

- 6.8 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 5$ and $\pi = 0.3$. 98
- 6.9 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 5$ and $\pi = 0.5$. 99
- 6.10 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 10$ and $\pi = 0.1$. 100
- 6.11 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 10$ and $\pi = 0.3$. 100
- 6.12 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 10$ and $\pi = 0.5$. 101
- 6.13 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 20$ and $\pi = 0.1$. 102
- 6.14 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 20$ and $\pi = 0.3$. 102
- 6.15 Normalized timely throughput and normalized computation time
as a function of RSMA Subset Count Y under $N = 20$ and $\pi = 0.5$. 103

List of Tables

4.1	Simulation parameters and default values	53
4.2	In H-MA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$	60
4.3	In H-MA and NOMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$	60
4.4	Mean and variance of FI for different algorithms and γ values under $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$	61
4.5	Mean and variance of FI for different algorithms and γ values under $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$	61
4.6	Mean and variance of FI for different algorithms and γ values under $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$	62
4.7	In H-MA and s-OMA comparison, percentage of increase in timely throughput for different π values and EPD utilization cases for under $\gamma = 0.6$, $\alpha = 0.1$ and $V = 100$. In all cases, s-OMA is operated without EPD.	65
4.8	In H-MA and NOMA comparison, percentage of increase in timely throughput for different π values and EPD utilization cases under $\gamma = 0.6$, $\alpha = 0.1$ and $V = 100$. In all cases, NOMA is operated without EPD.	66
4.9	Mean and variance of FI for different algorithms and γ values under EPD mechanism, $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$	67
4.10	Mean and variance of FI for different algorithms and γ values under EPD mechanism, $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$	68
4.11	Mean and variance of FI for different algorithms and γ values under EPD mechanism, $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$	68

6.1	Simulation parameters and default values	91
6.2	In RSMA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$	95
6.3	In H-NOMA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$	95
6.4	In RSMA and H-NOMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$	95
6.5	Mean and variance of FI for H-NOMA, RSMA algorithms and γ values under EPD mechanism, $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$. . .	96
6.6	Mean and variance of FI for H-NOMA, RSMA algorithms and γ values under EPD mechanism, $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$. . .	97
6.7	Mean and variance of FI for H-NOMA, RSMA algorithms and γ values under EPD mechanism, $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$. . .	97

List of Abbreviations

4G	Fourth-Generation Wireless Networks
5G	Fifth-Generation Wireless Networks
IoT	Internet of Things
URLLC	Ultra-Reliable and Low-Latency Communications
MA	Multiple Access
OMA	Orthogonal Multiple Access
SDMA	Space-Division Multiple Access
NOMA	Non-Orthogonal Multiple Access
RSMA	Rate-Splitting Multiple Access
FDMA	Frequency-Division Multiple Access
TDMA	Time-Division Multiple Access
H2H	Human-to-Human
H2M	Human-to-Machine
M2M	Machine-to-Machine
M2H	Machine-to-Human
MTC	Machine Type Communication
3GPP	3rd Generation Partnership Project
mmWave	millimeter-wave
VR	Virtual Reality
TI	Tactile Internet
FA	Factory Automation
MTC	Machine Type Communication
QoS	Quality of Service
MU-MIMO	Multi User-Multiple Input Multiple Output
CSI	Channel State Information

AoI	Age of Information
FIFO	First-in First-out
FBL	Finite Blocklength
RDPPA	Remaining Deadline based Parametric Prioritization Approach
EPD	Early Packet Dropping
EDF	Earliest Deadline First
CQSI	Channel Queue State Information
SoA	State of the Art
CR	Cognitive Radio
MEC	Mobile Edge Computing
WSAoI	Waighted Sum Age of Information
MDP	Markov Decision Process
SCA	Successive Convex Approximation
AP	Access Point
AWGN	Additive White Gaussian Noise
BLER	Block Error Rate
SNR	Signal to Noise Ratio
RV	Random Variable
i.i.d.	Independent and Identically Distributed
TM	Transmission Mode
CTM	Complete Transmission Mode
FTM	Fragmented Transmission Mode
SIC	Successive Interference Cancellation
FI	Fairness Index
s-OMA	soft-OMA
H-MA	Hybrid-MA
h-OMA	hard-OMA
p-OMA	P_{inf} -OMA
H-NOMA	Hybrid-MA (Same as H-MA)

List of Symbols

$\mathbb{Z}$	The set of whole numbers.
$\mathbb{Z}_+$	The set of positive whole numbers.
$\mathcal{N}$	The set of users
N	Number of users
t	Slot
τ	Slot duration in seconds
$\mathcal{B}$	Transmission bandwidth
η	Signal-to-Noise Ratio (SNR)
$\mathcal{V}_i$	Channel dispersion of user i
$Q^{-1}(\cdot)$	inverse Q -function.
P_0	Power budget for transmitter
N_0	Noise spectral density
$P_i(t)$	Power allocation for user i
$\mathbf{P}(t)$	Vector of $P_i(t), \forall i \in \mathcal{N}$
$\overline{P}_i$	Average power consumption of user i
$\overline{P}$	Overall average power consumption
W	Watt
kW	Kilowatt
γ_i	Average power constraint for user i
R	Radius of the circular coverage area
r_i	Distance between the transmitter and user i
h_i	Channel gain of user i
$m_i \in \mathbb{Z}_+$	Deadline for packets of user i
π_i	Packet arrival probability of user i
Λ	Set of available packet sizes in bits

$\lambda_i(t)$	Arrival rate for user i
$Q_i(t)$	Queue backlog of user i
$d_i(t)$	# slots left before expiration for user i
$q_i(t)$	# data bits left for packet of user i
$\mu_i(t)$	Departure rate for user i
$D_i(t)$	# dropped bits for user i
$E_i(t)$	# dropped packets by EPD for user i
$\overline{D}_i$	Packet drop rate for user i
$\overline{D}$	Overall packet drop rate
$\phi(t)$	Occupied TM for a packet
Φ_C	Complete TM
Φ_F	Fragmented TM
$\varphi(t)$	Queue reduction ratio
$\alpha_i(t)$	RDPPA parameter for user i
$\mathcal{F}_i(\cdot)$	Proposed cost function
$\overline{\mathcal{F}_i}$	Time average of $\mathcal{F}_i$
$X_i(t)$	Virtual queue of user i
$\mathbf{X}(t)$	Vector of $X_i(t), \forall i \in \mathcal{N}$
$L(\mathbf{X}(t))$	Quadratic Lyapunov Function
$\Delta(\mathbf{X}(t))$	Conditional Lyapunov Drift
V	Weight parameter of <i>drift-plus-penalty</i>
$\mathcal{M}(\mathbf{P}(t))$	Objective function
ϵ_O	BLER for OMA
$\mathcal{P}_i$	OMA power region for user i
P_i^O	Optimal OMA power for user i
ϵ_N	BLER for NOMA
$\{i, j\}$	NOMA user pair, such that $ h_i ^2 \leq h_j ^2$
$\mathcal{P}_{(i,j)}$	NOMA total power region for pair $\{i, j\}$
θ	Auxiliary variable representing total NOMA power
$P_{i,(i,j)}^N$	Optimal NOMA power of user i of pair $\{i, j\}$
$P_{j,(i,j)}^N$	Optimal NOMA power of user j of pair $\{i, j\}$
S	EPD Slot Count
$\hat{\mu}_i(S)$	EPD Transmission Rate Estimate

Y	User count of NOMA/RSMA subset
$\mu_{(i,j)}^c$	RSMA common stream rate of pair $\{i, j\}$
$\mu_{k,(i,j)}^p$	RSMA private stream rate of user k of pair $\{i, j\}$
$P_{(i,j)}^c$	RSMA common stream power
$P_{k,(i,j)}^p$	RSMA private stream power of user k of pair $\{i, j\}$
$P_{(i,j)}^p$	RSMA total private stream power of pair $\{i, j\}$
$P_{k,(i,j)}$	RSMA allocated power for user k of pair $\{i, j\}$
$a_{k,(i,j)}$	RSMA common rate allocated for user k of pair $\{i, j\}$
P_T	RSMA optimal total power
P_c	RSMA optimal common stream power
P_p	RSMA optimal total private stream power
$\mathbb{A}$	Search space for $a_{k,(i,j)}$, $k \in \{i, j\}$
$\overline{\mathbb{A}}$	Reduced search space for $a_{k,(i,j)}$, $k \in \{i, j\}$

Chapter 1

Introduction

1.1 Background and Motivation of the Thesis

The evolution from several decades old human-to-human (H2H) oriented communication ecosystem towards enabling human-to-machine (H2M) and machine-to-machine (M2M) communication types led to the emergence of mission and safety-critical applications requiring reliable and low-latency communications [1]. Widely deployed fourth-generation (4G) communication systems target the latency of 50 ms with 10^{-1} block error rate [2]. Over and above, the latency experienced with 4G can be up to a few seconds. Although the 4G led revolutionary achievements in wireless communications, they are not enough to meet future demands requiring much lower latency and much higher reliability. Therefore, development of wireless communication systems with high-reliability and low latency capabilities became an important challenge for the academia and industry [3]. In fifth-generation (5G) communication systems, the mission and safety critical communications are addressed with ultra-reliable and low-latency communications (URLLC) [1].

A communication system with capability of URLLC aims to provide low-latency and highly reliable communication services [4]. Recent standardization

studies on 5G communications highlight the extremely stringent latency and reliability requirements for the emerging wireless applications. For instance, latency requirements for motion control, mobile automation and electric power grid applications are denoted as up to 0.5 ms, 1 ms and 10 ms, respectively [5]. As further releases of 3GPP are published, 10^{-5} reliability and 1ms latency requirements for a data packet consisting of 32 bytes are boosted towards 10^{-6} reliability and below 1ms latency requirements [6]. Remote control for time-critical applications, factory automation and IoT with massive number of devices are some of these challenging applications, that require sub-millisecond delays. Moreover, extreme conditions of 0.1 ms latency with 99.999999% reliability are being considered lately for beyond 5G URLLC applications [4].

The effectiveness and usefulness of URLLC are analyzed under four different potential application categories [3]. The first one is ultra-reliable millimeter-wave (mmWave) communication. Millimeter-wave communication is a key enabler for 5G, since there are vast spectrum regions in high-frequency bands. The challenge of leveraging millimeter-wave communication consists of coping with variable link status, blockage, and high propagation loss. As a URLLC-oriented solution approach to this challenge, learning the best transmission policy according to the observed fluctuations in the user's transmission rates is proposed. This URLLC-oriented approach is observed to be better compared to a system leveraging only average metrics.

The second application category is virtual reality (VR), which demands $< 20ms$ motion-to-photon latency for the user to experience smooth and accurate movements. The challenge is to render high-definition video with the limited computational capability of the VR device. As a URLLC-oriented solution approach, the rendering task is offloaded to a nearby server using mmWave communication.

The third application category is mobile edge computing. In a latency-constrained task offloading scenario, the waiting time of a task in the mobile server becomes non-negligible. Since the task queue length on the mobile server is related to the task waiting time, extreme value theory is used to analyze the

task queue length for investigating task offloading performance.

Finally, multi-connectivity for ultra-dense networks is considered as a URLLC application category. In ultra-dense networks, there are some users located in an area that is covered by a set of base stations. The purpose is to connect each user with several base stations to increase overall system reliability while taking energy efficiency constraints into consideration. These application categories have different characteristics, and they can illustrate a potential basis for possible real-life applications. Promising features of URLLC are shown to be useful and effective in these distinct application categories. The high potential of URLLC for a vast range of applications and its challenging constraints highlight the significance of the research on URLLC.

There are numerous use cases of URLLC. Replacement of the existing cabled communications in systems with wireless capabilities is a URLLC use case [7]. In Industry 4.0, this advancement will revolutionize industrial automation by creating new cooperative M2M interactions. For instance, reduction of cost by improving installation and maintenance, increased robustness toward motion based industrial applications, and increased flexibility in system deployment can be achieved with the replacement of wired communication with radio links. Other applications of industrial automation are motion control and process monitoring in complex factory facilities.

Another use case of URLLC services is the tactile internet (TI), in which interactively generated information must be transmitted with low latency [8]. An ambitious application of TI is remote medical surgical consultation and remote surgery, where the consequences of a simple error might be irreversible. Intelligent transport is another use case of URLLC [7]. For instance, autonomous cars can share information between each other to effectively react to situations on the road. For such applications, the role of the low-latency communication is crucial. URLLC is especially important in IoT networks for low-latency and reliable transmission of time-critical sensor information under heterogeneous and resource-restricted environments [9].

Another critical aspect of IoT networks is the increasing number of connected devices [10], such as Factory Automation (FA) [11]. The number of M2M devices connected to the global network is expected to increase to 14.7 billion in 2023, which will form more than half of the globally connected network [12]. Moreover, the rate of increase in the M2M over the years is the highest compared to other device types [12]. Therefore, there is an evident need for the research on the time-critical and massive-connectivity aspects of IoT networks. In this thesis, we focus on the time-critical cellular IoT applications with massive connectivity, as a URLLC use case.

The emerging communication ecosystems facilitate massive M2M and machine-to-human (M2H) connectivity [1]. The cellular Internet of Things (IoT) is a framework for conceptualizing such massive connectivity while addressing fundamental challenges such as the ever-increasing number of interconnected devices, latency constraints, and high-throughput demands [1, 13, 14]. For instance, Machine Type Communication (MTC) based cellular IoT applications focus on the delivery of small-sized packets between a significantly large number of devices with latency, reliability, and connectivity constraints [15]. Moreover, real-time applications demanding URLLC, such as factory automation, highlight the concept of deadline [16]. Reliable communication for IoT with small-sized packets requires short blocklength transmission, which further increases the problem complexity [17]. Such challenging URLLC constraints in IoT applications motivate the research on latency-constrained communications under strict deadlines [18, 19].

The time-critical aspects of cellular IoT applications bring out the concept of deadline as the maximum allowable time duration for the successful delivery of a data packet. If the data packet is not fully transmitted within the deadline duration, then it is considered useless and dropped out of the system [20]. For the performance evaluation of deadline-constrained systems, the notion of timely throughput is proposed for the analysis of relation between latency and reliability [21]. It represents the long-term average rate of data packets that are successfully delivered within their deadlines [22]. Therefore, timely throughput is an analytical metric suitable for the throughput and quality of service (QoS) evaluation of

the deadline constrained systems. Communication protocols and resource allocation schemes can be designed to meet the requirements of deadline-constrained traffic by optimizing the timely throughput [16].

The fact that half of the connected devices on the global network is evolving towards IoT-focused M2M underlines the huge demand for massive connectivity [12]. The limitations of widely used OMA schemes introduce new challenges in terms of increasing the efficiency of available resources in order to satisfy the emerging massive connectivity demand. NOMA is able to adaptively allocate resources according to the traffic load and user channel state information (CSI), therefore, spectrum and energy efficiency can be increased under various conditions [23]. In terms of instantaneous throughput, NOMA is shown to be better than OMA [24]. Moreover, NOMA increases connectivity in the system by increasing the number of concurrent transmissions using the same spectral resource [25]. Yet, the advantage of NOMA in terms of system capacity depends on the diversity of user channel conditions and the number of connected users.

In timely throughput oriented systems, the main purpose is to increase the average number of successfully transmitted packets within their corresponding deadlines. For this purpose, allocation of available resources for a single user with OMA or multiple users with NOMA are both suitable options. Moreover, some users may not be able to perform NOMA due to the high computational complexity of successive interference cancellation (SIC) for the decoding process of NOMA. Based on the system state, allocation of available spectral resources for a single user with OMA or for multiple users with NOMA can be a suitable option for increasing the timely throughput. In order to address various needs of emerging applications, a Hybrid MA scheme consisting of adaptively switching between OMA and NOMA in the time domain is considered by the 3rd Generation Partnership Project (3GPP) [26, 27, 28, 29], which is a collaborative project to develop globally applicable specifications for mobile systems.

Rate-Splitting Multiple Access (RSMA) is another promising technique for next-generation networks such as 6G due to its capability of generalizing OMA,

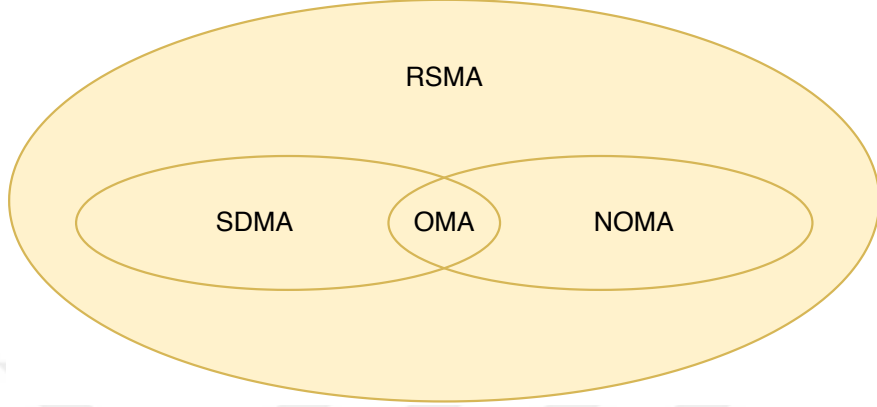


Figure 1.1: Generalization provided with RSMA

NOMA and Space Division Multiple Access (SDMA) to further improve the spectral efficiency and the robustness against interference under various conditions [30, 31]. Multi User-Multiple Input Multiple Output (MU-MIMO) is a key enabling technology for achieving high throughput in 6G. Interference management is a critical aspect for the successful operation of MU-MIMO, since imperfect CSI causes interference as the number of users increases. Fundamental limitations of multi-antenna networks, which originate from the imperfect CSI at the transmitter, is a strong motivation for the research on the RSMA [32]. Interference management capabilities of RSMA increases system performance in such challenging high throughput scenarios [33]. In [34], effective throughput is maximized in a MU-MIMO URLLC scenario with SIC errors and RSMA is shown to be better compared to other MA schemes. Superior spectral efficiency and interference management capabilities provided with RSMA further improves the latency performance of communication systems. One of the significant sources of latency in URLLC applications is the retransmission of unsuccessful packet deliveries. RSMA provides a reliable transmission scheme under interference, which improves the latency by reducing the packet retransmission [33]. Another important source of latency is the blocklength. RSMA is shown to be superior to other MA schemes, including NOMA, in the short blocklength communication [35]. Moreover, RSMA is shown to achieve the same Max-Min fairness rate as NOMA and SDMA with shorter blocklength, which indicates reduction in latency while satisfying the fairness constraints [36].

NOMA and SDMA are different generalizations of OMA, and RSMA generalizes all of these MA schemes, as illustrated in Figure 1.1. RSMA can adaptively adjust the amount of interference to be decoded with SIC and the amount of interference to be treated as noise, according to the system state. When the interference level is low, the interfering signal can be treated as noise, as applied in SDMA. When the interference level is high, the interfering signal can be removed using SIC, as applied in NOMA. Under moderate interference, a combination of treating interference as part of the noise and removing interference using SIC can be considered with RSMA to increase system performance. The following questions for an MA scheme can be addressed with RSMA:

- Being robust against imperfect CSI.
- Being adaptable to interference level.
- Being general enough to cover all other MA schemes and perform better than them all.

Therefore, the improvements provided with RSMA are significant in terms of the decoding complexity, robustness to imperfect CSI, flexible interference management and superior spectral efficiency [31]. As a use case, the idea of decoding the RSMA common stream for a subset of the users using SIC is proposed as a flexible RSMA scheme in [37]. This approach allows more efficient utilization of the system resources and the available bandwidth.

The spectrum utilization of MA schemes OMA, SDMA, NOMA and layer-1 RSMA are illustrated in Figure 1.2, where the application of SIC is denoted with red color. Single-user OMA solely utilizes the whole spectrum for a single user as demonstrated in Figure 1.2a, whereas two user SDMA, NOMA and RSMA leverages the same spectrum for two users in Figures 1.2b, 1.2c, 1.2d respectively. There are two non-orthogonal streams in SDMA and NOMA. There are three non-orthogonal streams in layer-1 two user RSMA, which consist of a common stream and two private streams of each user. The spectral efficiency of RSMA is higher than NOMA. There are SIC in both RSMA and NOMA, therefore the

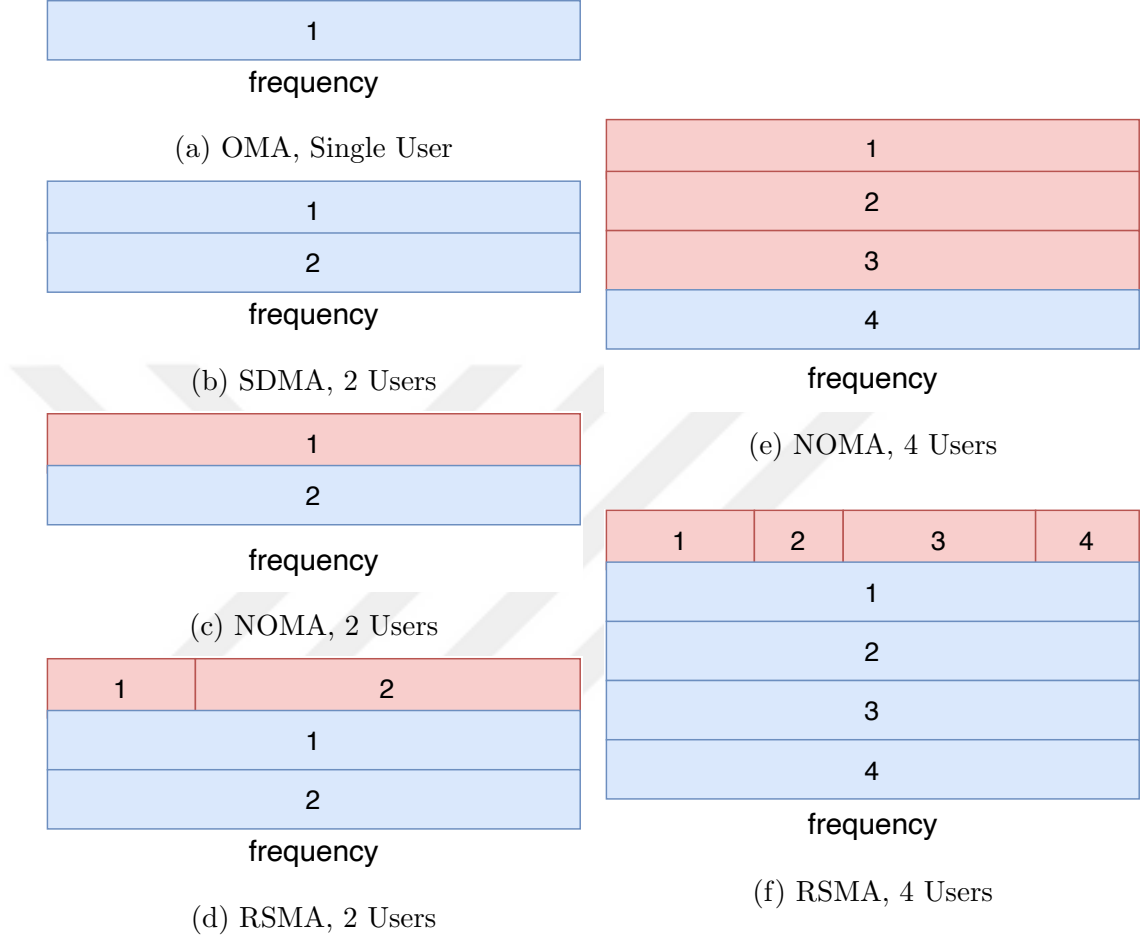


Figure 1.2: Comparison of OMA, SDMA, NOMA and RSMA schemes in terms of spectrum utilization.

receiver complexities of RSMA and NOMA are similar. The transmitter of RSMA needs to determine the common and private streams for each user, whereas the transmitter of NOMA needs to determine the non-orthogonal streams of each user. Therefore, the transmitter of RSMA is more complex than the transmitter of NOMA. The cost of increasing spectral efficiency with RSMA is the increased complexity of the transmitter in the two-user case.

In Figures 1.2e and 1.2f, spectral utilization of four-user NOMA and layer-1 RSMA schemes are illustrated, respectively. There are 3 SIC operations in NOMA scheme, whereas there are only one SIC operation in RSMA scheme. On the transmitter side, the transmitter of RSMA must determine the common

stream and the private streams of four users, whereas NOMA must determine non-orthogonal streams of four users. Therefore, as the user count increases, the receiver complexity of NOMA and the transmitter complexity of RSMA increase. These observations highlight the tradeoff between spectral efficiency and system complexity in using NOMA and RSMA.

Latency minimization is studied for both Hybrid MA and RSMA in the literature. There are studies focusing on optimizing average latency, satisfying strict deadlines, targeting to satisfy information freshness and optimizing the packet blocklength. To the best of our knowledge, there is a lack of study on Hybrid MA and RSMA in the scope of timely throughput maximization for cellular IoT applications under the Finite Blocklength (FBL) regime. Based on this motivation, the focus of this thesis is to maximize the timely throughput using Hybrid MA, which consists of both OMA and NOMA as transmission scheme options, and RSMA for cellular IoT applications.

1.2 Contributions

In this thesis, we address the problem of latency-constrained communications for cellular IoT applications with strict deadlines under time average power constraints using OMA and NOMA-based Hybrid MA scheme and RSMA scheme. We propose a dynamic algorithm that allocates user power in real-time to satisfy time average power constraints while maximizing the timely throughput. We process packets in the first-in-first-out (FIFO) manner. Therefore, only the head of the queue packet is considered for transmission. Other packets in the queue wait until the packets ahead of them are either fully transmitted or dropped. Since packets are dropped due to deadline expiration [20], detection of likely to be expired packets before the expiration of their deadline may further increase the timely throughput. With this motivation, we consider flexible packet dropping schemes to deal with this situation. Finally, we propose a heuristic to reduce the computational load of the proposed algorithm.

We considerably extend the stochastic network optimization framework proposed by Fountoulakis et al. in [38], which focuses on satisfying strict deadline requirements of the user data packets by minimizing the average drop rate due to deadline violations while satisfying time average power constraints. In the previous research [38], only OMA is considered as the transmission scheme. In this thesis, we articulate a solution using an OMA and NOMA based Hybrid MA, and RSMA. We extensively analyzed the potential of the better throughput and connectivity features of NOMA and RSMA for this problem. In the previous research [38], transmission decisions are made in packet-wise manner. In this study, we extend the problem definition with the fragmentation of the packets at the transmitter. In this case, any portion of data in the system is eligible for transmission. In the previous research [38], the size of arriving user packets are assumed to be fixed and time varying nature of wireless medium is modeled as a binary channel. In this thesis, we consider time-varying sizes for arriving user packets and time-varying fading channel model for the wireless medium. Along with the fragmentation of the packets, these extensions form a much more realistic problem scenario. Moreover, adoption of these significant extensions increases problem complexity and this is beneficial for conducting extensive analysis to reveal true potential of complex transmission schemes, such as NOMA and RSMA.

The previous research [38] addresses the packet drop rate minimization task by providing an objective function that is proportional to the remaining packet deadline information. The cost induced by this objective function linearly increases as remaining deadline of the packet diminishes and the objective function reaches the maximum value when the packet expires. In this thesis, we introduce a novel parameter to this objective function as a new degree of freedom. The purpose is to conduct in depth analysis of relation between remaining packet deadline and packet drop rate. The role of the introduced parameter is to adjust increment pattern of the induced cost as the remaining deadline diminishes, so that we can investigate impact of various cost patterns induced from the remaining deadline values. For instance, we can prioritize packets with smallest remaining deadlines among others by keeping cost values high only for the very small remaining deadline cases. As another example, we can induce similar cost values for packets

irrespective of their remaining deadline, so that wide range of packets in the system become eligible candidates for the transmission decision. Throughout this thesis, the prioritization of packets with the proposed technique is indicated as Remaining Deadline based Parametric Prioritization Approach (RDPPA).

The main contributions of this thesis are as follows:

- This thesis contributes to the gap in the literature on timely throughput maximization with OMA and NOMA based hybrid MA, and RSMA schemes.
- We use a realistic model which is appropriate for the cellular IoT scenario. We extend the stochastic network optimization framework for packets with deadlines under average power constraints, proposed by Fountoulakis et al. [38]. The scope of the extensions covers the time-varying arrival content, OMA and NOMA based hybrid MA, and RSMA schemes, fragmentation of packets, and modeling the wireless medium as a fading channel. Moreover, we consider short packet communication with FBL codes, which is appropriate for latency-critical cellular IoT applications.
- We introduce a novel degree of freedom to the objective function to adjust its increment pattern as the remaining deadline diminishes in order to investigate the relation between the remaining packet deadline and the packet drop rate. The prioritization of packets with the proposed technique is called RDPPA.
- We consider constraints on time average power utilization. We propose a dynamic algorithm using Lyapunov stochastic optimization to satisfy time average constraints while minimizing the packet drop rate.
- The proposed dynamic algorithm leverages optimal power allocations for single-user OMA, two-user NOMA and two-user RSMA transmission schemes. Using convex optimization techniques, we derive optimum transmission schemes according to the observed channel and queue state information.

- In order to further improve the performance of the proposed dynamic algorithm, we propose a flexible packet dropping scheme called Early Packet Dropping (EPD) for detecting likely to be expired packets and removing them from the system before their deadline expiration to eliminate unnecessary packet fragment transmissions.
- In order to reduce computational load of the proposed algorithm, we propose a heuristic to reduce the number of user pairs considered as candidates for two-user NOMA and two-user RSMA. The proposed heuristic is called as the computational load reduction heuristic.
- In delay-constrained wired systems, Earliest Deadline First (EDF) is the optimal scheduling algorithm [39]. We show that for fading channels, the drop rate is minimized using RDPPA when packets are prioritized considering the remaining deadline as well as the channel state.

Our key numerical results show that the Hybrid MA outperforms OMA-only based systems by increasing the timely throughput by up to 46% and on the average by more than 21% while satisfying time average power constraints. Proposed flexible packet drop mechanism EPD further improves the timely throughput with Hybrid MA compared to OMA-only without EPD by up to 53% and on the average by more than 24.5%. Utilization of RSMA with EPD further improves timely throughput by up to 5.95% and on the average by about 3.12% compared to Hybrid MA with EPD. We show that RDPPA minimizes the drop rate for fading channels when packets are prioritized considering the remaining deadline and the channel state. In this way, a non-earliest deadline packet of a user with a strong channel condition can be eligible for transmission in order to minimize the overall packet drop rate, instead of the earliest deadline packet of another user with a weak channel condition. RDPPA controls the trade-off between power allocations and Channel-Queue State Information (CQSI) in the system to minimize the overall packet drop rate. Finally, results of simulations conducted with different number of users indicate that the proposed heuristic significantly reduces computational load at the cost of a small loss in the timely throughput performance.

1.3 Outline

The remaining of this thesis is organized as follows. In Chapter 2, we introduce the problem of latency constrained communication using MA schemes. Then, studies on the literature focusing on timing in wireless communications using Hybrid MA and RSMA transmission schemes are presented. In Chapter 3, the system model, the EPD mechanism, the optimization problem for power allocation, the proposed solution for the problem formulation, optimal power assignment for single-user OMA and two-user NOMA are presented. Moreover, the computational load reduction heuristic is explained. In Chapter 4, the numerical results are presented with an elaborate analysis of system parameters' effects for Hybrid-MA. In Chapter 5, the system model, the optimization problem for power allocation, the proposed solution for the problem formulation, optimal power assignment for two-user RSMA are presented. In Chapter 6, the numerical results are presented for RSMA. Finally, the conclusions and future work are presented in Chapter 7.

Chapter 2

Literature Review

In this chapter, we start with introducing the problem of latency constrained communications using MA schemes. Secondly, the metrics considered for the solution of this problem are presented. Then, we survey studies on latency constrained communications using MA schemes. Finally, the contributions of this thesis are presented.

2.1 Problem of Latency Constrained Communication using MA Schemes

State of the art (SoA) on timing in wireless communication often optimizes average latency, operates under strict deadline constraints, aims to maintain information freshness, and optimizes packet block-length [16]. Latency is a metric to quantify the amount of time elapsed with respect to an event in the system. Queuing theory is the main tool for the theoretical analysis of latency in a system [16]. The real-time and interactive applications highlight the metric of strict deadlines, which quantifies the successful occurrence of a time-critical event within the specified amount of time. Therefore, the latency is not a suitable metric for such applications [16]. The remote monitoring type of application

demands the freshness of the received information. The metric considered for analyzing such applications is Age of Information (AoI) [16]. Finally, the length of a data block is important in terms of the transmission latency. Therefore, the packet block-length is another metric that is considered for the optimization of timing in wireless communication.

Communication with low latency is one of the key challenges for 5G and beyond. The low latency constraint requires the use of short packets to reduce the transmission delay. Reliable transmission of the short packets can be ensured with the FBL codes. The authors in [40] proposed a tight approximation for the transmission rate in the FBL regime, leading to a more accurate analysis of latency, reliability, and throughput. Topics of NOMA, hybrid NOMA, and short packet communication are studied in SoA as key solutions for low-latency IoT applications for beyond-5G cellular networks [13]. Moreover, RSMA is shown to be superior to other MA schemes, including NOMA, in the short blocklength communication [35]. Therefore, the potential of MA schemes for latency constrained communications is worth investigating.

With this motivation, the problem of latency constrained communications using MA schemes can be defined as a combination of the following key components. The first component of the problem definition is the MA scheme. The OMA, NOMA and SDMA schemes, and the OMA-NOMA based hybrid MA scheme are widely studied in the literature. The RSMA scheme has a lot of potential over all other MA schemes and is now being focused on by academia. In this thesis, we investigate hybrid-MA and RSMA, which are advanced MA schemes.

The second one is the timing metric, which is to be optimized as the objective of the problem. The SoA timing metrics in the literature are average latency, strict deadlines, information freshness and packet blocklength. In SoA, there is a lack of study on satisfying strict deadlines through advanced MA schemes, such as Hybrid-MA and RSMA. In this thesis, we focus on latency constrained communications with strict deadlines.

The third component is the FBL regime. In low-latency communications, the

effect of FBL codes on the performance must be considered to achieve accurate results. Therefore, we consider the FBL regime. Due to these challenging key components, the latency-constrained communications using advanced MA schemes under the FBL regime is highlighted as a significant research problem in the literature.

In the following two sections, we survey the studies on timing in wireless communication under FBL regime using NOMA, Hybrid MA, and RSMA schemes, respectively. The studies are categorized according to each timing metric.

2.2 Studies on Timing in Wireless Communications using Hybrid MA

2.2.1 Average Latency as Timing Metric

In the first type of study, the aim is to optimize average latency by minimizing the average queue length, based on Little's theorem [18, 19, 41]. Here, the main target is increasing average energy efficiency while minimizing average queue length, which is proportional to the average system latency by the Little's Theorem.

Choi et al. proposed a joint dynamic power control and user pairing algorithm under a Hybrid MA scheme for power efficient and delay-constrained communications in [41]. Based on channel and queue state information, the algorithm adjusts queue backlogs by appropriate user pairing and power allocation. This study is based on convex optimisation of power assignment for candidate transmissions among which a best candidate is selected by the dynamic algorithm, which is based on the *drift – plus – penalty* [42] technique. The *drift – plus – penalty* technique is a Lyapunov based stochastic network optimization method. It refers to a problem of minimizing the expected value of a function that combines a drift term and a penalty term subject to certain constraints. The drift term encapsulates the alteration in the queue backlog, whereas the penalty term reflects the

power expenditure in [41].

Zeng et al. proposes two dynamic power allocation schemes for uplink NOMA systems with statistical delay QoS guarantee in [43]. Statistical delay QoS refers to the probability distribution of the delay experienced by users in a communication system. The proposed dynamic power allocation scheme aims to make the system more suitable for IoT scenarios by improving its effective capacity and energy efficiency. Comparisons are made with existing NOMA and OMA schemes. Furthermore, the paper discusses the Dinkelbach method and its use in the proposed effective energy efficiency maximization scheme.

2.2.2 Strict Deadlines as Timing Metric

The second type of study aims to meet strict deadlines for the problem of latency-constrained communications.

Zafer et al. discusses the problem of energy-efficient transmission of data with deadline constraints over a time-varying wireless channel in [44]. In order to obtain the optimal policy for sending B bits of data within a T duration of a deadline over a Markov fading channel, the authors introduce the *BT – Problem* and use continuous-time stochastic control theory to solve it. Discrete time decisions are determined as samples obtained from the developed continuous time model. The setup is extended to take into account different deadlines for the packets. Moreover, a heuristic rule is proposed for random packet arrival procedures, according to the gained intuition. The authors think that this paper’s idea can be used for other ways to send data in wireless systems that are limited by quality of service.

Lei et al. proposes a deep learning-assisted optimization framework for energy-efficient scheduling in NOMA systems with practical deadline constraints in [45]. The high computational complexity of scheduling and resource allocation tasks is addressed with deep learning techniques. The authors address the challenges of effectively allocating resources to meet the data demands of all users within

specified deadlines. Moreover, they investigate how to navigate the substantial complexity of resource allocation while maintaining rapid execution in real-time applications. However, the FBL regime is not considered in this study.

Xu et al. addresses heterogeneous latency-constrained communications under the FBL regime with hybrid NOMA in [46]. Different latency requirements of two users are modeled as the non-orthogonal transmission of multiple short and urgent packets of a user along with a large non-urgent packet of the other user. The objective is to minimize the transmission energy while considering packet deadlines. The paper presents an algorithm that utilizes a sequential upper bound minimization approach to derive optimal transmission blocklength and power values for one user and suboptimal values for the other user.

Lyu et al. addresses the time-critical nature of MTC for industrial automation applications with hybrid NOMA/Frequency Division Multiple Access (FDMA) under the FBL regime in [47]. A constrained optimization problem is considered to present the performance of a NOMA-assisted small-packet transmission scheme for mission-critical MTCs. The objective is to maximize total revenue over the network under diverse requirements. The paper additionally presents a case study to evaluate the revenue generated across the entire network by weighing the effective throughput and energy consumption.

Zhu et al. proposes a hybrid MA scheme that combines NOMA and Time-Division MA (TDMA) schemes in the FBL regime in [48]. The authors look at two latency-sensitive application scenarios and come up with the best framework for each one of them. Latency-critical communication is modeled as the reliability of the one-shot transmission. The authors prove that the effective capacity and coding rate of the proposed NOMA-TDMA based transmission scheme are jointly concave with respect to blocklength and signal to interference ratio.

Choi et al. considers NOMA for uplink transmissions in the FBL regime with delay constraints in [49]. The paper suggests using a short-message transmission method called NOMA mode. The problem is modeled as transmitting a finite number of packets within a given number of slots with high probability under

a limited power budget constraint. The authors show that this approach can significantly lower the session error probability compared with OMA.

Fountoulakis et al. proposes a dynamic power assignment algorithm for the packet drop rate minimization under strict deadlines with a limited power budget in [38, 50]. Fixed sized packet arrivals are served in a packet per slot manner with an OMA transmission scheme through a wireless medium modeled as a binary channel. Lyapunov optimization is used to obtain an approximate solution, and a dynamic power assignment algorithm is developed with the *drift-plus-penalty* [42] technique for the maximization of timely throughput under average power constraints. The drift term encapsulates the average power consumption constraint. A penalty metric based on the remaining packet deadline until expiration is proposed. The induced penalty reaches the maximum value when the packet expires. This algorithm is as good as EDF when there is enough power, but it's better than EDF by properly balancing the packet drop rate under average power constraints.

We studied on satisfying strict deadlines for cellular IoT through OMA and NOMA-based hybrid MA scheme in [51]. We maximize timely throughput by minimizing packet drop rate under the average power constraints in the FBL regime. In this paper, we extend the stochastic network optimization framework for packets with deadlines under average power constraints, proposed by Fountoulakis et al. [38]. Time-varying arrival content, flexible packet dropping capability, hybrid MA scheme based on OMA and NOMA, as well as the fragmentation of packets, and the modeling of the wireless medium as a fading channel are some of the extensions. Moreover, we consider short packet communication in the FBL regime. We developed a dynamic power assignment algorithm to minimize the packet drop rate while satisfying average power constraints using Lyapunov stochastic optimization.

2.2.3 Information Freshness as Timing Metric

The third type of study is concerned with satisfying information freshness. The AoI is used as a general performance metric for this kind of studies [52]. The AoI

is defined as the time elapsed since the generation of the latest received packet at the destination.

Wang et al. examine the utilization of adaptive NOMA/OMA for timely status updates to multiple clients in wireless networks in [53]. They formulate a Markov decision process problem and develop optimal and suboptimal policies for the base station to switch between NOMA and OMA based on the instantaneous AoI metric. The paper's main contributions consist of the formulation of optimal policies for the two-client scenario, and development of suboptimal policies for the multi-client scenario, which achieve near-optimal performance.

Wang et al. propose an optimization problem to enhance the freshness of data packets in MA channels with energy harvesting sensor nodes under primary user interference constraints in [54]. They combine cognitive radio (CR) technology with heterogeneous MA sensor networks to minimize the average AoI and average peak AoI of sensor nodes. The authors employ two distinct approaches to address the optimization issue, and the outcomes indicate that the Dynamic Power Policy approach outperforms the Proportional Resource Allocation approach in optimizing the age of the secondary user nodes in the entire wireless network, thereby enhancing the system's timeliness and the freshness of data packets. The authors propose that future research endeavors should take into account multichannel access scenarios in CR wireless network scenarios.

Pan et al. consider NOMA for the task of timely information updates in [55]. They explore the concept of AoI and its relation to information freshness in information update systems. The average AoI and packet error rates are compared between conventional OMA systems and NOMA systems. The paper additionally discusses how the implementation of physical-layer network coding can enhance the performance of AoI in the medium SNR regime.

Wang et al. explores the use of hybrid NOMA/OMA for minimizing the weighted sum of expected AoI in [56]. The instantaneous AoI of both clients guides the base station to adaptively alternate between NOMA and OMA for

downstream transmission. In order to decide which scheme to use at the beginning of each time slot, the authors develop an optimal policy for the base station and prove its existence. They also propose a policy that is near-optimal in performance but has a lower computation complexity. To provide useful insights for practical system designs, the performance of different policies is compared and analyzed in numerical results.

Guo et al. discusses age-optimal power allocation policies for downlink short-packet communication in cellular wireless networks in [57]. The authors consider using a hybrid MA scheme with NOMA and OMA under the FBL regime. A stationary power allocation policy for NOMA and an adaptive NOMA/OMA policy based on Lyapunov optimization are proposed by the authors. The objective is to optimize power allocation in order to minimize the expected weighted sum AoI. The paper discusses the concept of AoI in capturing the freshness of information and derives new expressions for BLER of users under the FBL regime.

Wang et al. study hybrid NOMA/OMA for IoT uplink to serve time-sensitive and throughput-sensitive devices in IoT networks under the FBL regime in [58]. The effect of AoI and throughput requirements on NOMA user grouping is studied. To reduce competition for radio resources by sharing the same NOMA block with different types of devices, the authors propose a Mixed Block Strategy (MBS). The authors compare the performance of MBS to that of the uniform block strategy and demonstrate that MBS outperforms the latter under varying parameters in the FBL regime. The paper concludes by stressing the significance of considering diverse service requirements in the FBL regime for maximizing system performance in hybrid NOMA/OMA for IoT uplink networks.

Maatouk et al. consider NOMA and OMA to minimize the average AoI in 5G networks in [59]. The paper compares the two schemes based on average AoI. The numerical findings indicate that despite the enhanced spectral efficiency of NOMA in comparison to OMA, this does not necessarily translate into a lower average AoI in the network. The paper additionally discusses the significance of AoI in diverse domains, including but not limited to environmental monitoring and vehicular networks.

2.2.4 Packet Blocklength as Timing Metric

The fourth type of study is about minimizing packet block-length for reducing transmission latency.

Sun et al. introduces two-user downlink NOMA into short packet communication for achieving low-latency in [60]. The paper examines the potential of NOMA in enhancing fairness and spectral efficiency for low-latency downlink transmission. Effective throughput is increased subject to a FBL constraint for reducing latency. To overcome challenges related to achievable rate and interference cancellation, the authors propose a 1-D search algorithm. Transmission rate and power allocation for two-user NOMA under FBL are optimized.

Xu et al. investigate the problem of minimizing transmission energy for a two-user downlink with strictly heterogeneous latency constraints in [61]. The latency is modeled as packet blocklength, and the transmission energy is minimized under strictly heterogeneous latency requirements using Hybrid NOMA/TDMA under the FBL regime. Moreover, authors investigate the trade-off between block-length and reliability. Various interference mitigation schemes are studied due to the possible infeasibility of conventional successive interference cancellation techniques under heterogeneous receiver conditions. The paper shows that the proposed hybrid transmission scheme saves more energy than other transmission schemes.

The studies on latency-constrained communications with hybrid NOMA in the FBL regime mainly consider delay metrics such as average latency [41], AoI [57, 58, 59, 56, 55], packet block-length [60, 61] and packet deadline [46, 49]. There was a lack of study on maximizing the timely throughput for given strict deadlines using OMA and NOMA based hybrid MA in the FBL regime. To the best of our knowledge, our paper in [51] is the first study focusing on this gap in the literature.

2.3 Studies on Timing in Wireless Communications using RSMA

2.3.1 Average Latency as Timing Metric

There are studies focusing on timing in communications using RSMA. The first type of study focuses on optimizing latency.

Reifert et al. examine energy efficient RSMA algorithms under QoS constraints referring to mixed criticality in [62]. The tradeoff between power minimization and satisfying QoS constraints is addressed. Higher energy efficiency and QoS fulfillment are shown by the application of the algorithm to a cloud-radio access network. Approaches to reduce computation time are investigated for latency sensitive green 6G networks. However, the FBL regime is not considered in this paper. Furthermore, the authors emphasize that it is important to consider mixed criticality levels in future wireless communication networks.

Karacora et al. address the high reliability requirements as criticality levels and formulate multi-connectivity based solution approach using RSMA transmission scheme under the FBL regime in [63]. The objective of this paper is to develop algorithms for mission-critical use-cases of the forthcoming 5G/6G wireless communication. In a multi connectivity scenario, a wireless device can simultaneously connect to multiple access points in order to increase reliability. This approach is useful to mitigate the negative effects of the wireless channel. Moreover, multi-connectivity can reduce latency by improving throughput, providing multiple access paths to the devices.

Xiao et al. study the minimization of overall offloading latency for mobile edge computing (MEC) using Hybrid RSMA-TDMA transmission scheme under the FBL regime in [64]. Two user's offloading time is divided into two phases. In the first phase, both of the users perform offloading. In the second phase, only one of the users performs offloading. While minimizing the offloading delay,

total energy and transmission power constraints are taken into consideration. In scenarios where there is a limited energy budget, the proposed hybrid RSMA-TDMA transmission scheme is shown to be superior to other offloading schemes. Furthermore, the paper discusses potential practical applications of this approach.

Hua et al. study latency minimization in wireless caching networks using RSMA transmission scheme under the FBL regime in [65]. The problem formulation consists of cache placement and RSMA resource management. The problem is modeled as mixed-integer nonlinear programming and solved by decomposing the original problem into two sub-problems, which consist of short-term and long-term RSMA resource allocation. The problem is solved using dynamic programming and alternating optimization methods.

Zhang et al. study latency minimization for Internet of Vehicles (IoV) using RSMA under the FBL regime in [66]. The IoV scenario considered in this study jointly investigates FEderated Edge Learning (FEEL) and platoon control in downlink. High-velocity vehicles experience Doppler effects, which reduces the accuracy of channel state estimators. RSMA is considered to increase spectral efficiency in this challenging scenario. The proposed optimization problem aims to minimize the latency of the FEEL downlink. Therefore, the deviation of the vehicles within the platoon is minimized. The results indicate that the proposed system works well at making communication faster and making sure that IoV platoon control operates properly.

2.3.2 Information Freshness as Timing Metric

In the second type of study, AoI is considered as the timing metric for satisfying information freshness.

Akyürek et al. study energy efficiency in massive MTC in [67]. The authors aim to address the needs and demands of mMTC and offer ideas for how age-aware power allocation could be used in wireless networks. AoI is considered for the

development of age aware power allocation algorithm, which minimizes weighted-sum average AoI (WSAoI). An adaptive transmission scheme, which consists of OMA, NOMA and RSMA, is considered for the optimization of energy efficiency. The problem of WSAoI minimization is modeled as Markov Decision Process (MDP) and optimal and suboptimal policies are developed to select a transmission scheme. Moreover, an action elimination technique is proposed to reduce computational complexity. This is the first work combining AoI and RSMA in the literature.

2.3.3 Packet Blocklength as Timing Metric

In the third type of study, blocklength is optimized to improve timing performance.

Xu et al. introduce RSMA to short-packet communication in FBL regime in [35]. In order to maximize sum rate, optimal linear precoders are designed. A wide range of network loads and user deployments are examined in the paper to determine the relationship between the sum rate and block length of RSMA. In multi-antenna scenario, it is shown that RSMA outperforms NOMA and SDMA under the same blocklength. This result also shows that, RSMA achieves the same sum rate with NOMA and SDMA with smaller blocklength. Therefore, blocklength can be optimized with RSMA to satisfy low latency communication requirements.

Xu et al. examine the performance of RSMA for an uplink communication system in the FBL regime in [68]. The authors derive analytical expressions of the error probability and propose Successive Convex Approximations (SCA) based algorithm to solve the problem. RSMA is shown to improve error probability and throughput performance compared to NOMA. This result shows that, RSMA achieves more reliable communication given a blocklength. Therefore, the same reliability level as NOMA can be achieved with smaller blocklength using RSMA, and this is useful for low-latency communications.

The studies on latency constrained communications with RSMA in the FBL regime mainly concentrated on optimizing latency and blocklength. Thus, there is a lack of study on maximizing the timely throughput for given strict deadlines using RSMA in the FBL regime. To the best of our knowledge, this thesis is the first study focusing on this gap in the literature.

In the next chapter, the system model, the EPD mechanism, the optimization problem for power allocation, the proposed solution for the problem formulation, the optimal power assignment for single-user OMA and two-user NOMA are presented. Then, the computational load reduction heuristic is explained.

Chapter 3

Timely Throughput Maximization using Hybrid MA

In this chapter, the system model for Hybrid MA is presented at first. Secondly, flexible packet dropping mechanism EPD is presented. Then, optimization problem formulation and the proposed solution is presented. This is followed with the derivations of the optimal power assignment for single-user OMA and two-user NOMA. Finally, the computational load reduction heuristic is explained.

3.1 System Model

In this thesis, we consider a downlink broadcast scenario for cellular IoT applications in which a single-antenna access point (AP) transmitting time-critical data to N stationary single-antenna users within its coverage area, as illustrated in Figure 3.1.

The set of users is denoted as $\mathcal{N} \triangleq \{1, \dots, N\}$. Since AP is equipped with a single antenna, there is a single available output link. Therefore, the output link is allocated either for user $\{i\}$'s OMA transmission, or users $\{i, j\}$'s two-user NOMA

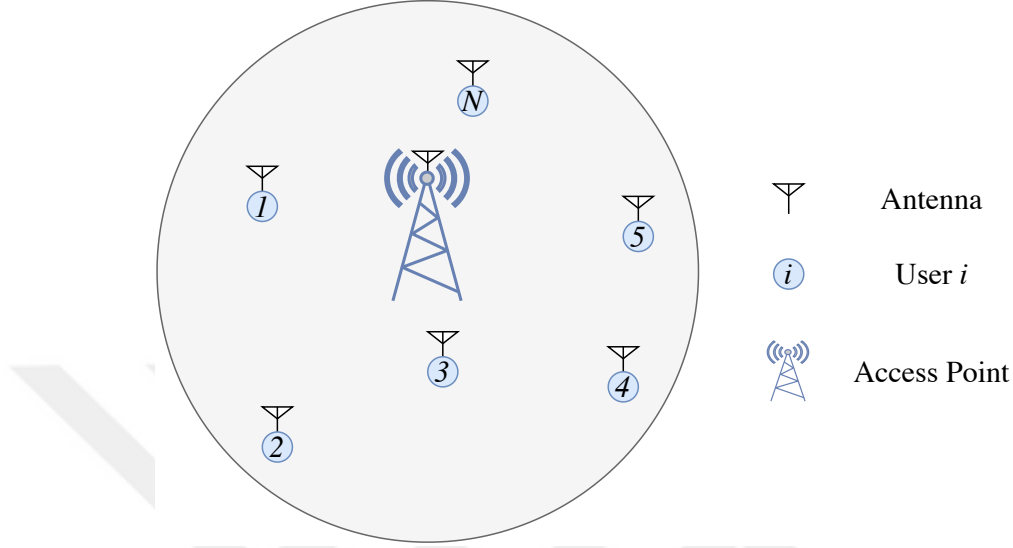


Figure 3.1: Illustration of an access point equipped with a single antenna and users in the coverage area. Each user is equipped with a single antenna, too.

transmission on each time slot. We consider a discrete-time system where the time duration of each slot is denoted as τ and $\mathcal{B}$ represents the transmission bandwidth. Therefore, transmission is performed within FBL of $\tau\mathcal{B}$. Conventional Shannon capacity is based on infinite blocklength; thus it is not applicable for this scenario. Polyanski *et al.* [40] proposed a framework for tightly approximating transmission rate $R^*(\eta)$ in the FBL regime for blocklength $\tau\mathcal{B}$ and block error rate (BLER) ϵ , as follows:

$$R^*(\eta) \approx \log_2(1 + \eta) - \sqrt{\frac{\mathcal{V}(\eta)}{\tau\mathcal{B}}} \cdot \frac{Q^{-1}(\epsilon)}{\ln 2} \quad (3.1)$$

where η is Signal-to-Noise Ratio (SNR), $\mathcal{V}(\eta) = 1 - (1 + \eta)^{-2}$ is the channel dispersion and $Q^{-1}(\cdot)$ is the inverse Q -function. Note that, (3.1) holds for Additive White Gaussian Noise (AWGN) channels. We can apply this rate calculation for downlink NOMA transmission by properly deriving SNR for NOMA user pair, considering Successive Interference Cancellation (SIC) [69]. Assume that AP communicates with a user pair using NOMA. The channel gains of strong and weak users are denoted as h_s and h_w , such that $|h_s|^2 > |h_w|^2$. Moreover, the allocated power by AP for strong and weak users are denoted as P_s and P_w . The strong user removes interference due to the weak user's message using SIC, then decodes its own message. The weak user treats the signal of the strong user as part of the noise and decodes its own message. Let η_s^N and η_w^N denote the SNR

of strong and weak users as,

$$\eta_s^N = \frac{P_s |h_s|^2}{N_0} \quad (3.2)$$

$$\eta_w^N = \frac{P_w |h_w|^2}{P_s |h_w|^2 + N_0} \quad (3.3)$$

where N_0 is the noise spectral density. Then, NOMA transmission rate in FBL regime can be calculated for strong and weak users as $R^*(\eta_s^N)$ and $R^*(\eta_w^N)$ using (3.1), respectively.

The maximum power for the transmitter is denoted as P_0 . $\mathbf{P}(t) = \{P_i(t)\}_{i \in \mathcal{N}}$ corresponds to the transmitter powers allocated for users in a slot t . We consider continuous power levels such that $0 \leq P_i(t) \leq P_0$ for $\forall i, t$. $\mathcal{P}^O(t)$ and $\mathcal{P}^H(t)$ denote the set of all available power assignments for OMA-only and Hybrid MA at a time t , respectively. $\mathcal{P}^O(t)$ is presented in (3.4) and $\mathcal{P}^H(t)$ is presented in (3.5),

$$\forall t, \mathcal{P}^O(t) \triangleq \left\{ \mathbf{P}(t) : \sum_{i=1}^N \mathbb{1}\{P_i(t) > 0\} \leq 1 \right\} \quad (3.4)$$

$$\forall t, \mathcal{P}^H(t) \triangleq \left\{ \mathbf{P}(t) : \sum_{i=1}^N \mathbb{1}\{P_i(t) > 0\} \leq 2 \right\} \quad (3.5)$$

where, $\mathbb{1}\{.\}$ denotes the indicator function, such that $\mathbb{1}\{TRUE\} = 1$ and $\mathbb{1}\{FALSE\} = 0$. Let $\overline{P}_i$ be the average power utilization overall time slots for user i and $\overline{P}$ be the overall average power consumption.

$$\overline{P}_i \triangleq \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} P_i(\tau), \quad \forall i \in \mathcal{N} \quad (3.6)$$

$$\overline{P} \triangleq \frac{1}{N} \sum_{i=1}^N \overline{P}_i \quad (3.7)$$

We consider an average power consumption constraint $\gamma_i \in (0, P_0]$ for each user i , such that $\overline{P}_i \leq \gamma_i$. R denotes the radius of the circular coverage area. Let r_i denote the distance between the transmitter and user i , which is uniformly selected: $r_i \sim U[0, R]$. Let h_i denote the channel gain of user i . The wireless communication link between the transmitter and users is modeled as a Rayleigh fading channel [41]. The Random Variables (RVs) representing the fast fading

component of each user's channel in a slot are independent and identically distributed (i.i.d.). Finally, we assume that the channels between the transmitter and users are static during a time slot, but they alter from slot to slot. This assumption is justified for cellular IoT applications, where mobility is typically low. We assume that channel state information is available at the beginning of each slot, based on the available channel estimation methods for low mobility applications [60, 70].

User i 's arriving data packets are stored in queue i . Let $m_i \in \mathbb{Z}_+$ be the deadline for the arriving packets of user i in terms of slot count. Let $a_i(t) \sim \text{Ber}(\pi_i)$ be a Bernoulli RV with arrival rate π_i representing the arrival probability of a packet for a user i at a slot t . Let $u_i(t) \sim U[\Lambda]$ be the bit count of the arriving packet for the user i in the t^{th} slot, and it is uniformly selected from a finite set of positive integers Λ , which represents available packet sizes. The RVs considered for the arrival processes are i.i.d. The arrival process of user i in the t^{th} slot is denoted as $\lambda_i(t)$, in bits per slot.

$$\lambda_i(t) \triangleq a_i(t) \cdot u_i(t) \quad (3.8)$$

The queue backlog for a user i in the t^{th} slot is denoted as $Q_i(t)$, where $Q_i(0) = 0$. The packets in a queue are processed in the FIFO manner, so that, only the packet at the head of the queue is considered for transmission. Let $d_i(t)$ be the number of slots left before expiration and $q_i(t)$ be the number of data bits left in the t^{th} slot for the packet at the head of the queue i .

The departure process represents the transmitted number of bits in a slot. Let $\Psi_i(t) = j$ be the user paired with the user i in the slot t . If $i = j$, OMA is employed for user i , else, users $\{i, j\}$ are paired for a NOMA transmission. Let $R_i(t, P_i(t), \Psi_i(t))$ be the data rate of user i in bits per second in the slot t , where $P_i(t)$ represents the power assigned to the user i and $\Psi_i(t)$ indicates the MA scheme occupied for the user i . The departure rate for a user i in the t^{th} slot is denoted as $\mu_i(t)$, in bits per slot, as

$$\mu_i(t) \triangleq \tau \cdot R_i(t, P_i(t), \Psi_i(t)) \quad (3.9)$$

where τ represents the slot duration in seconds. The drop event occurs due to either EPD mechanism or deadline violation. At first, the EPD mechanism

detects packets that are likely to be expired. To detect these packets, the EPD mechanism estimates the transmission duration of all packets in the user queues and compares them with the respective remaining packet deadlines. If a packet's estimated transmission duration is greater than its remaining deadline, then it is dropped by the EPD mechanism. The EPD mechanism is presented in Section 3.2. Let $E_i(t)$ be the number of packets of user i dropped at slot t by EPD mechanism. $Q_i(t)$, $q_i(t)$ and $d_i(t)$ are updated according to $E_i(t)$. Secondly, the drop event occurs due to deadline violation for the packet at the head of the queue i when $d_i(t) = 1$ and $\mu_i(t) < q_i(t)$. Let $D_i(t)$ be the number of dropped bits for user i due to deadline violation is denoted as

$$D_i(t) = \mathbb{1}\{d_i(t) = 1\} \cdot \max[q_i(t) - \mu_i(t), 0] \quad (3.10)$$

The queue dynamics is presented in (3.11), based on the arrival and departure processes, queue backlog, and the number of dropped bits. $\overline{D}_i$ represents the packet drop rate due to deadline violation for a user i and $\overline{D}$ is the overall average drop rate.

$$Q_i(t+1) \triangleq \max[Q_i(t) - \mu_i(t), 0] + \lambda_i(t) - D_i(t) - E_i(t) \quad (3.11)$$

$$\overline{D}_i \triangleq \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{1}\{D_i(\tau) > 0\}, \quad \forall i \in \mathcal{N} \quad (3.12)$$

$$\overline{D} \triangleq \frac{1}{N} \sum_{i=1}^N \left(\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{1}\{D_i(\tau) > 0\} + E_i(\tau) \right) \quad (3.13)$$

The queue dynamics of 3 users are illustrated for consecutive 3 time slots $T, T+1, T+2$ in Figure 3.2. The deadline for all arriving packets is considered as $m_i = 5$. Each packet is labeled with the respective remaining number of slots until expiration, which is represented as d . On slot T , there are 2 packets in each users' queue. There is no packet arrival for any one of the users. The head of queue packets of user 1 and user 2 are transmitted using NOMA. Although the head of queue packets of user 3 is not expired yet, it is considered as being likely to be dropped by the EPD mechanism and removed from the queue of user 3 before expiration. The transmitted and dropped packets are removed from the queues. The other packets left in the queues are preserved and their remaining

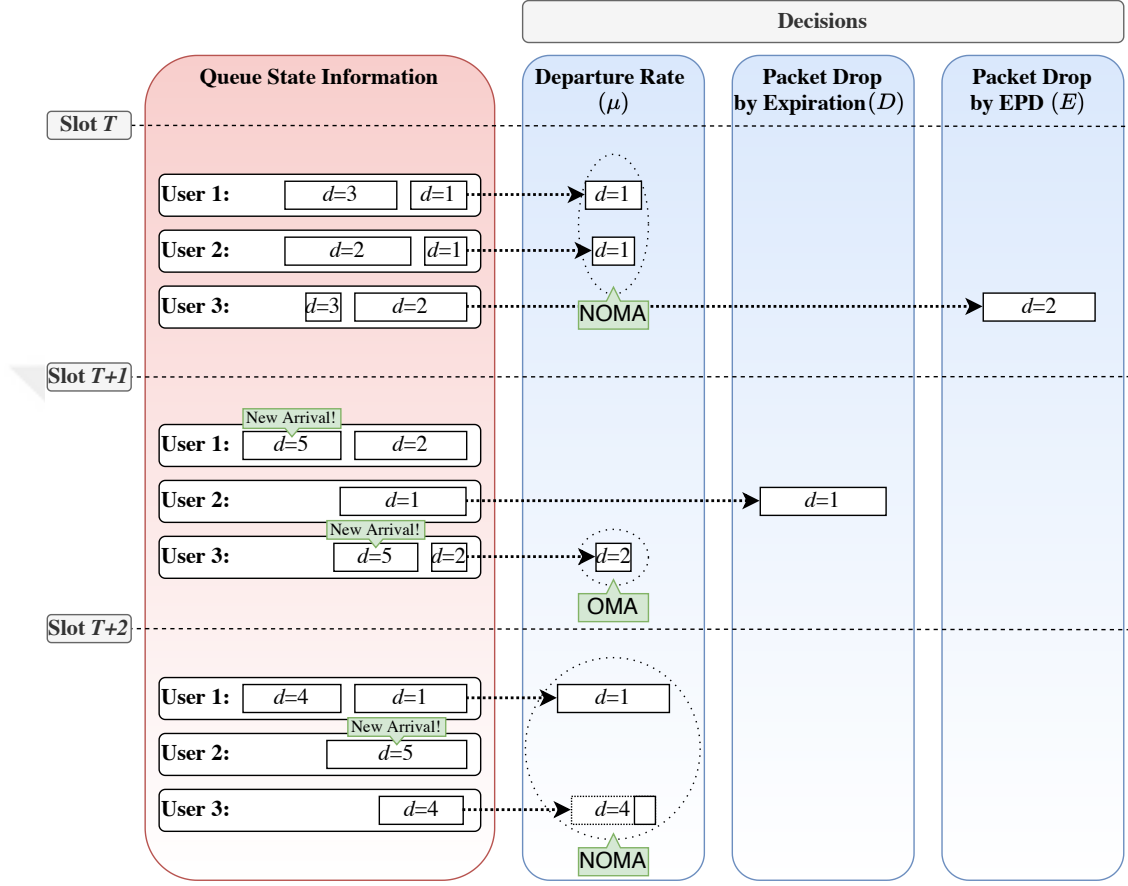


Figure 3.2: Illustration of queue dynamics for 3 consecutive time slots.

number of deadlines until expiration are reduced by 1. On slot $T + 1$, we have arriving packets for user 1 and user 3. The head of queue packet of user 3 is transmitted using OMA. The expired packet of user 2 is dropped. On slot $T + 2$, the head of queue packets of user 1 and a fragment of user 3's head of queue packet are transmitted using NOMA. Moreover, we have a new packet arrival for user 2.

In the next section, flexible packet dropping mechanism EPD is presented.

3.2 Early Packet Dropping Mechanism

In our system model, the packets in the queues are processed in a FIFO manner. According to the definition of timely throughput, a packet is considered to be useless when it is not completely transmitted until the deadline. Therefore, the drop event due to expiration occurs only when a head of queue packet could not be completely transmitted until deadline. If the required number of slots for the transmission of a head of queue packet is greater than or equal to the packet's number of remaining slots until deadline, then this packet is considered as likely to be dropped due to expiration. In this case, before a head of queue packet being transmitted or dropped due to expiration, other existing packets waiting in the queue may be more likely to become expired. The reason is, when these packets become the head of queue packet, their remaining deadline might be very close to expiration, which leads to the loss of timely throughput. Therefore, early dropping of the packets, that are likely to expire, before the deadline violation has the potential to increase timely throughput. With this motivation, we propose the Early Packet Dropping (EPD) algorithm to improve timely throughput.

The purpose of the EPD algorithm is to detect the packets that are impossible to be transmitted completely by the end of their deadlines in the best-case scenario. We define the best-case scenario for a packet as dedicating the transmitter with a full power budget of P_0 only for the OMA transmission of the packet throughout its remaining slots until expiration. We consider the full power budget of P_0 to achieve the maximum transmission rate for the best-case scenario. Since the user channel gain alters in each slot, we need to estimate the OMA transmission rate for future slots. Then, we derive the number of slots required to transmit the packet completely using the estimated OMA transmission rate. We consider a packet as likely to become expired if the derived slot count is greater than the remaining deadline of the packet. EPD aims to detect packets that are likely to become expired due to deadline violation in the best-case scenario and drops them immediately.

We estimate OMA transmission rate of the future slots by averaging the actual

OMA transmission rates regarding a number of most recently observed user channel gain values. Let $S \in \mathbb{Z}$ be the number of slots over which the EPD mechanism estimates the OMA transmission rate. $S \geq 0$ indicates the set of user channel states in the last $S + 1$ slots, including the current slot. Thus, $S = 0$ indicates only the currently observed user channel state. Finally, $S < 0$ indicates the case where EPD mechanism is inactivated. In this case, we set the estimated OMA transmission rate as ∞ , so that, EPD does not drop any packet, and we have $E_i(t) = 0, \forall i, t$ for $S < 0$ case. Let $\hat{\mu}_i(S)$ be the estimated OMA transmission rate for user i which is expressed as follows;

$$\hat{\mu}_i(S) = \begin{cases} \frac{1}{S+1} \sum_{s=0}^S \mu_i(P_0)|_{h_i(t-s)} & , S \geq 0 \\ \infty & , S < 0. \end{cases} \quad (3.14)$$

where $\mu_i(\cdot)$ is the OMA transmission rate in the FBL regime expressed in (3.24).

Algorithm 1 Pseudocode for EPD mechanism

Inputs: $Q_i(t), \forall i \in \mathcal{N}, S$
Outputs: $Q_i(t), E_i(t) \forall i \in \mathcal{N}$
Local Variables: $\{q', d'\}$
Operation:

- 1: **for** each user i **do**
- 2: Initialize $E_i(t) \leftarrow 0$
- 3: Calculate $\hat{\mu}_i(S)$
- 4: **for** each packet in $Q_i(t)$ starting from head of the queue **do**
- 5: $q' \leftarrow$ Bit size of the head of the queue packet from $Q_i(t)$.
- 6: $d' \leftarrow$ Remaining deadline of the head of queue packet from $Q_i(t)$.
- 7: **if** $\lceil \frac{q'}{\hat{\mu}_i(S)} \rceil > d'$ **then**
- 8: Drop the head of the queue packet.
- 9: $E_i(t) \leftarrow E_i(t) + q'$
- 10: $Q_i(t) \leftarrow Q_i(t) - q'$
- 11: Move on to the next packet.
- 12: **else**
- 13: Do not drop the packet. Move on to the next user.
- 14: **break**
- 15: **end if**
- 16: **end for**
- 17: **end for**
- 18: **return** $Q_i(t), E_i(t), \forall i \in \mathcal{N}$

The pseudo-code for the EPD mechanism is presented in Algorithm 1. In line

#1, we start operating EPD on the queue of user i . In lines #2 – 3, we initialize $E_i(t)$ as 0 and calculate $\hat{\mu}_i(S)$, the estimated OMA transmission rate for the user i . In lines #4 – 16, we examine the packets within the queue of user i . The order of examination starts from the head of queue packet. In lines #5 – 6, we get the bit size and the remaining deadline of the head of queue packet. In line #7, we check whether the packet is likely to become expired or not. If the packet is likely to become expired, the packet is dropped from $Q_i(t)$ and $E_i(t)$ is updated accordingly in lines #8 – 11. Otherwise, the packet is preserved in line #13 and EPD moves on to the queue of the next user. Note that, when a packet is dropped from a queue, the head of queue packet in that queue is changed, and EPD starts processing the new packet in line #4. Moreover, if all the packets in a queue are dropped, EPD moves on to the following users. Finally, when all the user queues are processed, the EPD algorithm returns with the updated $Q_i(t)$ and $E_i(t)$.

3.3 Optimization Problem Formulation for Power Allocation using Hybrid MA

The problem of minimizing packet drop rate with deadlines under users' average power constraints [38] using Hybrid MA is presented as follows:

$$\min_{\mathbf{P}(t)} \sum_{i=1}^N \overline{D}_i \quad (3.15a)$$

$$\text{s.t. } \overline{P}_i \leq \gamma_i, \forall i \in \mathcal{N} \quad (3.15b)$$

$$\mathbf{P}(t) \in \mathcal{P}^H(t) \quad (3.15c)$$

We consider two different Transmission Modes (TM) for a packet. The first one is for completely transmitting a packet per slot, which is called Complete TM (CTM). A binary decision is made to either fully transmit the content of a packet or not transmit it at all. The second one is called Fragmented TM (FTM), where a data packet is fragmented at the source for being transmitted in different slots

and reassembled at the destination. A packet is considered to be successfully transmitted only when all its fragments are successfully transmitted before the deadline expiration. Let $\phi(t) \in \{\Phi_C, \Phi_F\}$ be the occupied TM at a time t , where Φ_C and Φ_F represent CTM and FTM, respectively.

We define a metric called queue reduction ratio denoted by $\varphi(t, \phi(t))$ to represent the effect of occupied TM on the head of the queue packet. In CTM, the target is to completely transmit the packet in a slot. Thus, the queue reduction ratio represents whether the head of the queue packet belonging to user i is completely transmitted in the slot with $\mu_i(t)$ or not by taking the values $\varphi(.) = 0$ or 1, respectively. In FTM, the target is to transmit a fragment of the packet in a slot. Thus, the queue reduction ratio represents the remaining ratio of the packet after transmission such that, $\varphi(.) = 0$ represents completely transmitting the packet. $\varphi(.) = 1$ represents no transmission, and transmission of a fragment is linearly represented between these two extremes, i.e., $0 < \varphi(.) < 1$. Note that, $\varphi(.)$ is a continuous function in FTM. The power optimization is performed using the continuous function $\varphi(.)$ and the resultant value is quantized according to the closest bit count of the to be transmitted packet fragment. $\varphi(t, \phi(t))$ can be expressed as follows:

$$\varphi(t, \phi(t)) \triangleq \begin{cases} \mathbb{1}\{q_i(t) - \mu_i(t) > 0\} & , \text{if } \phi(t) = \Phi_C \\ (q_i(t) - \mu_i(t))/q_i(t) & , \text{if } \phi(t) = \Phi_F \end{cases} \quad (3.16)$$

The cost of a packet drop contributes to the minimization of the objective function over the infinite horizon, however, the decision variable $\mathbf{P}(t)$ is optimized slot-by-slot. The future values of CQSI are unknown due to their random nature. Therefore, it is not possible to predict future values of (3.15a). In [38], Fountoulakis et al. introduced the function $f_i(t)$ whose future values are affected by the current decision $P_i(t)$ and the relative difference between the packet deadline, m_i , and the number of remaining future slots, $d_i(t) - 1$, before the expiration of the packet at the head of the queue. In this thesis, we propose a novel function $\mathcal{F}_i(t, \alpha_i(t), \phi(t))$ as:

$$\mathcal{F}_i(t, \alpha_i(t), \phi(t)) \triangleq \left(\frac{m_i - (d_i(t) - 1)}{m_i} \right)^{\alpha_i(t)} \varphi(t, \phi(t)) \quad (3.17)$$

where $\alpha_i(t)$ is a non-negative exponent parameter proposed to adjust the importance of the remaining deadline in the objective function for user i at slot t . The $f_i(t)$ in [38] is equal to $\mathcal{F}_i(t, 1, \Phi_C)$, showing that $\mathcal{F}_i(t, \alpha_i(t), \phi(t))$ has two additional degrees of freedom, $\alpha_i(t)$ and $\phi(t)$. For simplicity of the notation, $\mathcal{F}_i(t, \alpha_i(t), \phi(t))$ is referred to as $\mathcal{F}_i$. Note that $0 \leq \mathcal{F}_i \leq 1$. While, $\mathcal{F}_i = 1$ represents the packet drop event, $\mathcal{F}_i = 0$ indicates that the packet of user i is served completely before expiration. Between these two extreme cases, the remaining deadline of user i 's packet, $d_i(t)$, is mapped to a penalty value which elevates as the remaining deadlines reduce. We propose RDPPA to investigate the relative importance of packets' remaining deadlines in terms of average dropping rate by rapidly and slowly elevating the penalty towards 1 for $\alpha_i(t) < 1$ and $\alpha_i(t) > 1$ cases, respectively. Therefore, $\mathcal{F}_i$ provides information that helps us to predict the future consequences of our actions. Let $\overline{\mathcal{F}_i}$ be the time average of $\mathcal{F}_i$. We define the following new problem:

$$\min_{\mathbf{P}(t)} \sum_{i=1}^N \overline{\mathcal{F}_i} \triangleq \sum_{i=1}^N \left(\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathcal{F}_i \right) \quad (3.18a)$$

$$\text{s.t. } \overline{P_i} \leq \gamma_i, \forall i \in \mathcal{N} \quad (3.18b)$$

$$\mathbf{P}(t) \in \mathcal{P}^H(t) \quad (3.18c)$$

The difference between the problem definitions for packet drop rate minimization in (3.15) and (3.18) is about how the cost is introduced to the objective function. In the problem definition (3.15), the cost is introduced to the objective function only when a packet is dropped. In the problem definition (3.18), the cost is introduced to the objective function according to $\mathcal{F}_i$, whose future values are determined according to the power allocation decision on the correct slot and the remaining deadline of the packet. When a packet is dropped, the introduced cost values are equal in (3.15) and (3.18). Otherwise, the introduced cost is zero in (3.15), whereas, the introduced cost with $\mathcal{F}_i$ in (3.18) is designed to quantify the consequences of our actions on the current slot. In order to solve (3.15), the slot-by-slot decisions need to be made while considering their effects in the infinite horizon. To solve (3.18), per-slot decisions can be made by the dynamic algorithm based on Lyapunov stochastic optimization [42]. Since the solution of

the per-slot decisions-based problem definition in (3.18) is simpler than that of the infinite horizon-based problem definition in (3.15), we focus on solving (3.18).

The problem definition presented in (3.18) is a minimization problem with constraints in the form of time averages, which can be solved using the *drift – plus – penalty* technique [42]. The time average constraints in (3.18b) are transformed into virtual-queues $(X_i(t), \forall i \in \mathcal{N})$ in (3.19), where arrivals are $P_i(t)$ and respective service rates are γ_i . The problem becomes a queue stability problem with a penalty metric. The strong stability of these virtual queues guarantees the respective time average constraints.

$$X_i(t+1) \triangleq \max[X_i(t) - \gamma_i, 0] + P_i(t) \quad (3.19)$$

Let $\mathbf{X}(t)$ be the vector of $X_i(t), \forall i \in \mathcal{N}$. Let $L(\mathbf{X}(t))$ be defined as the quadratic Lyapunov function and $\Delta(\mathbf{X}(t))$ be the conditional Lyapunov drift as,

$$L(\mathbf{X}(t)) \triangleq 1/2 \sum_{i=1}^N X_i^2(t) \quad (3.20)$$

$$\Delta(\mathbf{X}(t)) \triangleq \mathbb{E}[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t)) \mid \mathbf{X}(t)] \quad (3.21)$$

where the expectation is with respect to the random channel states and arrivals. At every time slot t , the problem in (3.18) is solved by determining $\mathbf{P}(t)$ in order to minimize the *drift – plus – penalty* expression [42],

$$\Delta(\mathbf{X}(t)) + V \sum_{i=1}^N \mathbb{E}[\mathcal{F}_i \mid \mathbf{X}(t)] \quad (3.22)$$

where $V > 0$ is a weight parameter to scale the tradeoff between the average power constraint and the penalty related to the deadline. The dynamic policy is obtained by applying the principle of opportunistically minimizing an expectation [42] on the upper bound analysis [38] of the *drift – plus – penalty* expression. We observe CQSI and determine $\mathbf{P}(t)$ by solving the following *drift – plus – penalty* problem at each time slot t :

$$\min_{\mathbf{P}(t)} \mathcal{M}(\mathbf{P}(t)) = \sum_{i=1}^N (V \mathcal{F}_i + X_i(t)(P_i(t) - \gamma_i)) \quad (3.23a)$$

$$\text{s.t. } \mathbf{P}(t) \in \mathcal{P}^H(t) \quad (3.23b)$$

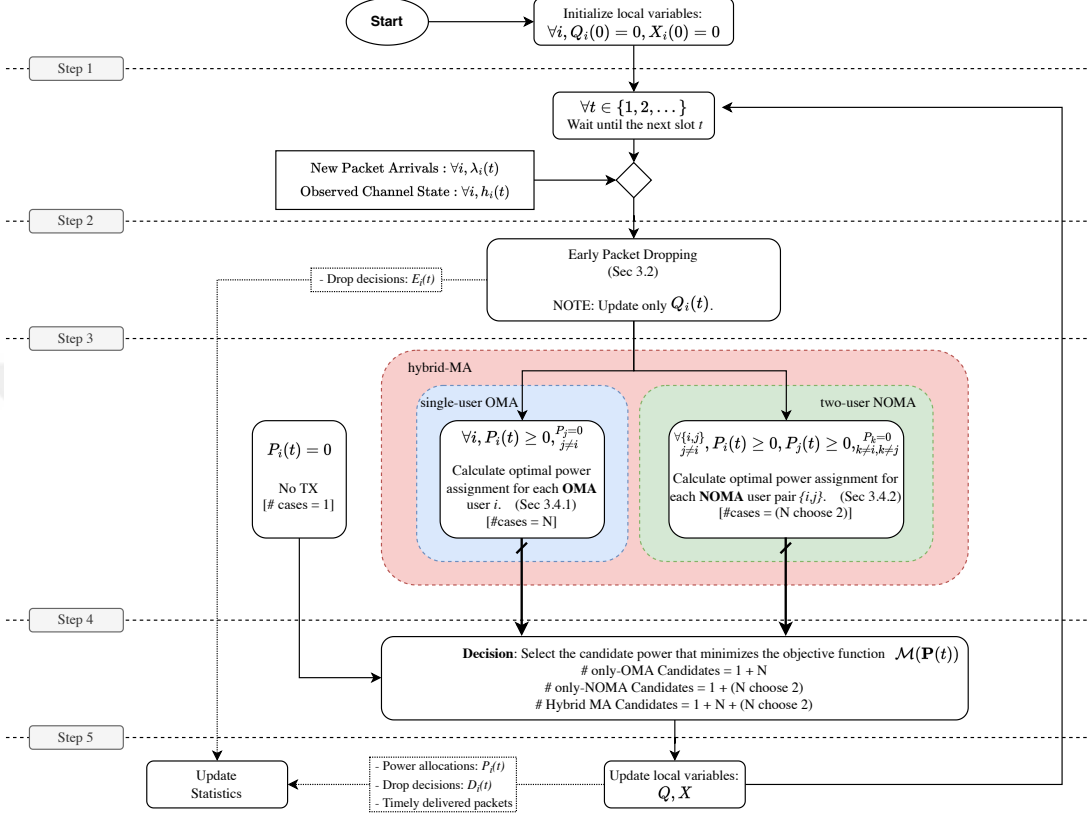


Figure 3.3: Flow chart of the proposed algorithm for timely throughput maximization using Hybrid-MA.

The Slater condition indicates that all constraints can be satisfied with a slackness value $\delta > 0$ under a decision policy [42, 71], such that $\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} P_i(\tau) \leq (\gamma_i - \delta)$, $\forall i \in \mathcal{N}$. The Slater condition also ensures the strong stability of virtual queues [42]. The *drift – plus – penalty* algorithm gives an $O(\epsilon)$ approximation to optimality under the Slater condition [42, 71]. Thus, the proper selection of the weight parameter V in order to satisfy time average constraints while minimizing the penalty is important for ensuring optimality with an $O(\epsilon)$ approximation, too. The convexity of (3.23) is analyzed in Section 3.4.

Flow chart of the proposed dynamic power allocation algorithm for timely throughput maximization with Hybrid-MA is presented in Fig 3.3. The algorithm starts with the initialization of the local variables. In Step 1, the algorithm waits for the beginning of the next slot t and then obtains new packet arrivals and

observes the channel state. In Step 2, EPD, which is presented in Section 3.2, is applied to drop packets that are detected as to-be-expired. In Step 3, we calculate the optimal power assignments for candidate OMA and NOMA transmissions. Due to our single transmitter model, there are N possible OMA transmissions and $\binom{N}{2}$ possible NOMA transmissions. Thus, the scheduling complexity is $O(N)$ for only-OMA and $O(N^2)$ for only-NOMA and hybrid-MA cases. In Step 4, the optimization metric in (3.23) is exhaustively minimized. One of the minimizing candidates is selected randomly for the sake of fairness among users. In each slot, hybrid-MA dynamically switches between OMA and NOMA by selecting the best candidate among OMA and NOMA transmissions with derived optimal power assignments. In Step 5, local variables and statistics are updated according to the given decision. Finally, the algorithm goes to Step 1 to wait for the beginning of the next slot.

In the next section, optimal power assignment derivations for OMA and NOMA are presented.

3.4 Optimal Power Assignment for Hybrid MA

In this section, optimal power assignment is derived for OMA and NOMA. Since $\mathbf{P}(t)$ is optimized slot-by-slot, in the rest of the paper the time parameter t is removed for simplification, and $\mathbf{P}$ indicates $\mathbf{P}(t)$. Similarly, the time index is removed from all other parameters that depend on time.

3.4.1 Optimal Power Assignment for OMA

In this part, optimal power assignment is explained given that OMA transmission scheme is occupied for user i , such that $\Psi_i = i$. Based on [40], departure rate $\mu_i(P_i)$ for a user i using OMA with FBL $\tau\mathcal{B}$ and BLER ϵ_O is given by:

$$\mu_i(P_i) = \tau\mathcal{B} \log_2 \left(1 + \frac{|h_i|^2 P_i}{\mathcal{B}N_0} \right) - \sqrt{\tau\mathcal{B}\mathcal{V}_i} \cdot \frac{Q^{-1}(\epsilon_O)}{\ln 2} \quad (3.24)$$

where N_0 is the noise power spectral density. Note that, we obtain a lower bound on $\mu_i(P_i)$ with the approximation of $\mathcal{V}_i \approx 1$, since $\mathcal{V}_i \leq 1$ [72]. Thus, the lower bound can be used in the solution for satisfying strict deadlines. Moreover, for URLLC which requires high reliability and low latency, SNR should be sufficiently high[72], and hence the approximation of $\mathcal{V}_i \approx 1$ is accurate in this operating region. Let P_i^{req} and P_i^{one} be the required power value for transmission of q_i bits and 1 bit, respectively, as

$$\mu_i(P_i^{req}) \triangleq q_i \quad (3.25)$$

$$\mu_i(P_i^{one}) \triangleq 1 \quad (3.26)$$

Moreover, let P_i^{min} be the minimum required power to successfully perform an OMA transmission under deadline constraint, so that, at least 1 and q_i bits must be transmitted for $d_i > 1$ and $d_i = 1$ cases, respectively. Let P_i^{max} be the maximum power value that can be assigned for OMA transmission user i . Then, P_i^{min} and P_i^{max} can be defined as follows:

$$P_i^{min} \triangleq P_i^{req} \mathbb{1}\{d_i = 1\} + P_i^{one} \mathbb{1}\{d_i > 1\} \quad (3.27)$$

$$P_i^{max} \triangleq \min(P_i^{req}, P_0) \quad (3.28)$$

Let $\mathcal{P}_i$ be the available power region for an OMA transmission of user i . $\mathcal{P}_i$ is $\{0\} \cup [P_i^{min}, P_i^{max}]$ when $P_i^{min} \leq P_i^{max}$. Otherwise, $\mathcal{P}_i$ is $\{0\}$, since the respective packet will definitely be dropped when $d_i = 1$. Thus, we avoid the waste of power for transmitting a fraction of a to-be-expired packet with OMA by constraining $\mathcal{P}_i$ properly. $\mathcal{P}_i$ is presented as,

$$\mathcal{P}_i = \begin{cases} \{0\} \cup [P_i^{min}, P_i^{max}] & , \text{ if } P_i^{min} \leq P_i^{max} \\ \{0\} & , \text{ otherwise} \end{cases} \quad (3.29)$$

Let the objective function for user i under OMA transmission scheme be $\mathcal{M}_i^O(P_i)$ and the resulting OMA power optimization problem is stated as:

$$P_i^O = \arg \min_{P_i} \mathcal{M}_i^O(P_i) \triangleq \mathcal{M}(\mathbf{P}) \quad (3.30a)$$

$$\text{s.t. } P_i \in \mathcal{P}_i, \quad \begin{matrix} P_j=0 \\ \forall j \neq i \end{matrix} \quad (3.30b)$$

where P_i^O is the optimal power value for user i under OMA transmission scheme. The solution to the optimization problem above is presented in Theorem 3.4.1.

The occupied TM is represented as $\big|_{\phi}$ within equations for simplification. Thus, FTM and CTM are indicated as $\big|_{\Phi_F}$ and $\big|_{\Phi_C}$, respectively.

Theorem 3.4.1. *Optimal power allocation for OMA can be achieved with FTM. P_i^O is given by:*

$$P_i^O = \begin{cases} P_i' & , \text{if } \mathcal{M}_i^O(0)\big|_{\Phi_F} > \mathcal{M}_i^O(P_i')\big|_{\Phi_F} \\ 0 & , \text{otherwise} \end{cases} \quad (3.31)$$

where

$$P_i' = \begin{cases} P_i^{\max} & , \text{if } P_i^{\max} \leq P_i^* \\ P_i^* & , \text{if } P_i^{\min} \leq P_i^* < P_i^{\max} \\ P_i^{\min} & , \text{if } P_i^* < P_i^{\min} \end{cases} \quad (3.32a)$$

$$P_i^* = \Gamma_i / (X_i \ln 2) - (\mathcal{B}N_0) / |h_i|^2 \quad (3.32b)$$

$$\Gamma_i = V(1 - (d_i - 1)/m_i)^{\alpha_i} (\tau \mathcal{B}) / q_i \quad (3.32c)$$

Proof. Assume that transmission is performed using OMA, such that $P_i \in [P_i^{\min}, P_i^{\max}]$ and $\forall j \neq i, P_j = 0$. The FTM based objective function $\mathcal{M}_i^O(P_i)\big|_{\Phi_F}$ can be written as

$$\begin{aligned} \mathcal{M}_i^O(P_i)\big|_{\Phi_F} &= -V \left(\frac{m_i - (d_i - 1)}{m_i} \right)^{\alpha_i} \frac{\tau \mathcal{B}}{q_i} \log_2 \left(1 + \frac{|h_i|^2 P_i}{\mathcal{B}N_0} \right) + X_i P_i + C_i^O \\ &= -\Gamma_i \cdot \log_2 \left(1 + \frac{|h_i|^2 P_i}{\mathcal{B}N_0} \right) + X_i P_i + C_i^O \end{aligned} \quad (3.33)$$

where C_i^O is a constant given by

$$C_i^O = V \sum_{j=1}^N \left(\frac{m_j - (d_j - 1)}{m_j} \right)^{\alpha_i} + \sum_{j=1}^N X_j (-\gamma_j) + \Gamma_i \sqrt{\frac{\mathcal{V}_i}{\tau \mathcal{B}}} \frac{Q^{-1}(\epsilon^O)}{\ln 2} \quad (3.34)$$

Note that $\mathcal{M}_i^O(P_i \in \mathcal{P}_i)\big|_{\Phi_F} \geq 0$ and the reduction ratio in (3.17) is non-negative for $P_i \in [P_i^{\min}, P_i^{\max}]$. It can be shown that $\mathcal{M}_i^O(P_i)\big|_{\Phi_F}$ is convex for $P_i \in [P_i^{\min}, P_i^{\max}]$. The global minimizer P_i^* for FTM is given as,

$$P_i^* = \frac{\Gamma_i}{X_i \ln 2} - \frac{\mathcal{B}N_0}{|h_i|^2}. \quad (3.35)$$

The CTM based objective function $\mathcal{M}_i^O(P_i)\big|_{\Phi_C}$ can be written as

$$\mathcal{M}_i^O(P_i)\big|_{\Phi_C} = V \left(\frac{m_i - (d_i - 1)}{m_i} \right)^{\alpha_i} \cdot (\mathbb{1}\{P_i < P_i^{req}\} - 1) + X_i P_i + C_i^O. \quad (3.36)$$

Note that FTM and CTM based objective functions in (3.33) and (3.36), respectively, are equal for $P_i \in \{0, P_i^{req}\}$. CTM based objective function in (3.36) increases linearly with P_i for $P_i \in [0, P_i^{req}]$. Since FTM based objective function in (3.33) is convex for $P_i \in [P_i^{\min}, P_i^{\max}]$, we can conclude that $\mathcal{M}_i^O(P_i)\big|_{\Phi_F} \leq \mathcal{M}_i^O(P_i)\big|_{\Phi_C}$ for $P_i \in [0, P_i^{\max}]$. Thus, FTM is always a better option in terms of optimizing power assignment for OMA over the all available power region. P_i^O in (3.31) can be obtained using P_i^* and available power region $P_i \in \mathcal{P}_i$. $\square$

3.4.2 Optimal Power Assignment for NOMA

In this section, power assignment is explained given that NOMA transmission scheme is occupied for the paired users $\{i, j\}$, such that $\Psi_i = j$ and $\Psi_j = i$. Assume that $|h_j|^2 > |h_i|^2$, thus, the channel of user i is weaker. In theory, allocation of higher power to the user with weaker channel is not necessary in NOMA [73]. Therefore, allocated power values are only constrained by with total power budget: $P_i + P_j \leq P_0$. Based on [40], departure rates $\mu_{i,(i,j)}(P_i, P_j)$ and $\mu_{j,(i,j)}(P_j)$ with FBL $\tau\mathcal{B}$ and BLER ϵ_N are given by:

$$\mu_{i,(i,j)} = \tau B \log_2 \left(1 + \frac{|h_i|^2 P_i}{|h_i|^2 P_j + B N_0} \right) - \sqrt{\tau \mathcal{B} \mathcal{V}_i} \frac{Q^{-1}(\epsilon_N)}{\ln 2} \quad (3.37)$$

$$\mu_{j,(i,j)} = \tau B \log_2 \left(1 + \frac{|h_j|^2 P_j}{B N_0} \right) - \sqrt{\tau \mathcal{B} \mathcal{V}_j} \frac{Q^{-1}(\epsilon_N)}{\ln 2} \quad (3.38)$$

where the channel dispersion parameters are approximated as $\mathcal{V}_i \approx 1$ and $\mathcal{V}_j \approx 1$, which is valid in the high SNR regime. Similarly, we obtain a lower bound on $\mu_{i,(i,j)}$ and $\mu_{j,(i,j)}$ with these approximations [72]. The lower bounds can be used in the solution for satisfying strict deadlines with NOMA under FBL regime. Let $\{P_{j,(i,j)}^{req}, P_{j,(i,j)}^{one}\}$ and $\{P_{i,(i,j)}^{req}, P_{i,(i,j)}^{one}\}$ be the required power values for transmitting

$\{q_j, 1\}$ and $\{q_i, 1\}$ bits, respectively, of user pair $\{i, j\}$ within a slot under the NOMA transmission scheme. Moreover, let $P_{k,(i,j)}^{min}$, $k \in \{i, j\}$ be the minimum required power to successfully perform a NOMA transmission under deadline constraint, as follows:

$$P_{k,(i,j)}^{min} = P_{k,(i,j)}^{req} \mathbb{1}\{d_k = 1\} + P_{k,(i,j)}^{one} \mathbb{1}\{d_k > 1\} \quad (3.39)$$

Thus, we have,

$$\mu_{j,(i,j)}(P_{j,(i,j)}^{req}) = q_j \quad (3.40)$$

$$\mu_{j,(i,j)}(P_{j,(i,j)}^{one}) = 1 \quad (3.41)$$

$$\mu_{i,(i,j)}(P_{i,(i,j)}^{req}, P_{j,(i,j)}^{min}) = q_i \quad (3.42)$$

$$\mu_{i,(i,j)}(P_{i,(i,j)}^{one}, P_{j,(i,j)}^{min}) = 1 \quad (3.43)$$

Let $P_{(i,j)}^{max}$ and $P_{(i,j)}^{min}$ be the maximum and minimum values that can be assigned for the NOMA transmission, respectively, as

$$P_{(i,j)}^{max} \triangleq \min(P_0, (P_{j,(i,j)}^{req} + P_{i,(i,j)}^{req})) \quad (3.44)$$

$$P_{(i,j)}^{min} \triangleq (P_{j,(i,j)}^{min} + P_{i,(i,j)}^{min}) \quad (3.45)$$

Let $\mathcal{P}_{(i,j)}$ be the available total power region for a NOMA transmission of user pair $\{i, j\}$. $\mathcal{P}_{(i,j)}$ is $\{0\} \cup [P_{(i,j)}^{min}, P_{(i,j)}^{max}]$ when $P_{(i,j)}^{min} \leq P_{(i,j)}^{max}$. Otherwise, $\mathcal{P}_{(i,j)}$ is $\{0\}$, since the respective packet will definitely be dropped when $d_i = 1$ or $d_j = 1$. Thus, we avoid waste of power for transmitting a fraction of a to-be-expired packet with NOMA by constraining $\mathcal{P}_{(i,j)}$ properly. $\mathcal{P}_{(i,j)}$ is presented as,

$$\mathcal{P}_{(i,j)} = \begin{cases} \{0\} \cup [P_{(i,j)}^{min}, P_{(i,j)}^{max}] & , \text{ if } P_{(i,j)}^{min} \leq P_{(i,j)}^{max} \\ \{0\} & , \text{ otherwise} \end{cases} \quad (3.46)$$

Let $\mathcal{M}_{(i,j)}^N(P_i, P_j) = \mathcal{M}(\mathbf{P})$ where $P_k = 0$, $\forall k \neq \{i, j\}$ be the objective function for $\{i, j\}$ under NOMA scheme. Assume that FTM is selected. Then, the optimization problem of power assignment for FTM based NOMA transmission of user pair $\{i, j\}$ can be expressed as:

$$\{P_{i,(i,j)}^N|_{\Phi_F}, P_{j,(i,j)}^N|_{\Phi_F}\} = \arg \min_{P_i, P_j} \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F} \quad (3.47a)$$

$$s.t. \ P_i + P_j \in \mathcal{P}_{(i,j)} \quad (3.47b)$$

where $\{P_{i,(i,j)}^N|_{\Phi_F}, P_{j,(i,j)}^N|_{\Phi_F}\}$ are optimal power values under FTM based NOMA transmission schemes for users i and j , respectively. $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F}$ can be expressed as:

$$\begin{aligned} \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F} = & -\Gamma_i \log_2 \left(1 + \frac{|h_i|^2 P_i}{|h_i|^2 P_j + \mathcal{B}N_0} \right) \\ & -\Gamma_j \log_2 \left(1 + \frac{|h_j|^2 P_j}{\mathcal{B}N_0} \right) + X_i P_i + X_j P_j + C_{(i,j)}^N \end{aligned} \quad (3.48)$$

where $C_{(i,j)}^N$ is a constant. Note that $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F} \geq 0$ for $P_k \geq P_{k,(i,j)}^{\min}$, $k \in \{i, j\}$. Moreover, the reduction ratio in (3.17) is non-negative for $P_k \in [P_{k,(i,j)}^{\min}, P_{k,(i,j)}^{\max}]$, $k \in \{i, j\}$. Since $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F}$ in (3.48) is not convex, the optimization problem in (3.47) is not convex either. In order to solve this problem, an auxiliary variable $\theta = P_i + P_j$ is introduced and the solution process of the problem is divided into two consecutive sub-problems. At first, the optimal value of θ is calculated. Then, the optimal value of P_j is calculated for given θ . The first sub-problem is to find the optimal value of θ as follows:

$$\theta^N = \arg \min_{\theta} g(\theta, P_j) \quad (3.49a)$$

$$s.t. \theta \in [P_{(i,j)}^{\min}, P_{(i,j)}^{\max}] \quad (3.49b)$$

where θ^N is the optimal value and $g(\theta, P_j)$ is given as follows:

$$\begin{aligned} g(\theta, P_j) = & -\Gamma_i \log_2 \left(\frac{|h_i|^2 \theta + \mathcal{B}N_0}{|h_i|^2 P_j + \mathcal{B}N_0} \right) + X_i \theta \\ & -\Gamma_j \log_2 \left(1 + \frac{|h_j|^2 P_j}{\mathcal{B}N_0} \right) + (X_j - X_i) P_j + C_{(i,j)}^N \end{aligned} \quad (3.50)$$

The solution for the optimal θ^N is presented below.

Theorem 3.4.2. *Optimal θ^N for FTM based NOMA transmission scheme is given by:*

$$\theta^N = \begin{cases} P_{(i,j)}^{\max} & , \text{ if } P_{(i,j)}^{\min} \leq P_{(i,j)}^{\max} \leq \theta^* \\ \theta^* & , \text{ if } P_{(i,j)}^{\min} \leq \theta^* \leq P_{(i,j)}^{\max} \\ P_{(i,j)}^{\min} & , \text{ if } \theta^* \leq P_{(i,j)}^{\min} \leq P_{(i,j)}^{\max} \end{cases} \quad (3.51)$$

where

$$\theta^* = \frac{\Gamma_i}{X_i \cdot \ln 2} - \frac{\mathcal{B} \cdot N_0}{|h_i|^2} \quad (3.52)$$

Proof. It can be shown that $g(\theta, P_j)$ is convex for $\theta \in [P_{(i,j)}^{min}, P_{(i,j)}^{max}]$. The global minimizer θ^* is given as,

$$\theta^* = \frac{\Gamma_i}{X_i \ln 2} - \frac{\mathcal{B}N_0}{|h_i|^2}. \quad (3.53)$$

θ^N in (3.51) can be obtained using θ^* and available total power region $\theta \in \mathcal{P}_{(i,j)}$. $\square$

The second sub-problem is to find the optimal value of P_j , given the auxiliary variable θ^N , as follows:

$$P'_{j,(i,j)}|_{\Phi_F} = \arg \min_{P_j} g(\theta^N, P_j) \quad (3.54a)$$

$$s.t. \ P_i + P_j = \theta^N \quad (3.54b)$$

where $P'_{j,(i,j)}|_{\Phi_F}$ is the optimal value. The solution of the second sub-problem to find $P'_{j,(i,j)}|_{\Phi_F}$ is presented below.

Theorem 3.4.3. Suppose that $\Gamma_j \geq \Gamma_i$. Then, the optimal $P'_{j,(i,j)}$ for FTM based NOMA transmission scheme for $\theta \in [P_{(i,j)}^{min}, P_{(i,j)}^{max}]$ is given by:

$$P'_{j,(i,j)}|_{\Phi_F} = \begin{cases} P_j^{min} & , \text{if } \frac{dg(P_j)}{dP_j}|_{P_j=P_j^{min}} \geq 0 \\ P_j^\dagger & , \text{if } \frac{dg(P_j)}{dP_j}|_{P_j=P_j^\dagger} \leq 0 \\ P_j^\ddagger & , \text{otherwise} \end{cases} \quad (3.55)$$

where $P_j^\dagger = \min((\theta - P_{i,(i,j)}^{min}), P_{j,(i,j)}^{req})$, $\frac{d}{dP_j} g(P_j)|_{P_j=P_j^*} = 0$ and $P_j^\ddagger = \min(P_j^*, P_j^\dagger)$.

Proof. It can be shown that $g(\theta^N, P_j)$ is convex for $\Gamma_j \geq \Gamma_i$. The minimizer P_j^* is given as

$$P_j^* = \begin{cases} \mathcal{B}N_0 \left(\frac{\Gamma_j |h_j|^2 + \Gamma_i |h_i|^2}{(\Gamma_i - \Gamma_j) |h_i|^2 |h_j|^2} \right) & , \text{if } X_i = X_j \\ \max(\zeta_1, \zeta_2) & , \text{otherwise} \end{cases} \quad (3.56)$$

where ζ_1 and ζ_2 are the roots of the polynomial $\mathbf{a} \cdot P_j^2 + \mathbf{b} \cdot P_j + \mathbf{c} = 0$ such that

$$\mathbf{a} = \ln 2 (X_j - X_i) \quad (3.57a)$$

$$\mathbf{b} = \Gamma_i - \Gamma_j + \ln 2 (X_j - X_i) \mathcal{B}N_0 \left(\frac{1}{|h_i|^2} + \frac{1}{|h_j|^2} \right) \quad (3.57b)$$

$$\mathbf{c} = \mathcal{B}N_0 \left(\frac{\Gamma_i}{|h_i|^2} - \frac{\Gamma_j}{|h_j|^2} \right) + \frac{(X_j - X_i) \ln 2 (\mathcal{B}N_0)^2}{|h_i|^2 + |h_j|^2} \quad (3.57c)$$

Under $\Gamma_j \geq \Gamma_i$ condition, $P'_{j,(i,j)}|_{\Phi_F}$ in (3.55) can be obtained using P_j^* and available power region $P_j \in \{0\} \cup [P_{j,(i,j)}^{min}, P_j^\dagger]$. $\square$

The CTM based objective function $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$ can be written as:

$$\begin{aligned} \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C} = & X_i P_i + X_j P_j + C_{(i,j)}^N \\ & + \sum_{k \in \{i,j\}} \left(\frac{\Gamma_k q_k}{\tau \mathcal{B}} \right) \left(\mathbb{1} \{ P_k < P_{k,(i,j)}^{req} \} - 1 \right) \end{aligned} \quad (3.58)$$

Note that $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$ when the pair $\{P_i, P_j\}$ is $\{0, 0\}$ or $\{P_{i,(i,j)}^{req}, P_{j,(i,j)}^{req}\}$. Moreover, $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$ increases linearly with P_i and P_j . Since FTM based objective function in (3.48) is convex for $\Gamma_j \geq \Gamma_i$, we can conclude that $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_F} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$ for $\Gamma_j \geq \Gamma_i$. Therefore, Theorem 3.4.3 provides optimal power $P'_{j,(i,j)}$ under the assumption that $\Gamma_j \geq \Gamma_i$. The solution for remaining cases $\Gamma_j < \Gamma_i$ is derived using the approach proposed in [38]. The optimization problem of power assignment for CTM based NOMA transmission can be expressed as:

$$\arg \min_{P_i, P_j} \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C} \quad (3.59a)$$

$$s.t. \ P_i + P_j \in \mathcal{P}_{(i,j)} \quad (3.59b)$$

where $\{P'_{i,(i,j)}|_{\Phi_C}, P'_{j,(i,j)}|_{\Phi_C}\}$ are optimal power values under CTM based NOMA transmission schemes for users i and j , respectively. The solution of the respective optimization problem in (3.59) is given below.

Theorem 3.4.4. *Let users $\{i, j\}$ be such that $|h_j| > |h_i|$. The optimal $P'_{j,(i,j)}|_{\Phi_C}$ and $P'_{i,(i,j)}|_{\Phi_C}$ for NOMA transmission scheme under CTM is given by: When*

$$(P_{i,(i,j)}^{req} + P_{j,(i,j)}^{req} \leq P_{(i,j)}^{\max}),$$

$$\{P'_{i,(i,j)}|_{\Phi_C}, P'_{j,(i,j)}|_{\Phi_C}\} = \{P_{i,(i,j)}^{req}, P_{j,(i,j)}^{req}\}, \quad (3.60)$$

otherwise,

$$\{P'_{i,(i,j)}|_{\Phi_C}, P'_{j,(i,j)}|_{\Phi_C}\} = \{0, 0\}. \quad (3.61)$$

Proof. Note that $P_j^{req} = P_{j,(i,j)}^{req}$ and $P_i^{req} < P_{i,(i,j)}^{req}$. We have,

$$\mathcal{M}_j^O(P_j)|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C} \text{ when } P_i < P_{i,(i,j)}^{req} \text{ and } P_j = P_{j,(i,j)}^{req} = P_j^{req} \quad (3.62)$$

$$\mathcal{M}_i^O(P_i)|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C} \text{ when } P_i = P_{i,(i,j)}^{req} > P_i^{req} \text{ and } P_j < P_{j,(i,j)}^{req}. \quad (3.63)$$

Then, we can conclude that CTM based OMA is better than CTM based NOMA when only one of users' packet is completed in a NOMA pair. Therefore, these cases can be ignored for CTM based NOMA in the Hybrid MA scenario. Moreover, since $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$ linearly increases with P_i and P_j , we can conclude that $\mathcal{M}_{(i,j)}^N(0, 0)|_{\Phi_C} \leq \mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$ when $P_j < P_{j,(i,j)}^{req}$ and $P_i < P_{i,(i,j)}^{req}$. Therefore, we only need to consider the binary decision of completely serving both packets or not transmitting them at all for CTM based NOMA in the Hybrid MA scenario. We can determine the optimal decision by selecting one of $\{0, 0\}$ and $\{P_{i,(i,j)}^{req}, P_{j,(i,j)}^{req}\}$ power allocation options for minimizing $\mathcal{M}_{(i,j)}^N(P_i, P_j)|_{\Phi_C}$. The decision in (3.60)-(3.61) can be obtained under total power constraint, such that $P_i + P_j \leq P_{(i,j)}^{\max}$. $\square$

In this thesis, power allocation decision under NOMA transmission scheme in a Hybrid MA scenario for a user pair $\{i, j\}$, such that $|h_j| > |h_i|$, is as follows:

$$\{P_{i,(i,j)}^N, P_{j,(i,j)}^N\} = \begin{cases} \{P'_{i,(i,j)}, P'_{j,(i,j)}\} & , \text{if } \mathcal{M}^N(0, 0) > \mathcal{M}^N(P'_{i,(i,j)}, P'_{j,(i,j)}) \\ \{0, 0\} & , \text{otherwise} \end{cases} \quad (3.64)$$

where $\{P'_{i,(i,j)}, P'_{j,(i,j)}\}$ is as follows:

$$\{P'_{i,(i,j)}, P'_{j,(i,j)}\} = \begin{cases} \{(\theta^N - P'_{j,(i,j)}|_{\Phi_F}), P'_{j,(i,j)}|_{\Phi_F}\} & , \text{if } (\Gamma_j \geq \Gamma_i) \\ \{P'_{i,(i,j)}|_{\Phi_C}, P'_{j,(i,j)}|_{\Phi_C}\} & , \text{otherwise.} \end{cases} \quad (3.65)$$

In (3.65), we have $\theta^N \in [P_{(i,j)}^{min}, P_{(i,j)}^{max}]$, as presented in Theorem 3.4.3. Therefore, the resultant $\{P'_{i,(i,j)}, P'_{j,(i,j)}\}$ in (3.65) cannot be directly used as the final power allocation decision. In (3.64), we compare $\mathcal{M}^N(P'_{i,(i,j)}, P'_{j,(i,j)})$ with the $\theta^N = 0$ case, which is denoted as $\mathcal{M}^N(0, 0)$. Thus, we fully cover $\mathcal{P}_{(i,j)}$ in (3.64) to make our final decision for $\{P_{i,(i,j)}^N, P_{j,(i,j)}^N\}$, which is the power allocation decision under NOMA transmission scheme.

In the following section, the computational load reduction heuristic is presented.

3.5 Computational Load Reduction Heuristic

The computational load of finding the best Hybrid-MA transmission scheme consists of exhaustively searching among N candidate single-user OMA transmissions and $\binom{N}{2}$ candidate two-user NOMA transmissions, which are illustrated in Steps 3 and 4 of the proposed algorithm's flow chart in Figure 3.3, respectively. In this search operation, a big part of the computational load comes from candidates for two-user NOMA transmission. As N increases, the number of NOMA candidates $\binom{N}{2}$ significantly increases compared to the number of OMA candidates N . This situation motivated us to propose a heuristic to reduce the number of candidates considered for two-user NOMA transmission while preserving timely throughput as high as possible. So that, the overall computational load of the algorithm can be reduced at the cost of a small sacrifice in the timely throughput performance.

There are studies focusing on user group scheduling problem for NOMA in the literature. In [74], a NOMA user group scheduling algorithm under temporal fairness constraint among users is proposed. In [75], proportional fairness is considered for this purpose. In [45], deep learning techniques are leveraged for a power-efficient NOMA user grouping problem under a deadline-aware problem setting. In [76], the authors propose an opportunistic user pair scheduling algorithm, which both considers the effect of users' channel states and target rates,

for improving outage performance of a downlink NOMA scenario. In [77], optimal scheduling is derived using mixed integer optimization techniques. These studies indicate that user pair scheduling is an important problem for NOMA. In this thesis, we propose a simple and effective user pairing heuristic to reduce candidate number of NOMA user pairs in order to reduce computational load.

The purpose of our proposed approach is to reduce the search space for finding the best NOMA user pair among $\binom{N}{2}$ candidates. For this purpose, we need to find a way to quantify the likelihood of a user for being considered as a NOMA candidate with respect to other users. We made empirical experiments using the objective function in (3.23) to find a relation with the users, which are selected for the NOMA transmission. In these experiments, we observed a correlation between the optimal OMA performance of a user compared to others and selection of the user for the optimal NOMA user pair. Note that, the OMA objective function in (3.33) takes into consideration both the penalty due to packet dropping and the average power constraint. Therefore, users' optimal performance with the OMA objective function on (3.33) can be used for quantifying the likelihood of a user for being considered as a NOMA transmission candidate regarding the other users.

These observations inspired us to develop a computational load reduction technique based on the optimal OMA performances of all users. We propose a heuristic to determine a subset of $Y < N$ users among which a NOMA user pair can be formed. Thus, this approach will reduce the search space from $\binom{N}{2}$ to $\binom{Y}{2}$. Note that, we do not reduce the order of complexity, instead, we reduce the computational load. Let the reduced set of users be denoted as $\mathcal{H}$. In order to determine the candidate subset of users for NOMA, we sort the users according to their optimal OMA performance in (3.33), and we select the best Y number of them. This method is denoted as *BEST*, which refers to the set of BEST users according to their optimal OMA performances. The pseudocode for the *BEST* heuristic is presented in Algorithm 2.

At each slot t , $\mathcal{M}_i^O(P_i^O)\big|_{\Phi_F}$ values are calculated first. Note that, $\mathcal{M}_i^O(P_i^O)\big|_{\Phi_F}$ values must be calculated to determine the best OMA transmission scheme; thus,

Algorithm 2 Pseudocode for *BEST* heuristic

Inputs: $\mathcal{M}_i^O(P_i^O)\big|_{\Phi_F}, \forall i \in \mathcal{N}, Y$

Outputs: $\mathcal{H}$

Operation:

- 1: $\mathcal{H} \leftarrow$ Sort the users $i \in \mathcal{N}$ in the increasing order of $\mathcal{M}_i^O(P_i^O)\big|_{\Phi_F}$
 $\triangleright$ The order of users with the same $\mathcal{M}_i^O(P_i^O)\big|_{\Phi_F}$ value are randomly shuffled.
 - 2: $\mathcal{H} \leftarrow$ Only first Y elements of $\mathcal{H}$
 - 3: **return** $\mathcal{H}$
-

we use already available values for this step. Secondly, the *BEST* heuristic is run to derive $\mathcal{H}$, which consists of $Y \leq N$ users. Then, optimal NOMA power assignments are derived only for user pairs from $\mathcal{H}$ and the user pair among Y candidates that minimizes the objective function in (3.23) is selected as the decision for NOMA transmission.

In order to analyze the effectiveness of the proposed heuristic *BEST*, we propose another heuristic called *RAND*, which randomly selects users for $\mathcal{H}$. The pseudocode for the *RAND* heuristic is presented in Algorithm 3.

Algorithm 3 Pseudocode for *RAND* heuristic

Inputs: Y

Outputs: $\mathcal{H}$

Operation:

- 1: $\mathcal{H} \leftarrow$ Randomly select Y users from $\mathcal{N}$
 - 2: **return** $\mathcal{H}$
-

In the next chapter, numerical results and discussions for Hybrid-MA are presented.

Chapter 4

Numerical Results and Discussions for Hybrid MA

In this chapter, we comparatively evaluate the MA performances in terms of timely throughput. Let α , γ , m , π be the system parameters for the RDPPA algorithm, average power constraint, deadline and packet arrival rate, respectively. For the simulations, let $\alpha = \alpha_i(t)$, $\gamma = \gamma_i$, $m = m_i$, $\pi = \pi_i$ for all t and $i \in \mathcal{N}$. In this case, timely throughput can be represented as $\pi - \overline{D}$. The target BLER is considered as 10^{-5} . Due to SIC, NOMA related target BLER is set as $\epsilon_N = 5 \cdot 10^{-6}$, so that overall system target BLER is ensured [35]. A cellular model which is considered by Choi et. al [41] in a study for satisfying latency constraints under energy efficiency objective in IoT networks with Hybrid MA, is used in this paper. A channel model for IoT applications in urban areas [41, 78] is used in this study. The default values of all parameters used in obtaining the numerical results are given in Table 4.1, unless otherwise stated. The simulations are implemented using MATLAB. In order to assess fairness among $\overline{D}_i$, we considered Jain's Fairness Index (FI) [79] as

$$FI = \frac{(\sum_i \overline{D}_i)^2}{(N \sum_i \overline{D}_i^2)}. \quad (4.1)$$

Table 4.1: Simulation parameters and default values

V (Weight parameter)	100
Λ (Packet size)	$(160 \cdot w)_{w=1}^{20}$ bits
α	0.1
γ (Average power constraint for all users)	0.6 W
π (Arrival rate for all users)	0.3
R (Cell radius)	50 m
$\#Slots$ (Simulation slot count)	10^4
τ (Slot duration)	0.1 ms
$\mathcal{B}$ (Bandwidth)	1 MHz
P_0 (Power budget)	3 W
N (User count)	5
m (Deadline slot count)	5
Path loss	$35.3 + 37.6 \log_{10}(r_i)$ dB
Fast fading component	$CN(0, 1)$
N_0 (Noise power spectral density)	-174 dBm/Hz
ϵ_O (OMA Target BLER)	10^{-5}
ϵ_N (NOMA Target BLER)	$5 \cdot 10^{-6}$

Let $\mathbb{D}^O(\pi, N)$ represent the set of all achievable per user drop rates using OMA for a given $\{\pi, N\}$ pair. Then, the lower bound $\inf \mathbb{D}^O(\pi, N) = \max(0, \pi N - 1)/N$ is the minimum average unserved arrival rate per user. As N increases, $\inf \mathbb{D}^O(\pi, N)$ increases towards π . Therefore, in order to clearly analyze the impact of other parameters on the system, we select $N = 5$ in Table 4.1. The problem becomes intractable for low values of m due to the resultant unavoidable unfairness among users. For high values of m , the problem becomes relaxed. In accordance with the selection for N , we select $m = 5$ in Table 4.1. The algorithms proposed in this thesis are as follows. FTM-based OMA is indicated as soft, others as hard.

- opt. soft-OMA (s-OMA): Optimal power assignment for OMA with FTM is considered and Theorem 3.4.1 is used. This algorithm performs RDPPA.
- opt. Hybrid-MA (H-MA): Optimal power assignment for Hybrid MA is considered. The content of “Hybrid” consists of the NOMA and OMA using (3.64) and Theorem 3.4.1, respectively. This algorithm performs RDPPA.
- opt. NOMA: Optimal power assignment for only NOMA is considered,

using (3.64). This algorithm performs RDPPA.

The following algorithms are compared with the above algorithms as baseline references:

- opt. hard-OMA [38] (h-OMA): Optimal power assignment for OMA with CTM is considered. The CTM based objective function (3.36) is minimized. The dynamic power allocation approach presented in [38] is used.
- EDF[39]: It uniformly selects one of the users with the shortest remaining expiration time and performs OMA using available power budget without any power constraint.
- P_{inf} -OMA (p-OMA): This algorithm performs OMA using an infinite power budget P_{inf} without any power constraint for a uniformly selected user. p-OMA drop rate achieves $\inf \mathbb{D}^O(\pi, N)$ for any given $\{\pi, N\}$.

In order to analyse the potential of EPD mechanism, firstly the results without EPD are presented in Section 4.1. In this case, EPD mechanism in Step 2 of the flow chart of the algorithm given in Figure 3.3 is ignored by setting $S < 0$, thus $E_i(t) = 0, \forall i, t$ in this case. The system parameters α and V are optimized in this setting. Secondly, EPD will be applied along with the optimized system parameters α and V . The purpose is to present the benefit provided by EPD for improving the timely throughput. The results with the EPD mechanism are presented in Section 4.2.

4.1 Dynamic Power Allocation without EPD

In Figs. 4.1, 4.2, 4.3, $\bar{D}$ with different values of α under $\pi \in \{0.1, 0.3, 0.5\}$ are presented, respectively. Moreover, the results with $V \in \{10, 100, 1000\}$ are presented. Since h-OMA, EDF and p-OMA algorithms are independent of α , results

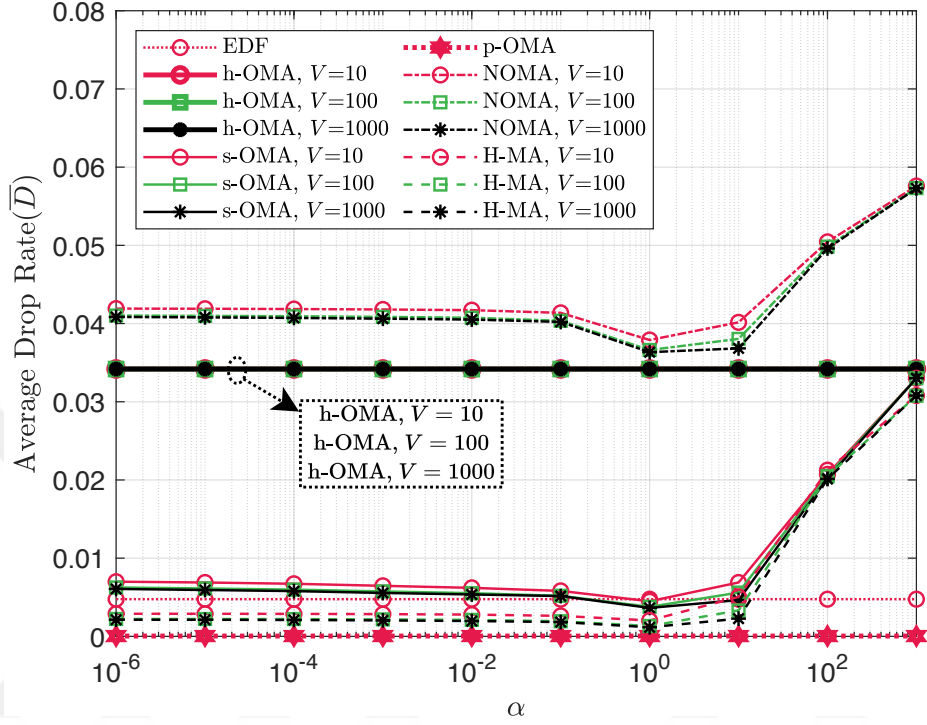


Figure 4.1: Average drop rate as a function of α with $\pi = 0.1$.

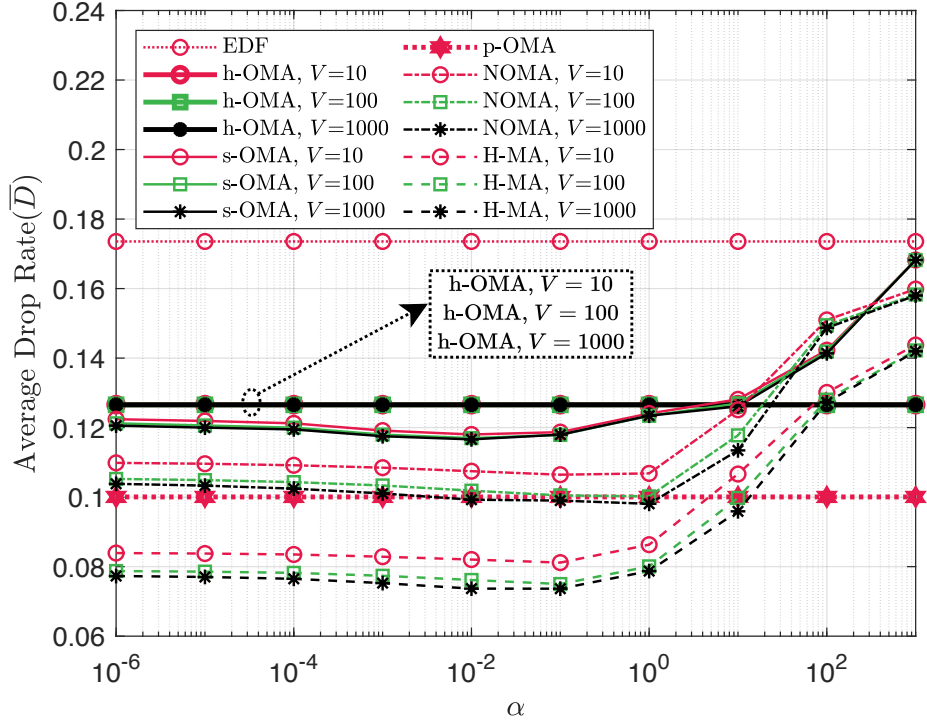


Figure 4.2: Average drop rate as a function of α with $\pi = 0.3$.

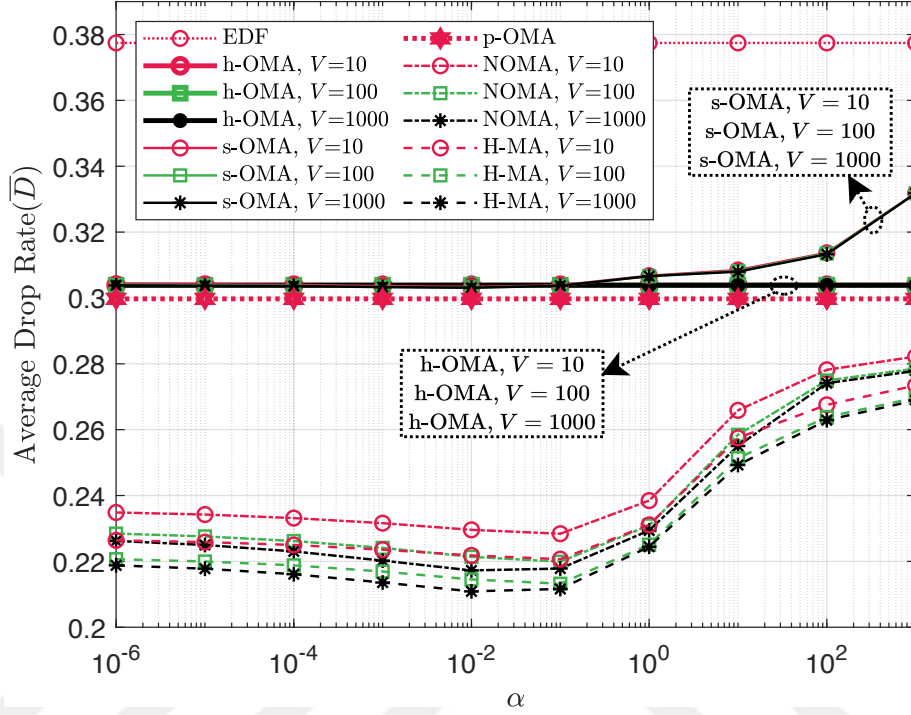


Figure 4.3: Average drop rate as a function of α with $\pi = 0.5$.

for each of these algorithms are the same for all α values. Simulations are performed for 1000 random seeds and their averages are reported.

Observations on Figs. 4.1, 4.2, 4.3 show the consistent relation between α and $\bar{D}$ under different traffic levels. Moreover, $\bar{D}$ converges at around $V = 100$ and increasing V towards 1000 does not make much difference on $\bar{D}$. Therefore, $V = 100$ is considered as the most suitable choice in terms of the tradeoff between $\bar{D}$ and γ . In Fig. 4.1, p-OMA can fully serve the arrival rate of $\pi = 0.1$. H-MA and s-OMA perform close to it with $\alpha = 1$. In this low traffic level, EDF performs well, too. Interestingly, h-OMA and NOMA perform considerably worse compared to other algorithms. The reason is that, low arrival rate $\pi = 0.1$ results in a smaller number of packets with small sizes. In such a case, binary decision of CTM is not a good choice and the existence probability of a packet pair suitable for an effective NOMA transmission is reduced. In Fig. 4.2 with $\pi = 0.3$, p-OMA cannot fully serve the arriving traffic. However, H-MA with $\alpha = 0.1$ outperforms the minimum achievable $\bar{D}$ with OMA. Since the packet diversity increases with the increasing arrival rate, h-OMA starts to perform close to the best performance

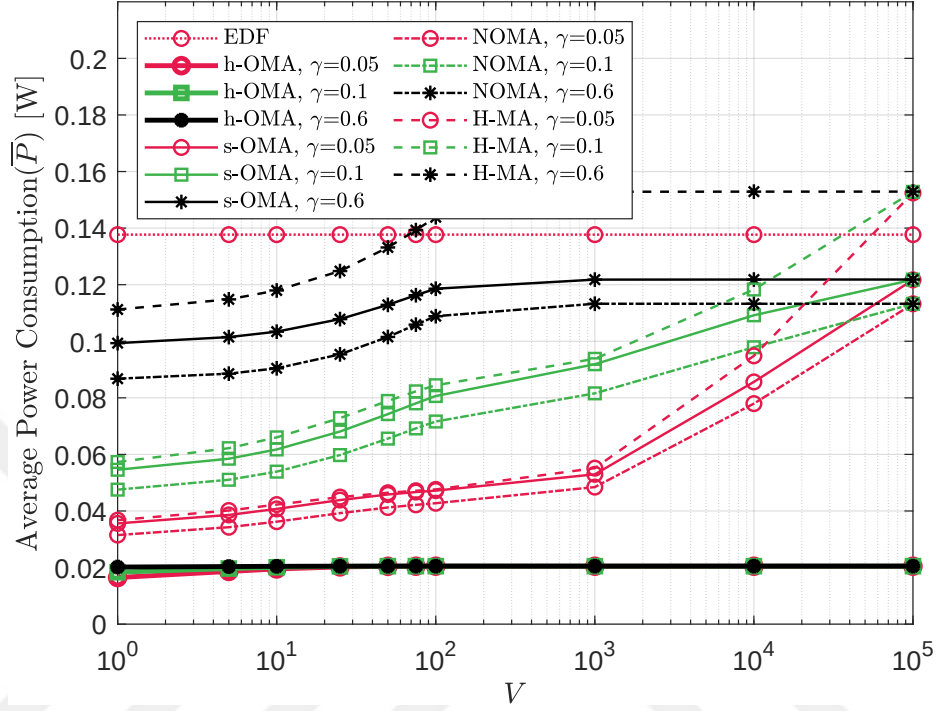


Figure 4.4: Average power consumption as a function of γ with $\pi = 0.1$.

of s-OMA with $\alpha = 0.01$. Moreover, NOMA starts to perform closer to H-MA. Finally, in Fig. 4.3 with $\pi = 0.5$, H-MA with $\alpha = 0.1$ performs about 30% better than p-OMA. These results show that with increasing traffic load and packet size diversity, NOMA capable H-MA and NOMA significantly outperform other algorithms.

In all traffic levels, H-MA performs the best, showing that Hybrid MA is a robust approach and RDPPA with proper selection of α increases timely throughput. The best performance for H-MA and s-OMA under different traffic levels are obtained when $\alpha \in \{0.01, 0.1, 1\}$. Therefore, $\alpha = 0.1$ is considered for the rest of the simulations. Finally, EDF performs significantly worse as the traffic level increases, showing that simple algorithms like EDF are not viable.

In Figs. 4.4, 4.5, 4.6, $\bar{P}$ with different values of γ and $\pi \in \{0.1, 0.3, 0.5\}$ are presented, respectively. Since EDF and p-OMA are independent of V , their results are the same for all V values. The results of p-OMA under $\pi \in \{0.1, 0.3, 0.5\}$ are 6.063 kW, 7.9509 kW and 9.3139 kW, respectively. Since the results of p-OMA

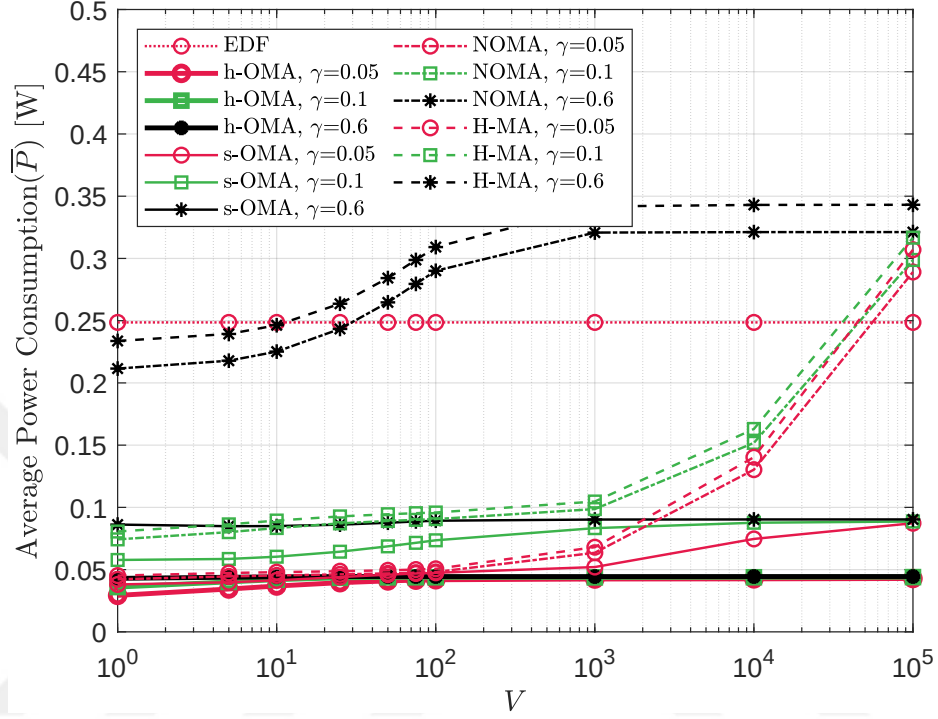


Figure 4.5: Average power consumption as a function of γ with $\pi = 0.3$.

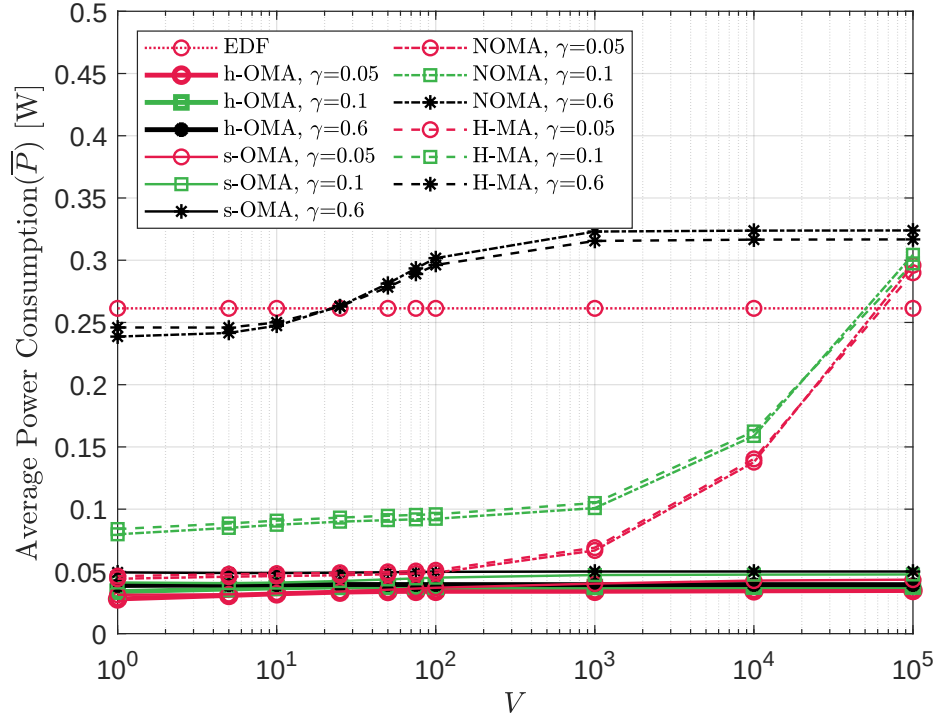


Figure 4.6: Average power consumption as a function of γ with $\pi = 0.5$.

are significantly greater than the maximum power value of $P_0 = 3$ W, their plots are not presented on the Figs. 4.4, 4.5, 4.6. H-MA is the most sensitive algorithm to the given average power utilization constraint and NOMA consumes slightly less power. Significantly high values of V force the algorithms to ignore the provided average power constraint in Figs. 4.4, 4.5, 4.6. For all the arrival rates $\pi \in \{0.1, 0.3, 0.5\}$, H-MA, NOMA and s-OMA start to violate the provided average power constraint as V increases from 100, which is observed also for $\pi = 0.3$ in Fig. 4.5. Based on these observations, it can be concluded that $V = 100$ is suitable for balancing the tradeoff between $\bar{D}$ and $\bar{P}$.

Another important role of the V parameter is about the optimality of the *drift-plus-penalty* algorithm. As presented in Section 3.3, the *drift-plus-penalty* algorithm closely approximates optimality under the Slater condition [42, 71], which indicates the satisfaction of all time average constraints. Therefore, proper selection of V is also important for balancing the tradeoff between the optimality of the *drift-plus-penalty* algorithm and timely throughput maximization. In Figs. 4.4, 4.5, 4.6, $\bar{P}$ satisfies the average power constraints γ with $V = 100$ and diverse traffic levels. Thus, we can conclude that $V = 100$ is suitable for closely achieving the highest timely throughput performance while satisfying optimality. Finally, the optimality of the *drift-plus-penalty* algorithm is satisfied for $V < 100$, however, the timely throughput performance is reduced compared to $V = 100$.

The percentage of increase in timely throughput achieved by H-MA compared to s-OMA and NOMA for different π and γ values under $\alpha = 0.1$ and $V = 100$ is presented in Tables 4.2 and 4.3, respectively. In Table 4.2, the increase in timely throughput using H-MA with respect to s-OMA is small for $\pi = 0.1$. As traffic level increases with $\pi = 0.5$, H-MA increases timely throughput up to 46%. Based on these results, we can conclude that H-MA increases timely throughput compared to s-OMA on the average by nearly 21.27% while satisfying average power constraints for all arrival rates. On the other hand, the increase in timely throughput using H-MA with respect to NOMA is high for $\pi = 0.1$ and it decreases as traffic level increases with $\pi = 0.5$. The reason for this inversely proportional behavior is that, low traffic levels can be handled by s-OMA, whereas the existence probability of a suitable NOMA user pair is proportional to the

Table 4.2: In H-MA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	3.63%	15.92%	36.42%
0.1	3.50%	18.59%	40.71%
0.6	3.47%	23.38%	45.79%

Table 4.3: In H-MA and NOMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	66.72%	13.03%	4.25%
0.1	65.42%	13.04%	3.76%
0.6	64.42%	12.71%	2.40%

increasing traffic level. Finally, these results show that H-MA is also robust under varying traffic levels.

In Table 4.5, the mean and variance of the $\overline{D}_i$'s FI over 1000 simulations under the design parameters $\alpha = 0.1$ and $V = 100$ are presented for traffic level $\pi = 0.3$. The OMA based algorithms, h-OMA and s-OMA, are slightly more fair in terms of average packet dropping rate compared to NOMA based H-MA and NOMA algorithms, since their average FI is higher and the variance of FI is lower. The same relation is observed also for $\pi = 0.1$ and $\pi = 0.5$ traffic levels in Tables 4.4 and 4.6, respectively. Thus, we can conclude that, NOMA based H-MA significantly increases overall performance at the cost of a slight decrease in FI .

In Fig. 4.7, the average incomplete bit-rates ($\bar{I}$) with different values of N under arrival rate $\pi = 0.3$ are plotted. $\bar{I}$ refers to the average data rate, in bits per slot, of dropped packets for which all fragments originated from FTM could not be successfully transmitted by the packet deadline. $\bar{I}$ value of H-MA is lower than $\bar{I}$ value of NOMA for small values of N , which shows the benefit of Hybrid MA in terms of avoiding packet dropping by suitable decisions between OMA and NOMA transmission. $\bar{I}$ for s-OMA, NOMA and H-MA converges to 0 as N

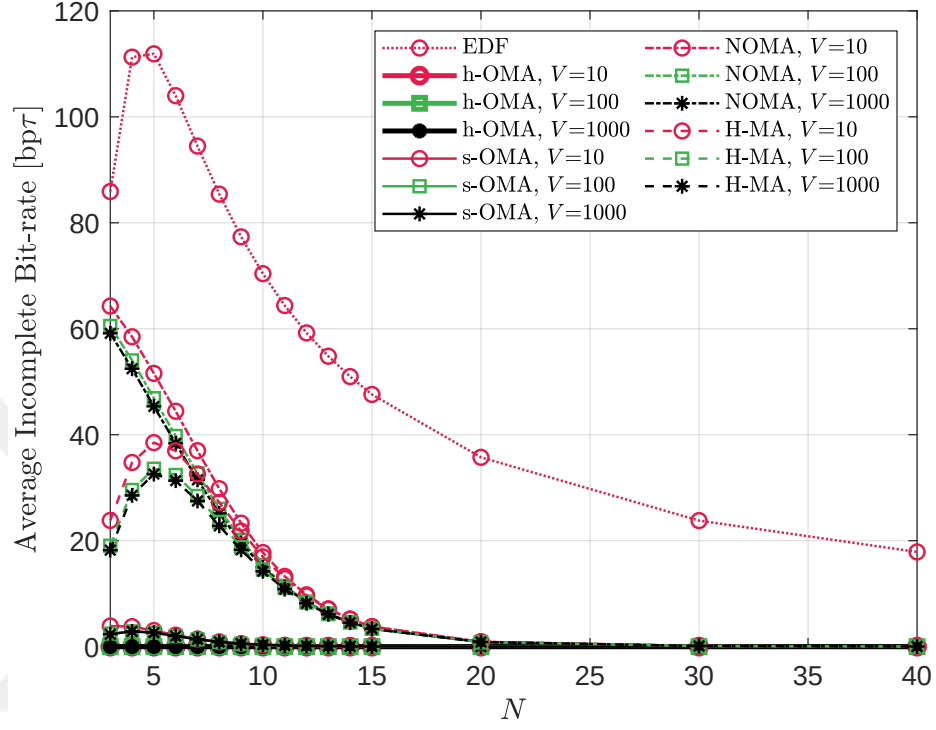


Figure 4.7: Average incomplete bit-rate as a function of N with $\pi = 0.3$.

Table 4.4: Mean and variance of FI for different algorithms and γ values under $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.86292	0.86292	0.86293	0.01310	0.01311	0.01311
s-OMA	0.71412	0.72200	0.86293	0.01528	0.01556	0.01543
NOMA	0.92265	0.92813	0.93506	0.00178	0.00165	0.00154
H-MA	0.77739	0.79863	0.82947	0.01901	0.01907	0.01803

Table 4.5: Mean and variance of FI for different algorithms and γ values under $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.91076	0.92004	0.92665	0.00481	0.00378	0.00312
s-OMA	0.89435	0.90751	0.92444	0.00496	0.00374	0.00238
NOMA	0.87862	0.88264	0.90010	0.00612	0.00612	0.00579
H-MA	0.85498	0.86147	0.88792	0.01171	0.01183	0.01087

Table 4.6: Mean and variance of FI for different algorithms and γ values under $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.97091	0.97978	0.98588	0.00054	0.00026	0.00012
s-OMA	0.97026	0.97949	0.98592	0.00054	0.00027	0.00012
NOMA	0.91289	0.91341	0.92093	0.0049273	0.00499	0.00478
H-MA	0.90627	0.90890	0.92179	0.00636	0.00626	0.00569

increases, due to the increase in the existence probability of a small packet from Λ on a strong channel. For small values of N , such as $N = 5$, we observe high $\bar{I}$ for NOMA and H-MA. Although s-OMA, NOMA and H-MA algorithms reduce $\bar{D}$ compared to the baseline algorithms, they increase $\bar{I}$ due to the dynamic nature of the opportunistic scheduling.

4.2 Dynamic Power Allocation with EPD

In this section, the EPD mechanism is introduced to the system with $\alpha = 0.1$ and $V = 100$, which are the values optimized in Section 4.1. In Figs 4.8 - 4.10, $\bar{D}$ results are plotted as a function of EPD slot count S under $\pi \in \{0.1, 0.3, 0.5\}$, respectively. Simulations are performed for 1000 random seeds and their averages are reported. The set of available EPD slot count values is selected as $S \in \{-2, 0, 1, \dots, 20\}$. Note that, $S = -2$ refers to no EPD mechanism and $S \geq 0$ indicates an active EPD mechanism as presented in Section 3.2. The first thing to notice in Fig 4.8 is the difference in $\bar{D}$ between $S = -2$ and $S \geq 0$ cases. As EPD slot count increases from $S = -2$ towards $S = 0$, $\bar{D}$ increases for NOMA, whereas it either decreases or does not change for the other algorithms. The reason for that is, the wrong decisions made by the EPD mechanism at this low traffic level affect the overall timely throughput. Moreover, since NOMA can operate with exactly two users, the wrong drop decisions made by EPD may block NOMA due to a lack of packets suitable for NOMA pairing. The second thing to notice is the algorithms' sensitivity with respect to S for $S \geq 0$. NOMA, s-OMA and H-MA are not sensitive to S in this low traffic level. The only algorithms

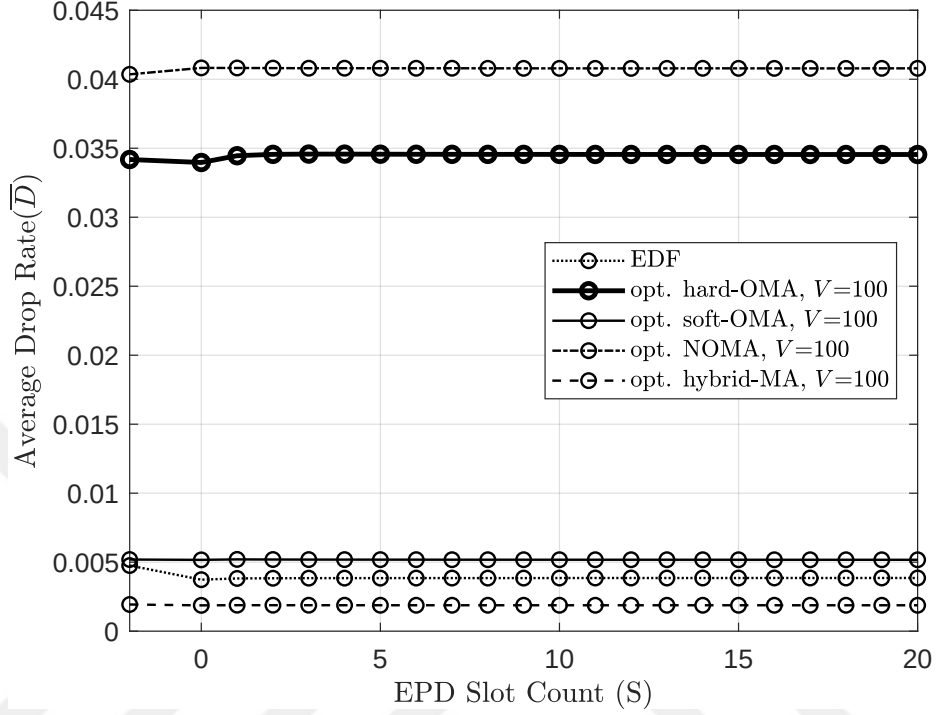


Figure 4.8: Average drop rate with $\pi = 0.1$ as a function of S in EPD mechanism.

that apparently perform the best $\bar{D}$ for $S = 0$ are EDF and h-OMA, which are algorithms based on CTM. The reason is, CTM and OMA based algorithms can compensate for the wrong decisions of EPD by aiming to transmit a packet fully in a slot. Therefore, the probability that EPD wrongly detects an incomplete packet as likely to-be-expired reduces. These observations indicate that EPD is not effective at low traffic levels, especially for the algorithms utilizing FTM and NOMA. Meanwhile, CTM and OMA based algorithms perform slightly better with EPD even at low traffic levels.

As the traffic level increases towards $\pi = 0.3$ and $\pi = 0.5$ as shown in Figs 4.9 and 4.10, timely throughput performances of all the algorithms increase with S . The reason is, high traffic levels increase the number of user packets left in the queues with the EPD mechanism, so that the probability of an algorithm finding a proper packet to serve on a slot increases. In this case, the timely throughput increases with the more effective utilization of the available network resources. Secondly, we observe that the algorithms' sensitivity with respect to S slightly increases at high traffic levels. Interestingly, the best $\bar{D}$ performance

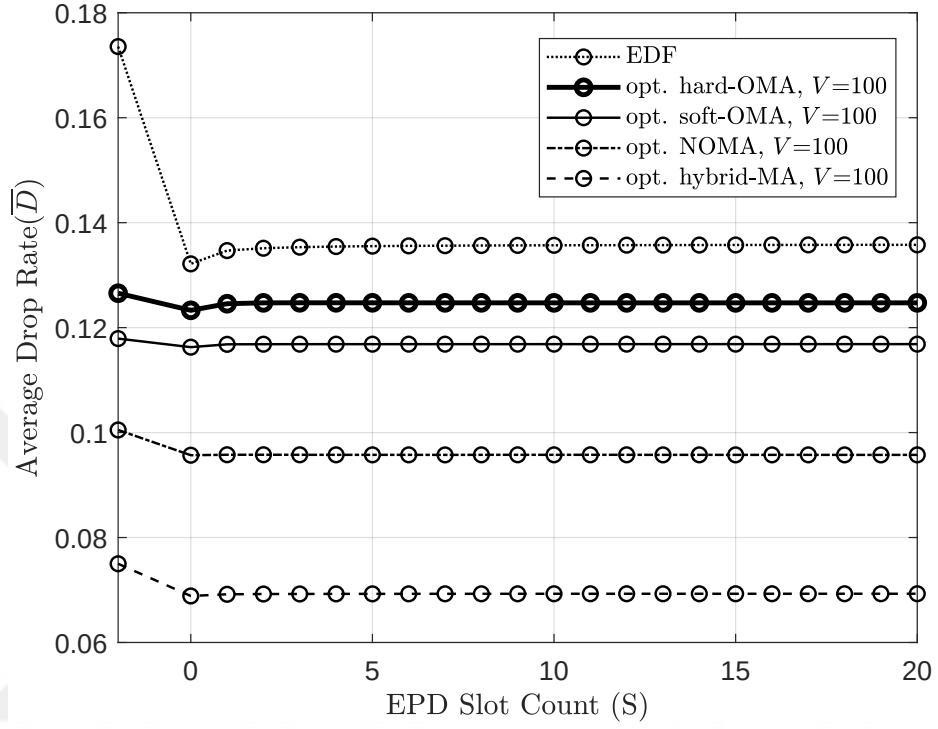


Figure 4.9: Average drop rate with $\pi = 0.3$ as a function of S in EPD mechanism.

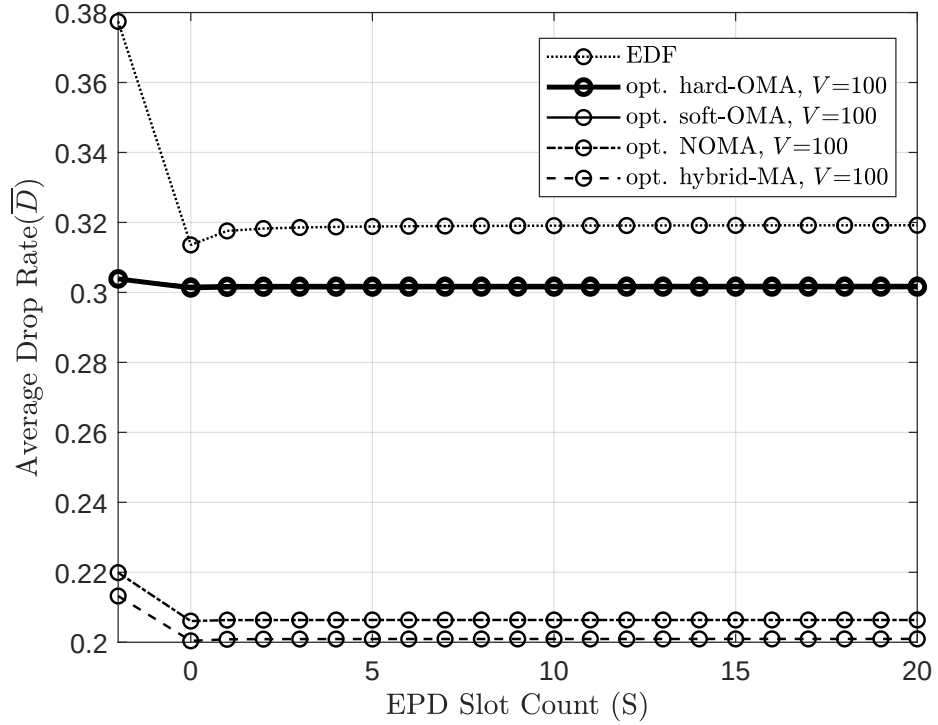


Figure 4.10: Average drop rate with $\pi = 0.5$ as a function of S in EPD mechanism.

Table 4.7: In H-MA and s-OMA comparison, percentage of increase in timely throughput for different π values and EPD utilization cases for under $\gamma = 0.6$, $\alpha = 0.1$ and $V = 100$. In all cases, s-OMA is operated without EPD.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	3.94%	19.50%	42.04%
0.1	3.74%	22.25%	46.63%
0.6	3.63%	27.21%	52.78%

for most of the algorithms, including H-MA, is observed with $S = 0$ at $\pi = 0.3$ and $\pi = 0.5$ traffic levels. Note that, $S = 0$ refers to considering only the channel gain observed in the current slot. Thus, there is no averaging over a range of slots. In the $S = 0$ case, the variance of the $\hat{\mu}_i(S)$ over the slots is expected to be at the highest among all $S \geq 0$ cases. For instance, estimated OMA transmission rate $\hat{\mu}_i(S)$ might result in a small value due to a deep fading encountered by user i . In such a case, the respective packets of user i might be wrongly detected as likely to-be-expired and be dropped by the EPD mechanism. Although the channel gain of user i might increase in the next slot, there will be no chance for the dropped packets. Therefore, EPD might decide to drop a packet just because of a very low channel gain observed in a slot. In the low traffic scenario of $\pi = 0.1$, we observed that EPD reduces the timely throughput performance in such a situation, especially for algorithms utilizing FTM and NOMA. However, the results at high traffic levels of $\pi = 0.3$ and $\pi = 0.5$ show that EPD improves the timely throughput by increasing the chance of transmission for packets waiting in the queue. This happens due to early dropping of packets that are likely to become expired. Thus, we can conclude that EPD further improves the timely throughput as the traffic level increases.

The percentage of increase in timely throughput achieved by H-MA with EPD compared to s-OMA and NOMA without EPD for different π and γ values under the optimized $\alpha = 0.1$ and $V = 100$ values are presented in Tables 4.7 and 4.8, respectively. In Table 4.7, H-MA with EPD increases timely throughput up to 53% compared to s-OMA at the high traffic level of $\pi = 0.5$. Moreover, H-MA with EPD achieves improvement in timely throughput compared to s-OMA on

Table 4.8: In H-MA and NOMA comparison, percentage of increase in timely throughput for different π values and EPD utilization cases under $\gamma = 0.6$, $\alpha = 0.1$ and $V = 100$. In all cases, NOMA is operated without EPD.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	67.22%	16.52%	8.54%
0.1	65.80%	16.53%	8.13%
0.6	64.68%	16.21%	7.31%

the average by nearly 24.5% while satisfying average power constraints for all arrival rates. However, the effect of EPD on the improvement achieved with H-MA compared to s-OMA depends on the arriving traffic level: while there is no improvement at low traffic rates, the timely throughput improves as the traffic rate increases. The improvement in the timely throughput with EPD when H-MA is compared with NOMA is clear at all traffic rates. Finally, improvement achieved in timely throughput at various traffic levels using H-MA with EPD compared to s-OMA and NOMA in Tables 4.7 and 4.8 show that H-MA with EPD is robust under varying traffic loads.

$\bar{I}$ under traffic level of $\pi = 0.3$ with EPD is presented in Fig 4.11. The only parametric difference between Figs 4.7 and 4.11 is the utilization of the EPD mechanism. These two figures reveal that EPD significantly reduces $\bar{I}$. Thus, we can conclude that EPD reduces the wasted network resources for incomplete packets. The proposed dynamic power allocation algorithm leverages the achieved reduction in the wasted power to further improve the timely throughput. These results indicate that flexible packet dropping approaches such as EPD have a significant potential for improving the timely throughput.

In Tables 4.9-4.10-4.11, the mean and variance of FI for different algorithms and γ values under the EPD mechanism, the optimized parameters of $\alpha = 0.1$ and $V = 100$, and the different traffic levels $\pi \in \{0.1, 0.3, 0.5\}$, respectively. Note that, the utilization of the EPD mechanism is the only difference between the FI results presented in Tables 4.4-4.5-4.6 and Tables 4.9-4.10-4.11, respectively. The mean FI of H-MA slightly increases with the EPD mechanism under all traffic levels. Note that, the EPD mechanism avoids the waste of power on packets that

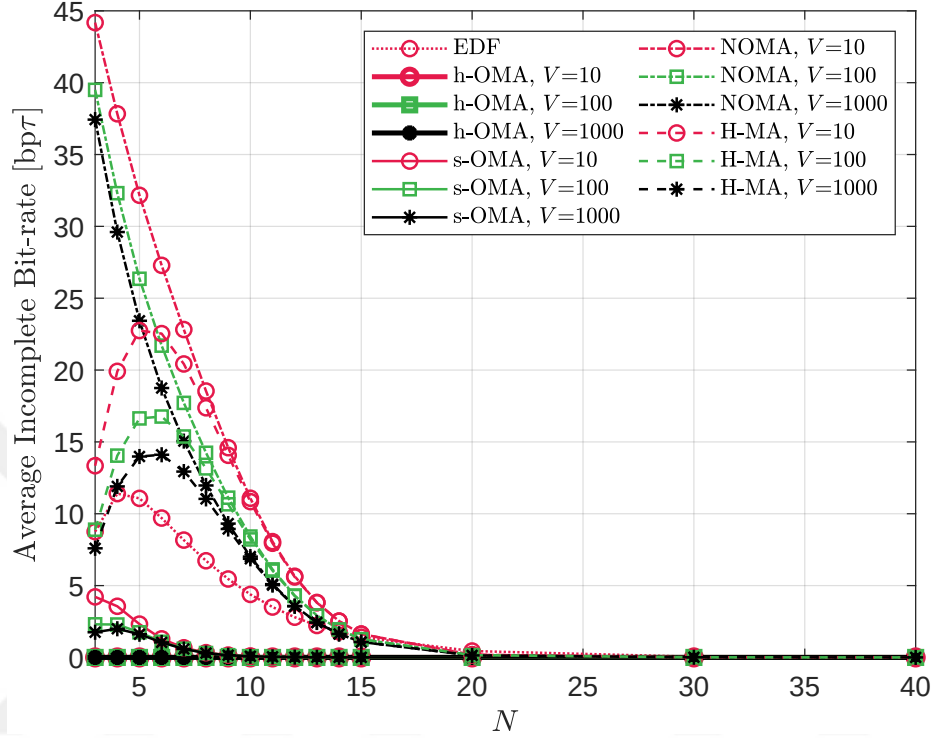


Figure 4.11: Average incomplete bit-rate as a function of N with $\pi = 0.3$ and EPD utilization.

are likely to become expired by dropping them before the actual expiration. The observed increase in FI performance of H-MA with the EPD mechanism is due to this avoidance of the waste of power. Therefore, we can conclude that the EPD mechanism significantly increases timely throughput while preserving the fairness among users.

Table 4.9: Mean and variance of FI for different algorithms and γ values under EPD mechanism, $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.86272	0.86272	0.86271	0.01311	0.01312	0.01313
s-OMA	0.71869	0.72649	0.74856	0.01476	0.01485	0.01442
NOMA	0.93182	0.93655	0.94251	0.00154	0.00145	0.0013281
H-MA	0.78548	0.80558	0.83928	0.01876	0.01908	0.01685

Table 4.10: Mean and variance of FI for different algorithms and γ values under EPD mechanism, $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.9081	0.91805	0.92529	0.0049545	0.0038118	0.0030852
s-OMA	0.89329	0.90497	0.9201	0.0053729	0.0041756	0.0026639
NOMA	0.8877	0.89434	0.91331	0.005341	0.0052389	0.0050603
H-MA	0.86126	0.86952	0.89436	0.011474	0.011491	0.010656

Table 4.11: Mean and variance of FI for different algorithms and γ values under EPD mechanism, $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
h-OMA	0.97059	0.98013	0.98671	0.00051918	0.0002373	0.00010329
s-OMA	0.97335	0.98022	0.98654	0.00044825	0.00024265	0.00010372
NOMA	0.91557	0.91746	0.92775	0.0047089	0.0047613	0.0046832
H-MA	0.90849	0.91176	0.92646	0.0064252	0.0062855	0.0056765

4.3 Computational Load Reduction Heuristic

In this section, the performance of computational load reduction heuristic *BEST* is presented in comparison with *RAND*. Simulations are performed with $\alpha = 0.1$ and $V = 100$, which are the values optimized in Section 4.1. EPD is utilized in all simulations. Simulations are performed for 100 seeds and their average performance is presented. We formulated the computation time of the proposed algorithm as a function of Y as follows:

$$\text{ComputationTime}(Y) = N + \binom{Y}{2}. \quad (4.2)$$

In order to clearly present the computation time as a function of Y along with the performance gain achieved with *BEST* compared to *RAND*, we prepared plots to combine normalized timely throughput and normalized computation time for different values of Y . On the left y-axis, normalized timely throughput is presented for *BEST* and *RAND* heuristics. On the right y-axis, normalized computation time is presented according to the computation time formulation

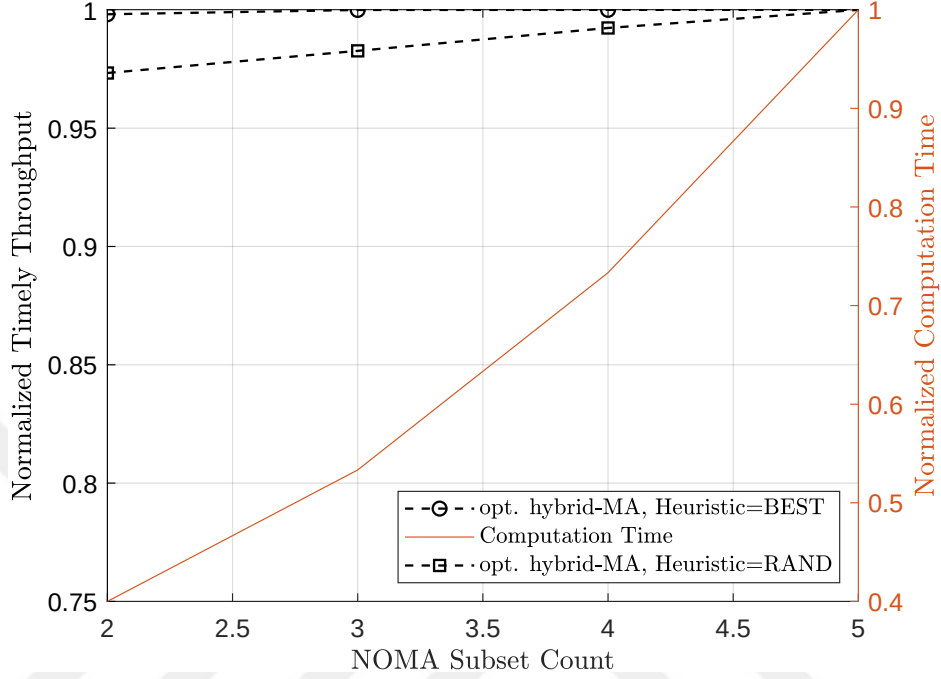


Figure 4.12: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 5$ and $\pi = 0.1$.

presented in (4.2). The x-axis of the plots represents the value of Y , which denotes the user count of the NOMA subset.

In Figs 4.12-4.13-4.14, normalized timely throughput for *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 5$ and the traffic levels $\pi \in \{0.1, 0.3, 0.5\}$, respectively. *BEST* mostly preserves timely throughput down to $Y = 3$, which refers to about 50% reduction in the computation time for all traffic levels. For $Y = 2$ case, normalized timely throughput with *BEST* gets the lowest value of 0.95 as traffic level increases towards 0.5. In all traffic levels, the timely throughput performance of *RAND* continuously deteriorates as NOMA subset count decreases from $Y = 5$ to $Y = 2$. These results show that *BEST* significantly preserves timely throughput compared to *RAND* for small number of users such as $N = 5$.

In Figs 4.15-4.16-4.17, normalized timely throughput for *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 10$ and the traffic level $\pi \in \{0.1, 0.3, 0.5\}$, respectively. Similarly,

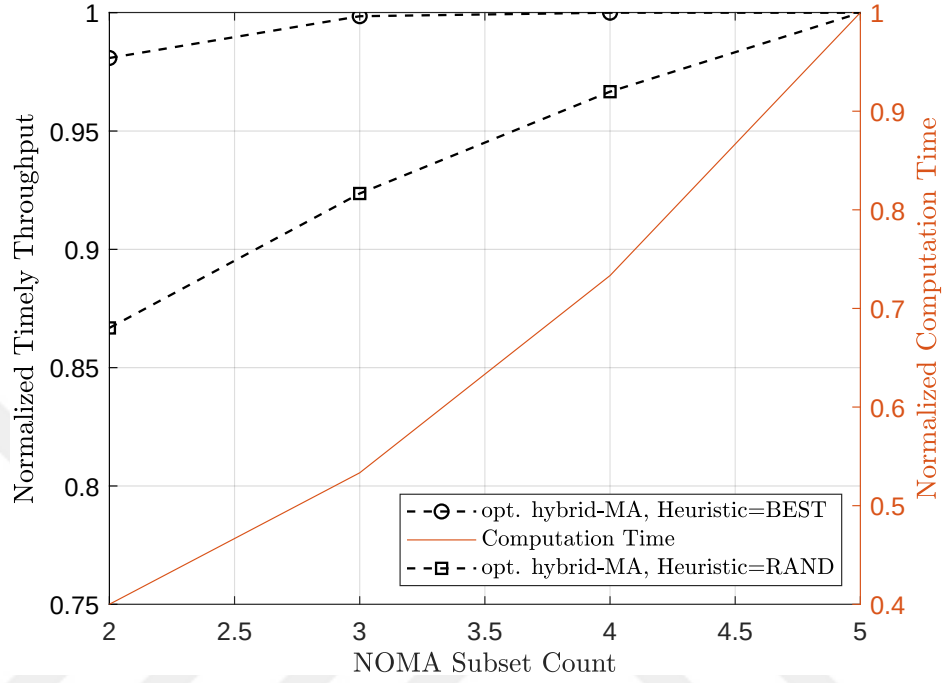


Figure 4.13: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 5$ and $\pi = 0.3$.

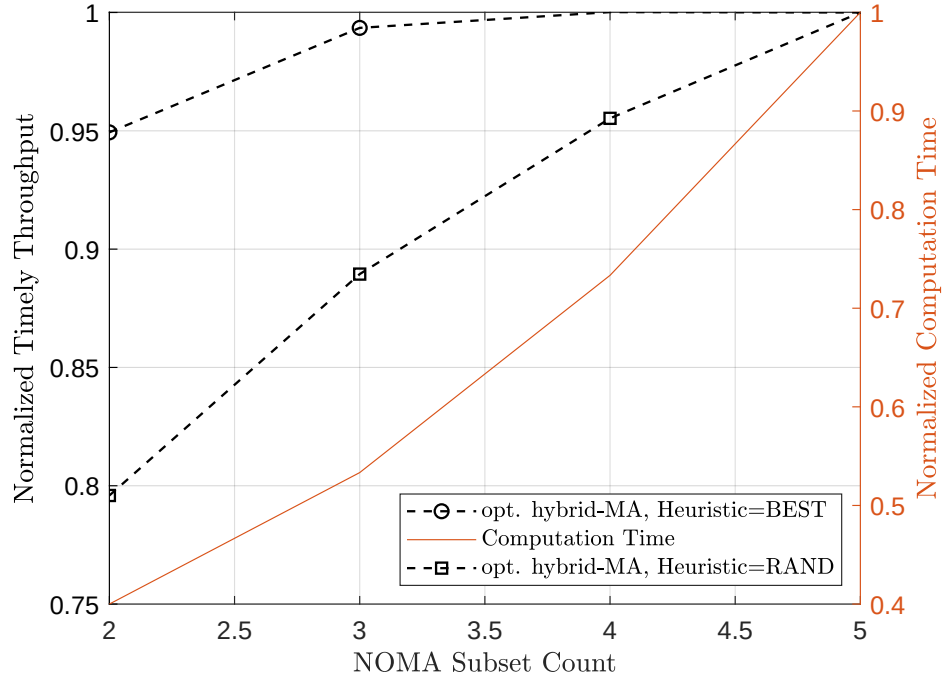


Figure 4.14: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 5$ and $\pi = 0.5$.

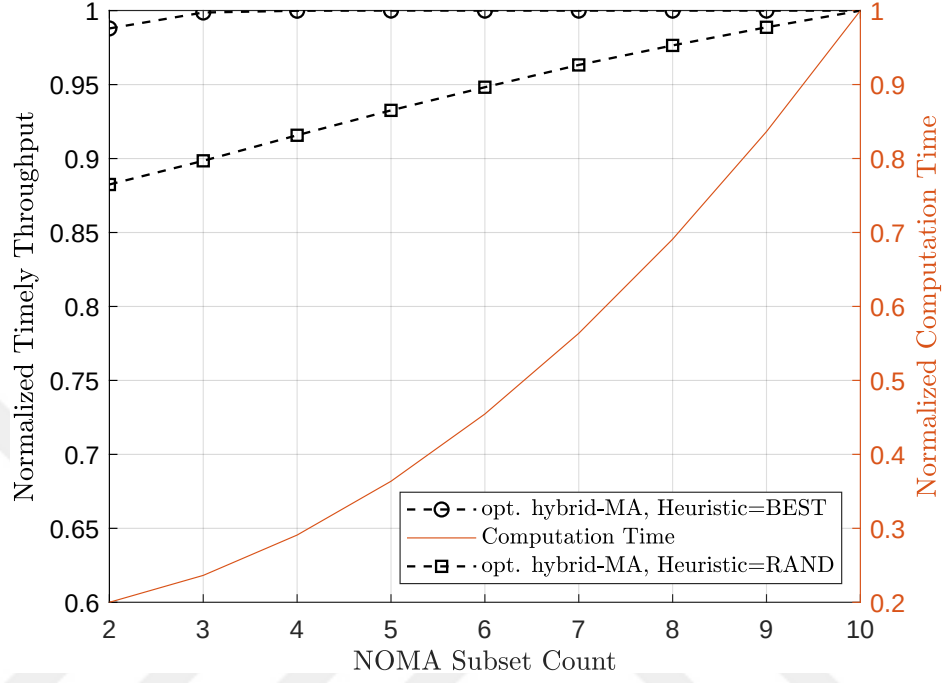


Figure 4.15: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 10$ and $\pi = 0.1$.

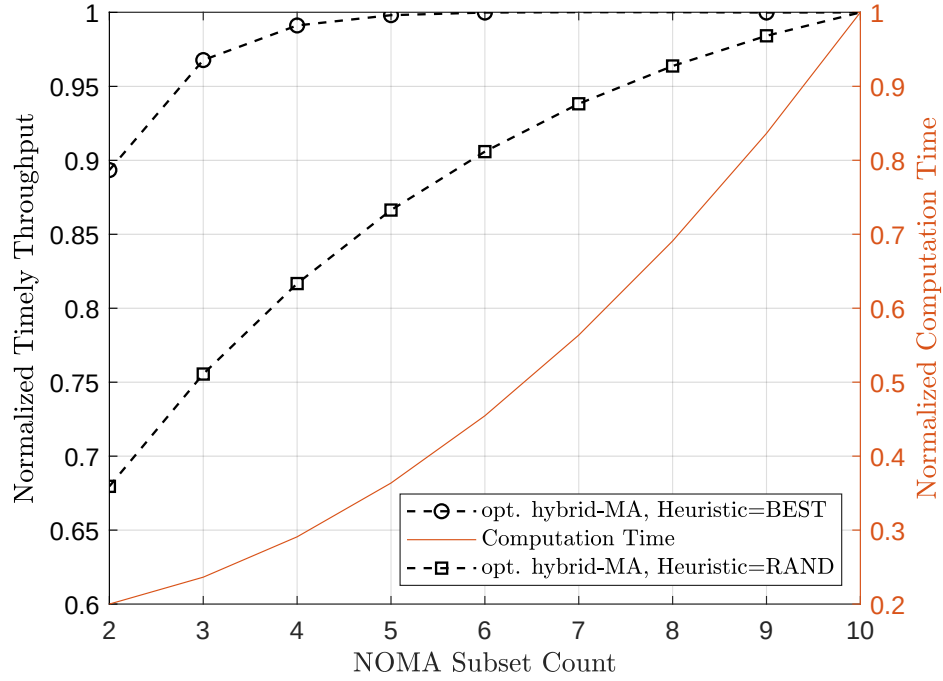


Figure 4.16: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 10$ and $\pi = 0.3$.

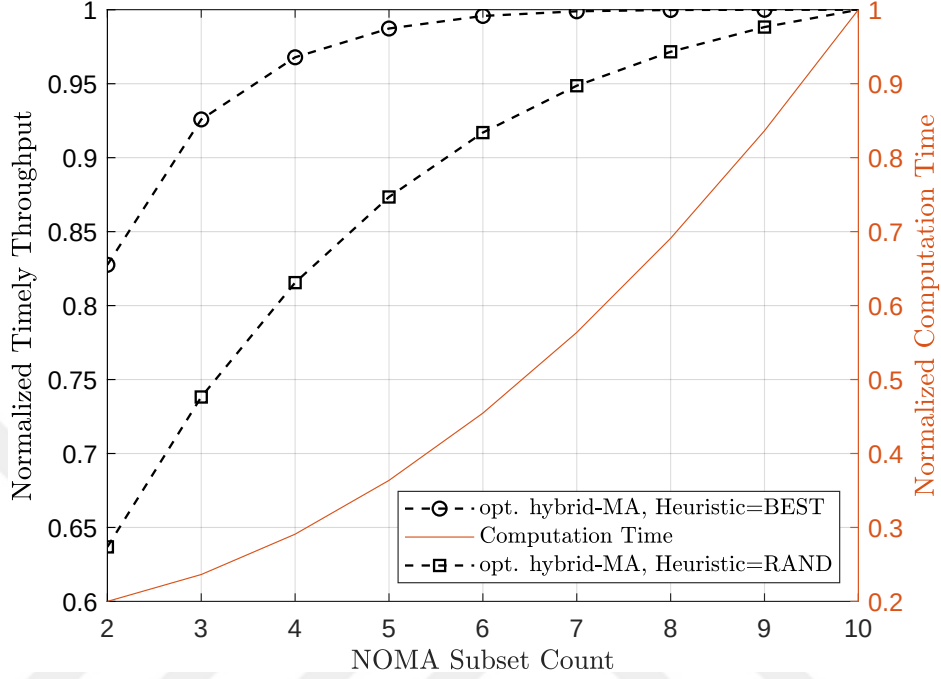


Figure 4.17: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 10$ and $\pi = 0.5$.

BEST performs significantly better than *RAND* in all cases. For low traffic level of $\pi = 0.1$, *BEST* mostly preserves timely throughput performance at $Y = 3$, which refers to about 75% reduction in computation time. As the traffic level increases towards $\pi = 0.3$ and $\pi = 0.5$, *BEST* can preserve timely throughput at $Y = 4$ and $Y = 6$, respectively. These cases refer to 70% and 55% reduction in computation time, respectively. Therefore, as the number of users increases in the system, the performance of *BEST* starts to proportionally alter with respect to the load of the system. This observation is meaningful, since *BEST* does not reduce the order of computation complexity, it reduces computation load.

In Figs 4.18-4.19-4.20, normalized timely throughput for *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 20$ and the traffic level $\pi \in \{0.1, 0.3, 0.5\}$, respectively. Similarly, *BEST* performs better than *RAND* in all cases. For low traffic level of $\pi = 0.1$, *BEST* preserves timely throughput performance at $Y = 4$, which refers to about 90% reduction in computation time. As the traffic level increases towards $\pi = 0.3$ and $\pi = 0.5$, *BEST* can preserve timely throughput at $Y = 7$ for both cases, and

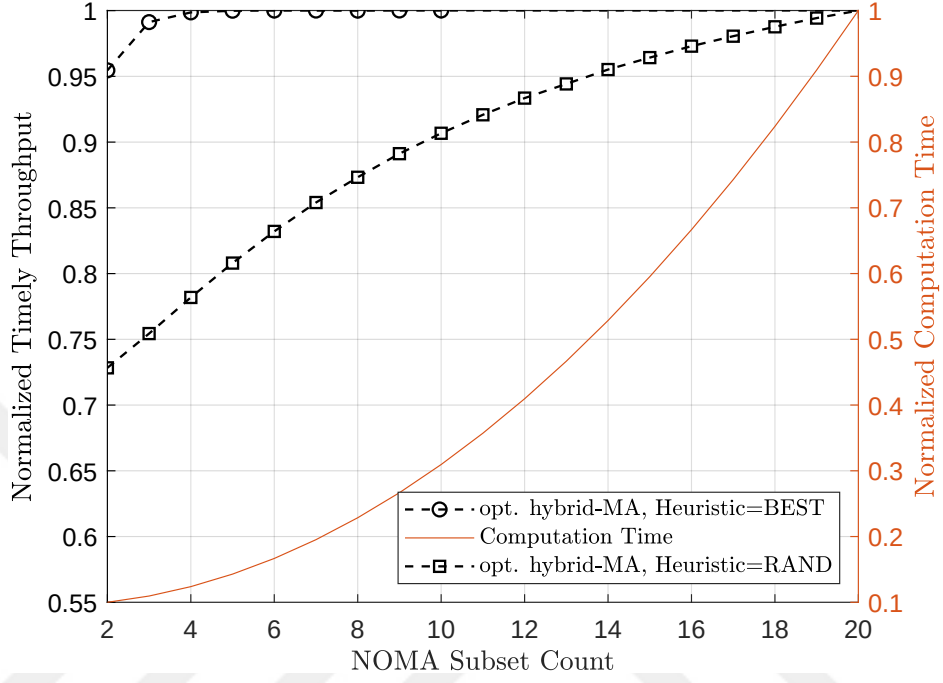


Figure 4.18: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 20$ and $\pi = 0.1$.

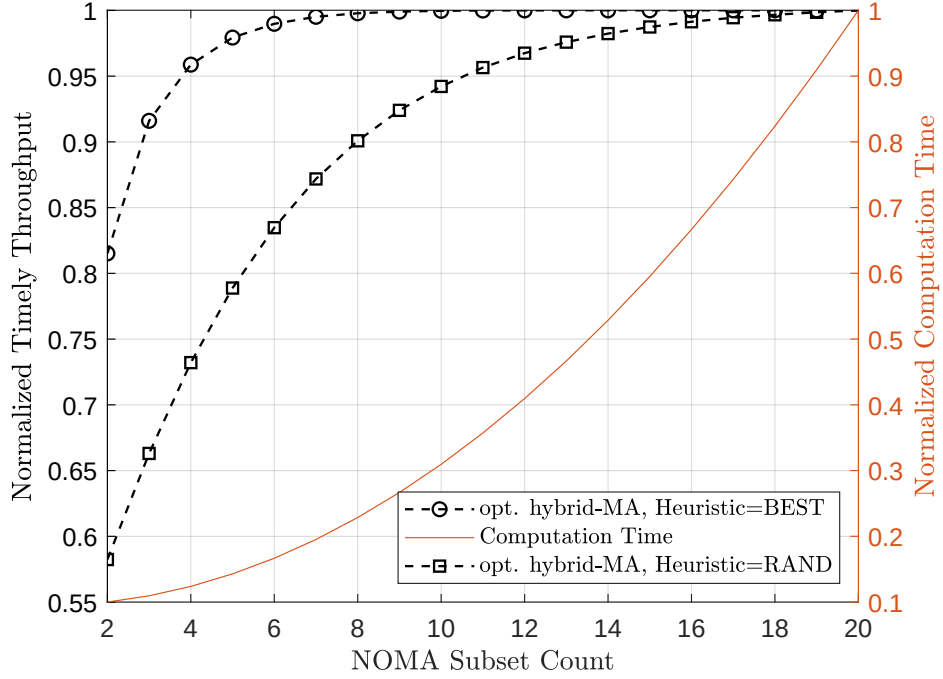


Figure 4.19: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 20$ and $\pi = 0.3$.

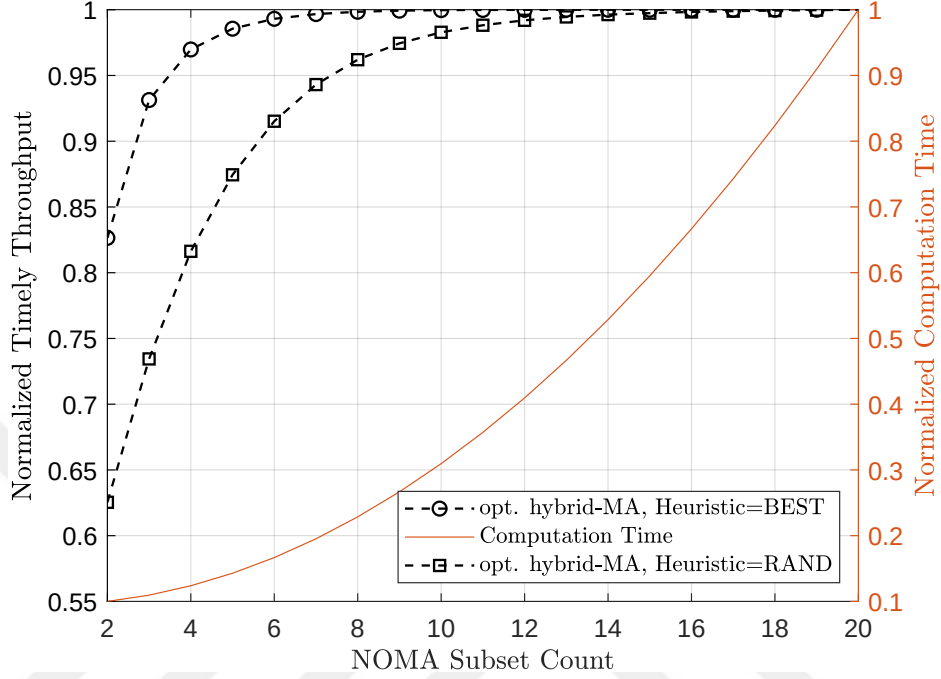


Figure 4.20: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 20$ and $\pi = 0.5$.

it refers to about 80% reduction in computation time. Therefore, as the number of users further increases in the system, the performance of *BEST* starts to saturate. The reason for this observation is about the increment in the existence probability of a short packet in the system as traffic level increases from $\pi = 0.3$ to $\pi = 0.5$. As the existence probability of a short packet in the system exceeds a threshold, further increase in the system load does not decrement the performance of *BEST*, since these short packets can be easily found within the head of queues, which are selected for NOMA subsets. Another interesting result of this situation is the slight performance increase in timely throughput obtained with *BEST* and *RAND* as traffic rate increases from $\pi = 0.3$ to $\pi = 0.5$.

In Figs 4.21-4.22-4.23, normalized timely throughput for *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 40$ and the traffic level $\pi \in \{0.1, 0.3, 0.5\}$, respectively. Similarly, *BEST* performs better than *RAND* in all cases. In all traffic levels, *BEST* preserves timely throughput performance at $Y = 7$, which refers to about 90% reduction in computation time. Therefore, we can conclude that as the number of

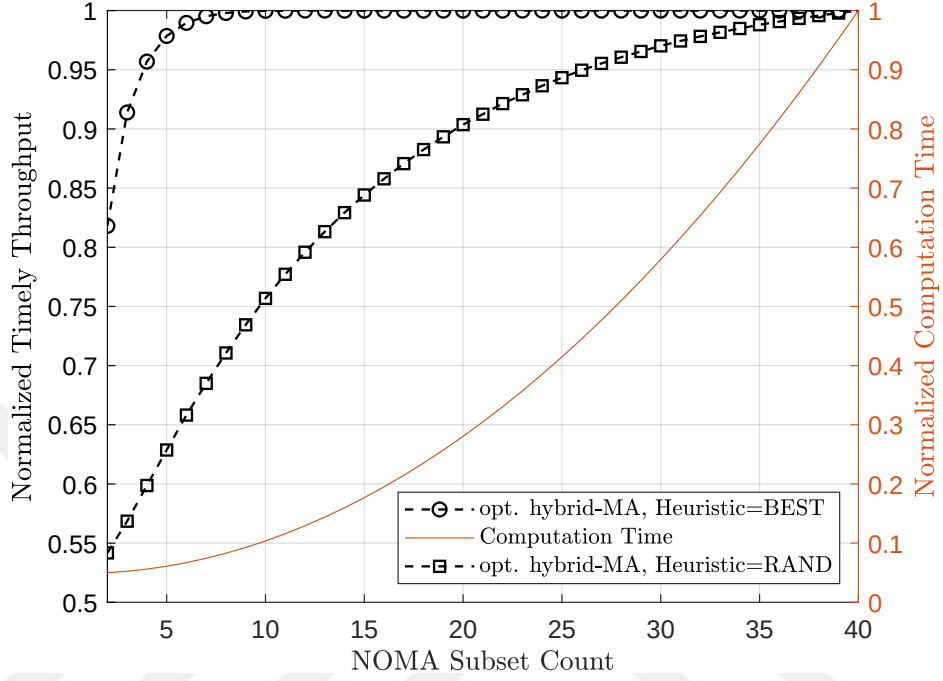


Figure 4.21: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 40$ and $\pi = 0.1$.

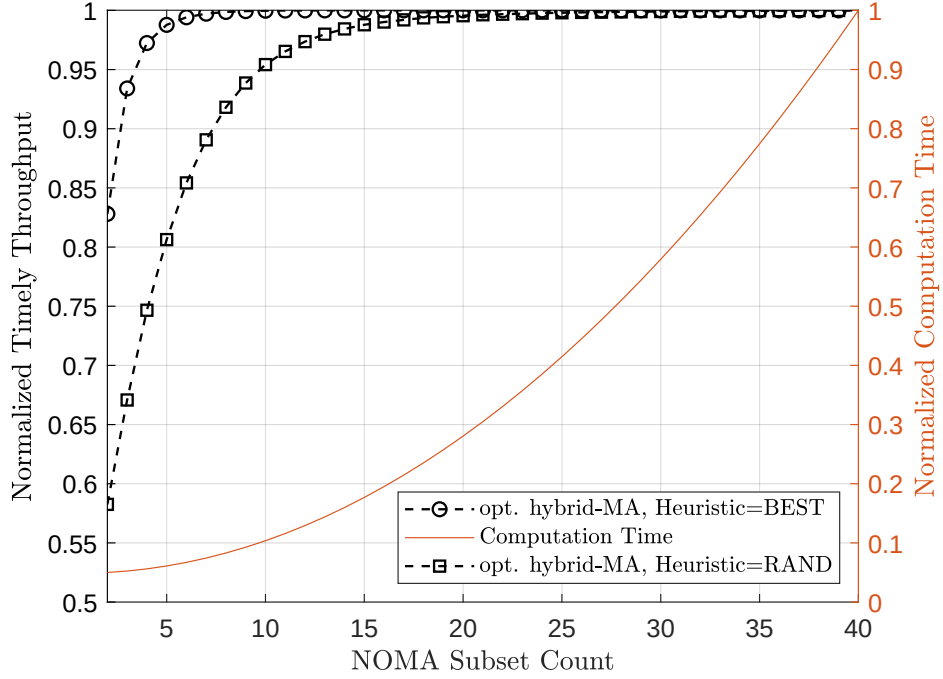


Figure 4.22: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 40$ and $\pi = 0.3$.

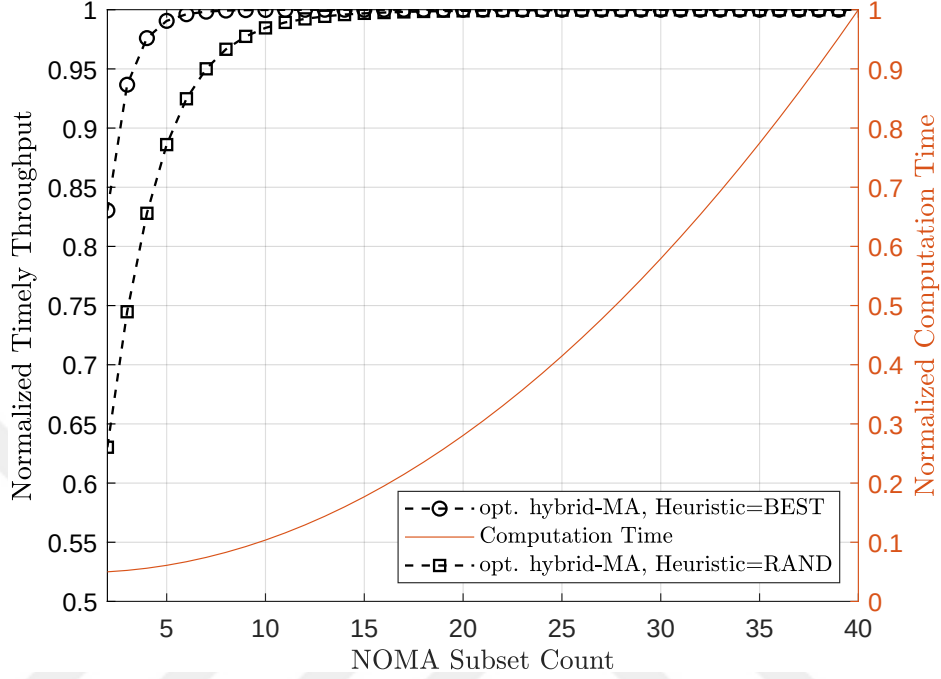


Figure 4.23: Normalized timely throughput and normalized computation time as a function of NOMA Subset Count Y under $N = 40$ and $\pi = 0.5$.

users further increases in the system, the performance of *BEST* starts to saturate under all traffic levels. The reason for this observation is about the existence probability of a short packet in the system being greater than a threshold, at which further increase in the system load does not decrement the performance of *BEST*. Thus, these short packets can be easily found within the head of queues, which are selected for NOMA subsets, under all traffic levels. Another interesting result of this situation is the performance increase in *RAND* as the traffic level increases from $\pi = 0.1$ to $\pi = 0.5$. Moreover, another interesting result of the situation is the performance of *BEST* and *RAND* getting closer to each other as traffic level increases towards $\pi = 0.5$.

Simulation results show that the proposed heuristic *BEST* significantly reduces computation time for Hybrid-MA by within a range of 50% to 90% under different user count N and arrival rate π values, while mostly preserving timely throughput. As the system load increases with elevating user count N and traffic level π , the performance of *BEST* starts to saturate at some point due to increasing existence probability of short packets at the head of queues. In other words,

the single output link in our transmitter model starts to reach its performance limits. *BEST* is a simple and effective heuristic. Moreover, *BEST* can be easily applied to RSMA, too.



Chapter 5

Timely Throughput Maximization with RSMA

In this chapter, RSMA-oriented system model, problem formulation and proposed solution are presented firstly. Note that, they are similar to the ones explained in Chapter 3. For the sake of completeness, they are briefly explained in this chapter, and the symbols defined in Chapter 3 are used in this chapter as is. Then, optimal power assignment for two-user RSMA is presented.

5.1 System Model and Problem Formulation

We consider a discrete time slotted downlink scenario where a base station (BS) with coverage radius R communicates with a set of users denoted as $\mathcal{N} \in \{1, \dots, N\}$. BS is equipped with a single antenna. At each slot t , BS broadcasts using available bandwidth $\mathcal{B}$ throughout the slot duration τ using RSMA. Since Shannon capacity is not applicable for the FBL, we consider the tight approximation on transmission rate R^* , which is presented in (3.1). The total power budget on a slot is denoted as P_0 . Allocated power for user i is denoted as $\{P_i(t)\}_{i \in \mathcal{N}} \leq P_0$. Let $\bar{P}_i \triangleq \lim_{t \rightarrow \infty} t^{-1} \sum_{\tau=0}^{t-1} P_i(\tau) \leq \gamma_i$ be the time average

power for user i , where $\gamma_i \in (0, P_0]$ is the average power consumption constraint for user i . Let $\bar{P} \triangleq N^{-1} \sum_{i=1}^N \bar{P}_i$ be the overall average power consumption.

The wireless channel between BS and stationary users is modelled as a Rayleigh fading channel, where the fast fading components in a slot are assumed to be i.i.d. Let $r_i \sim U[0, R]$ be the uniformly selected distance between user i and BS. We assume that time varying channels are static over a slot and the channel state information is perfectly known for each slot, which is justified under low mobility of cellular IoT applications [60, 70].

The data arrival for user i at slot t in bits is denoted as $\lambda_i(t) \triangleq a_i(t)u_i(t)$, where $a_i(t) \sim \text{Ber}(\pi_i)$ is a Bernoulli random variable with packet arrival probability $\pi_i = \text{Pr}(a_i(t) = 1)$, and $u_i(t) \sim U[\Lambda]$ is the packet bit count uniformly selected from a finite set of positive integers Λ . The deadline for each arriving packet of user i is denoted as m_i . Let $\mu_i(t)$ be the number of bits transmitted to user i at slot t . Transmission is considered on a slot for a user's head of the queue packet, of which the number of remaining data bits and the slot count until expiration are denoted as $q_i(t)$ and $d_i(t)$, respectively.

A packet can be dropped, either by EPD, which is introduced in Section 3.2, or due to deadline expiration. EPD detects packets to be expired and drops them before deadline violation to increase timely throughput. The number of user i packets dropped by EPD in the slot t is denoted by $E_i(t)$. The number of dropped bits due to expiration for user i in slot t is denoted as $D_i(t) = \mathbb{1}\{d_i(t) = 1\} \max[q_i(t) - \mu_i(t), 0]$. Queue backlog of user i in slot t is denoted as $Q_i(t)$, where $Q_i(0) = 0$. The queue dynamics is modeled as $Q_i(t) \triangleq \max[Q_i(t) - \mu_i(t), 0] + \lambda_i(t) - D_i(t) - E_i(t)$. Finally, the packet drop rate due to expiration for user i is represented as $\bar{D}_i \triangleq \lim_{t \rightarrow \infty} t^{-1} \sum_{\tau=0}^{t-1} \mathbb{1}\{D_i(\tau) > 0\}$, and the overall drop rate is denoted as $\bar{D} \triangleq N^{-1} \sum_{i=1}^N \lim_{t \rightarrow \infty} t^{-1} (\sum_{\tau=0}^{t-1} \mathbb{1}\{D_i(\tau) > 0\} + E_i(\tau))$. We model the packet drop rate minimization problem under time average power constraints

as:

$$\min_{\mathbf{P}(t)} \sum_{i=1}^N \overline{D}_i \quad (5.1a)$$

$$\text{s.t. } \overline{P}_i \leq \gamma_i, \forall i \in \mathcal{N} \quad (5.1b)$$

$$\sum_i \mathbb{1}\{P_i(t) > 0\} \leq 2 \quad (5.1c)$$

A data packet is fragmented at the source for being transmitted in different slots, and it is reassembled at the destination. A packet is considered to be successfully transmitted only when all its fragments are successfully transmitted before deadline expiration. Thus, only FTM is considered for RSMA. Since the cost of a packet drop event is observed at the moment of a dropping event, a slot by slot solution for (5.1) is not feasible under the random nature of the CQSI. In Section 3.3, we proposed a method to quantify the cost of taken decisions on packet dropping events, and reformulated (5.1) accordingly as

$$\min_{\mathbf{P}(t)} \sum_{i=1}^N \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathcal{F}_i \quad (5.2a)$$

$$\text{s.t. } \overline{P}_i \leq \gamma_i, \forall i \in \mathcal{N} \quad (5.2b)$$

$$\sum_i \mathbb{1}\{P_i(t) > 0\} \leq 2 \quad (5.2c)$$

where $\mathcal{F}_i \triangleq (\frac{q_i(t) - \mu_i(t)}{q_i(t)}) (\frac{m_i - (d_i(t) - 1)}{m_i})^{\alpha_i(t)}$ is the novel cost function which is introduced in (3.17) for quantifying future consequences of taken decisions, and $\alpha_i(t)$ is a non-negative parameter which is proposed to adjust the importance of remaining deadline in the objective function. Slot by slot-based solutions can be applied to (5.2). In Section 3.3, we applied the *drift – plus – penalty* technique [42] to solve (5.2) and we developed a dynamic power assignment algorithm. At every slot t , the following *drift – plus – penalty* problem in (5.3) is solved for $\mathbf{P}(t)$ using the observed CQSI:

$$\min_{\mathbf{P}(t)} \mathcal{M}(\mathbf{P}(t)) = \sum_{i=1}^N (V \mathcal{F}_i + X_i(t)(P_i(t) - \gamma_i)) \quad (5.3a)$$

$$\text{s.t. } \sum_i \mathbb{1}\{P_i(t) > 0\} \leq 2 \quad (5.3b)$$

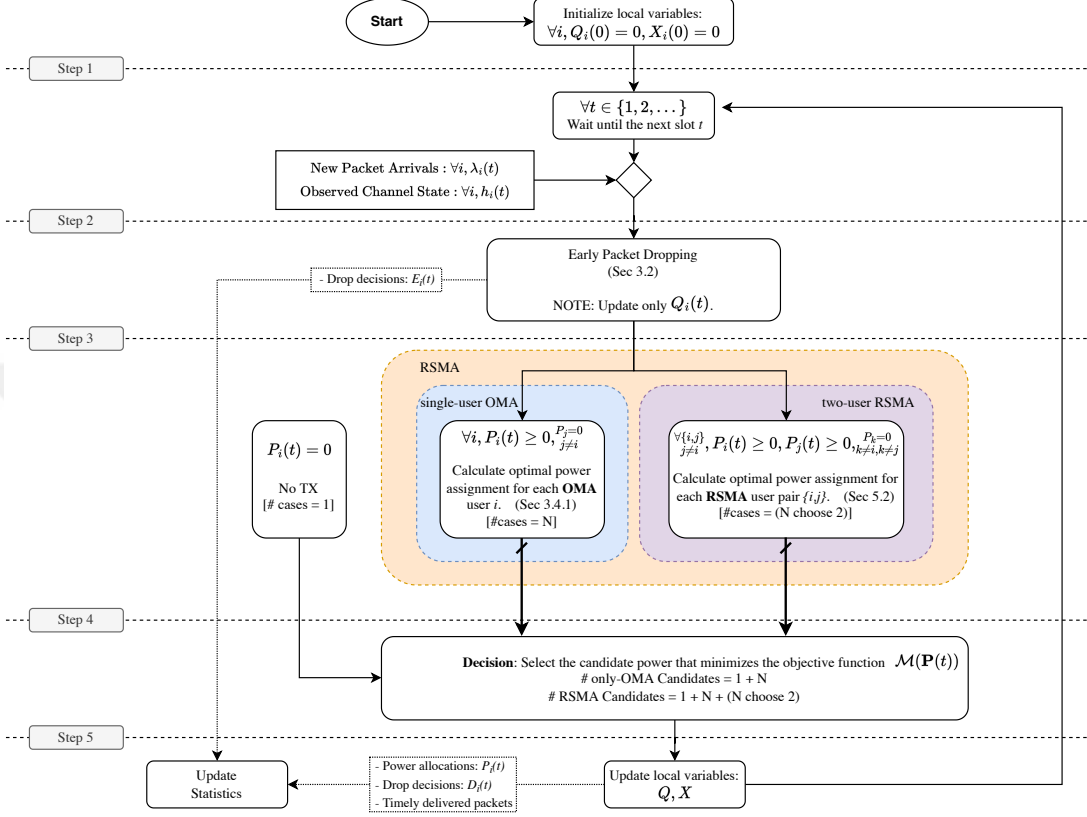


Figure 5.1: Flow chart of the proposed algorithm for timely throughput maximization using RSMA.

where $V > 0$ is a weight parameter which is used to scale the penalty $\mathcal{F}_i$ regarding time average constraints, and $X_i(t)$ is a virtual queue representing the time average constraints in (5.2b).

The flow chart of the proposed dynamic power allocation algorithm for timely throughput maximization with RSMA is presented in Fig 5.1. The algorithm commences with the initialization of the local variables. The step 1 involves waiting for the next slot to begin, acquiring fresh packets, and observing the channels state. In the second step, EPD is utilized to drop packets that are detected as being about to expire. In step 3, we derive the optimal power assignments for candidate OMA and two-user RSMA transmissions. The constraint in (5.3b) refers to N and $\binom{N}{2}$ candidate single-user OMA and two-user RSMA transmissions, respectively. In the fourth step, the optimization metric in (5.3) is exhaustively minimized. One of the minimizing candidates is selected randomly

for the sake of fairness among users. Local variables and statistics are updated in step 5 according to the decision. The algorithm proceeds to Step 1 to await the onset of the next slot.

In the following section, optimal power assignment for two-user RSMA is derived.

5.2 Optimal Power Assignment for two-user layer-1 RSMA

In this section, we derive the optimal power assignment for two-user layer-1 RSMA with paired users $\{i, j\}$, such that $i \neq j$. Assume without loss of generality that $|h_j|^2 > |h_i|^2$. Since we perform a theoretical performance analysis for RSMA, SIC power constraint is not necessary [73]. In RSMA, the transmission rate of the common stream is determined based on the user with the weak channel, so that, both users can decode the common stream message. The rate of the common and private streams of user $k \in \{i, j\}$ are expressed in (5.4) and (5.5), respectively:

$$\mu_{(i,j)}^c = \tau \mathcal{B} \left(\log_2 \left(1 + \frac{|h_i|^2 P_{(i,j)}^c}{|h_i|^2 P_{(i,j)}^p + \sigma^2} \right) - C_i^F \right) \quad (5.4)$$

$$\mu_{k,(i,j)}^p = \tau \mathcal{B} \left(\log_2 \left(\frac{|h_k|^2 P_{(i,j)}^p + \sigma^2}{|h_k|^2 (P_{(i,j)}^p - P_{k,(i,j)}^p) + \sigma^2} \right) - C_k^F \right) \quad (5.5)$$

where $\sigma^2 = \mathcal{B}N_0$, N_0 is noise power spectral density, $C_k^F = \sqrt{\frac{\mathcal{V}_k}{\tau \mathcal{B}}} \frac{Q^{-1}(\epsilon^R)}{\ln 2}$, ϵ_R is the BLER for RSMA, $P_{(i,j)}^c$ is the common stream power, $P_{k,(i,j)}^p$ is the user k private stream power and $P_{(i,j)}^p = \sum_k P_{k,(i,j)}^p$. In order to make sure that deadlines are satisfied with high likelihood, we assume that SNR is high, which results in $\mathcal{V}_k \approx 1$. Let $a_{k,(i,j)}$ be the common rate allocated for user k . Common stream rate, $\mu_{(i,j)}^c$, and total transmission rate for user k , $\mu_{k,(i,j)}$, can be calculated as,

$$\mu_{(i,j)}^c = \sum_k a_k \quad (5.6)$$

$$\mu_{k,(i,j)} = a_{k,(i,j)} + \mu_{k,(i,j)}^p, \quad k \in \{i, j\} \quad (5.7)$$

In this thesis, we define the RSMA-oriented allocated power to the user k , which is denoted as $P_{k,(i,j)}$, as the sum of respective private stream power, $P_{k,(i,j)}^p$ and the portion of common stream power proportional to the allocated common rate for the user, $P_{(i,j)}^c(a_{k,(i,j)}/\mu_{(i,j)}^c)$, as follows,

$$P_{k,(i,j)} = P_{k,(i,j)}^p + P_{(i,j)}^c \frac{a_{k,(i,j)}}{\mu_{(i,j)}^c}, \quad k \in \{i, j\} \quad (5.8)$$

Let the non-negative decision parameters $\{a_{i,(i,j)}, a_{j,(i,j)}, P_{(i,j)}^c, P_{i,(i,j)}^p, P_{j,(i,j)}^p\}$ be represented as, $\{a_i, a_j, P_c, P_{p,i}, P_{p,j}\}$ for simplicity. Let $P_p = P_{p,i} + P_{p,j}$ be the total private stream power and $P_T = P_c + P_p$ be the total-allocated power. Then, RSMA based objective function is presented as follows:

$$\begin{aligned} \mathcal{M}_{(i,j)}^R = & - \sum_{k \in \{i,j\}} \Gamma_k \left(\frac{a_k}{\tau \mathcal{B}} + \log_2 \left(\frac{|h_k|^2 P_p + \sigma^2}{|h_k|^2 (P_p - P_{p,k}) + \sigma^2} \right) \right) \\ & \sum_{k \in \{i,j\}} X_k \left(P_{p,k} + \left(\frac{a_k}{a_j + a_i} \right) P_c \right) + C_{(i,j)}^R \end{aligned} \quad (5.9)$$

where Γ_k is defined in (3.32c) and $C_{(i,j)}^R$ is a constant. The RSMA based optimization problem can be stated as follows:

$$\arg \min_{a_i, a_j, P_c, P_{p,i}, P_{p,j}} \mathcal{M}_{(i,j)}^R(a_i, a_j, P_c, P_{p,i}, P_{p,j}) \quad (5.10a)$$

$$\text{s.t. } P_T \leq P_0, \quad (5.10b)$$

$$\mu_{k,(i,j)} \leq q_k, \quad \forall k \quad (5.10c)$$

where (5.10b) is the total power budget constraint and (5.10c) represents the maximum rate constraint.

We present a two-step solution for the above optimization problem. At first, we present optimal power assignment for given $\{a_i, a_j\}$. Secondly, we determine the reduced search space for the optimal $\{a_i, a_j\}$.

5.2.0.1 Optimal Power Assignment for Given Common Rate

In the first step, we derive the optimal total allocated power value P_T for given common rates $\{a_i, a_j\}$. Let $b_k = q_k - a_k$ be the number of bits required to be transmitted through the private stream of user k , such that $\mu_{k,(i,j)}^p \leq b_k$. Let $P_{k,(i,j)}^{P,req}$ be the required private stream power value for user pair $k \in \{i, j\}$, such that

$$\mu_{k,(i,j)}^p \Big|_{(P_{p,i}=P_{i,(i,j)}^{P,req}, P_{p,j}=P_{j,(i,j)}^{P,req})} = b_k. \quad (5.11)$$

The relation between b_k and the respective required power levels is as follows,

$$b_j = \tau \mathcal{B} \left(\log_2 \left(1 + \frac{h_j P_{j,(i,j)}^{P,req}}{h_j P_{i,(i,j)}^{P,req} + \sigma^2} \right) - C^F \right) \quad (5.12)$$

$$b_i = \tau \mathcal{B} \left(\log_2 \left(1 + \frac{h_i P_{i,(i,j)}^{P,req}}{h_i P_{j,(i,j)}^{P,req} + \sigma^2} \right) - C^F \right) \quad (5.13)$$

In order to satisfy (5.10c), we solve the system of equations in (5.12)-(5.13) to calculate required private stream power value $P_{k,(i,j)}^{P,req}$ for user pair $k \in \{i, j\}$, as follows,

$$P_{j,(i,j)}^{P,req} = - \frac{\left(2^{\left(\frac{b_j}{\tau \mathcal{B}} + C^F \right)} - 1 \right) \left(\left(2^{\left(\frac{b_i}{\tau \mathcal{B}} + C^F \right)} - 1 \right) \frac{\sigma^2}{h_i} + \frac{\sigma^2}{h_j} \right)}{\left(2^{\left(\frac{b_j}{\tau \mathcal{B}} + C^F \right)} - 1 \right) \left(2^{\left(\frac{b_i}{\tau \mathcal{B}} + C^F \right)} - 1 \right) - 1} \quad (5.14)$$

$$P_{i,(i,j)}^{P,req} = - \frac{\left(2^{\left(\frac{b_i}{\tau \mathcal{B}} + C^F \right)} - 1 \right) \left(\left(2^{\left(\frac{b_j}{\tau \mathcal{B}} + C^F \right)} - 1 \right) \frac{\sigma^2}{h_j} + \frac{\sigma^2}{h_i} \right)}{\left(2^{\left(\frac{b_j}{\tau \mathcal{B}} + C^F \right)} - 1 \right) \left(2^{\left(\frac{b_i}{\tau \mathcal{B}} + C^F \right)} - 1 \right) - 1} \quad (5.15)$$

Note that, $P_{k,(i,j)}^{P,req} > 0$ only if $\prod_k (2^{(b_k/(\tau \mathcal{B}) + C^F)} - 1) < 1$, in (5.14)-(5.15). Otherwise, there exists no positive finite value to represent the upper limit of P_p . Thus, the upper limit of P_p is considered to be ∞ under $\prod_k (2^{(b_k/(\tau \mathcal{B}) + C^F)} - 1) \geq 1$. Let $P_{(i,j)}^{P,req}$ be total required private stream power as,

$$P_{(i,j)}^{P,req} = \begin{cases} \sum_k P_{k,(i,j)}^{P,req} & , \text{ if } \prod_k (2^{(b_k/(\tau \mathcal{B}) + C^F)} - 1) < 1 \\ \infty & , \text{ otherwise} \end{cases} \quad (5.16)$$

The required common stream power $P_{(i,j)}^{C,req}$ to transmit common rates $\{a_i, a_j\}$ along with private streams can be derived using (5.4). Total required power $P_{(i,j)}^{T,req} = P_{(i,j)}^{P,req} + P_{(i,j)}^{C,req}$ can be expressed as,

$$P_{(i,j)}^{T,req} = \begin{cases} 2^{\left(\frac{a_i+a_j}{\tau B}\right)} \left(P_{(i,j)}^{P,req} + \frac{\sigma^2}{h_i} \right) - \frac{\sigma^2}{h_i} & , \text{if } \prod_k (2^{(b_k/(\tau B)+C^F)} - 1) < 1 \\ \infty & , \text{otherwise} \end{cases} \quad (5.17)$$

Let $P_{(i,j)}^{T,max} = \min(P_0, P_{(i,j)}^{T,req})$ be the maximum available total power according to the constraints (5.10b)-(5.10c). The minimum total power $P_{(i,j)}^{T,min}$ is the power required to transmit only common rates $\{a_i, a_j\}$, as follows,

$$P_{(i,j)}^{T,min} = \frac{\sigma^2}{|h_i|^2} \left(2^{\left(\frac{a_i+a_j}{\tau B} + C^F\right)} - 1 \right) \quad (5.18)$$

Then, we have $P_{(i,j)}^{T,min} \leq P_T \leq P_{(i,j)}^{T,max}$. Note that, due to the limitation of total stream power with $P_{(i,j)}^{T,min} \leq P_T \leq P_{(i,j)}^{T,max}$, the transmitted number of bits will never exceed the remaining number of packet bits. In this case, the constraint (5.10c) is always satisfied.

We can rewrite (5.9) in terms of $\{a_i, a_j, P_T\}$ as follows:

$$\begin{aligned} \mathcal{M}_{(i,j)}^R(a_i, a_j, P_T) &= - \sum_{k \in \{i,j\}} \Gamma_k \left(\frac{a_k}{a_j + a_i} \right) \log_2 \left(\frac{|h_i|^2 P_T + \sigma^2}{|h_i|^2 P_p + \sigma^2} \right) \\ &\quad - \sum_{k \in \{i,j\}} \Gamma_k \log_2 \left(\frac{|h_k|^2 P_p + \sigma^2}{|h_k|^2 (P_p - P_{p,k}) + \sigma^2} \right) \\ &\quad + \sum_{k \in \{i,j\}} X_k \left(P_{p,k} + \left(\frac{a_k}{a_j + a_i} \right) (P_T - P_p) \right) + C_{(i,j)}^R \quad (5.19) \\ &= - \log_2 (|h_i|^2 P_T + \sigma^2) \sum_k \frac{a_k}{a_i + a_j} \Gamma_k \\ &\quad + \sum_k X_k \left(P_{p,k} + \left(\frac{a_k}{a_j + a_i} \right) P_T \right) + C' \end{aligned}$$

where C' represents the parameters independent of $\{a_i, a_j, P_T\}$. Note that $\mathcal{M}_{(i,j)}^R(a_i, a_j, P_T)$ is a convex function of P_T . Then, the optimal P_T is expressed

as follows:

$$P_T = \begin{cases} P_{(i,j)}^{T,max} & , \text{if } P_{(i,j)}^{T,max} \leq P_T^*(a_i, a_j) \\ P_T^*(a_i, a_j) & , \text{if } P_{(i,j)}^{T,min} \leq P_T^*(a_i, a_j) \leq P_{(i,j)}^{T,max} \\ P_{(i,j)}^{T,min} & , \text{otherwise} \end{cases} \quad (5.20)$$

where $P_T^*(a_i, a_j)$ is the global minimizer as,

$$P_T^*(a_i, a_j) = \frac{1}{\log(2)} \left(\frac{\Gamma_j a_j + \Gamma_i a_i}{X_j a_j + X_i a_i} \right) - \frac{\sigma^2}{|h_i|^2} \quad (5.21)$$

Optimal P_c and P_p values can be derived using the optimal P_T and the given $\{a_i, a_j\}$ values, as follows,

$$P_p^*(a_i, a_j) = \frac{P_T(a_i, a_j) + \frac{\sigma^2}{h_i}}{2^{\left(\frac{a_i + a_j}{\tau \cdot \mathcal{B}}\right)}} - \frac{\sigma^2}{h_i} \quad (5.22)$$

$$P_c^*(a_i, a_j) = \left(P_T(a_i, a_j) + \frac{\sigma^2}{h_i} \right) \frac{\left(2^{\left(\frac{a_i + a_j}{\tau \cdot \mathcal{B}}\right)} - 1 \right)}{2^{\left(\frac{a_i + a_j}{\tau \cdot \mathcal{B}}\right)}} \quad (5.23)$$

Note that, the objective function in (5.9) is a concave function with respect to $P_{p,i}$ and $P_{p,j}$ given the optimal P_T and P_p values. Therefore, the optimal $\{P_{p,i}, P_{p,j}\}$ values can be determined by comparing two extreme cases when the value of P_p is given. These extreme cases can be expressed as aiming to completely transmit user i or user j private stream packets, respectively. In either case, the remaining power is considered for the other user.

Let $P_{p,j}^\dagger$ and $P_{p,i}^\dagger$ represent the private stream power of the user j under these two extreme cases as follows,

$$b_j = \tau \mathcal{B} \left(\log_2 \left(\frac{h_j P_p + \sigma^2}{h_j (P_p - P_{p,j}^\dagger) + \sigma^2} \right) - C^F \right) \quad (5.24)$$

$$b_i = \tau \mathcal{B} \left(\log_2 \left(\frac{h_i P_p + \sigma^2}{h_i P_{p,i}^\dagger + \sigma^2} \right) - C^F \right) \quad (5.25)$$

The first case is presented in (5.24), where the user j 's private stream is prioritized. Thus, $P_{p,j}^\dagger$ can be derived as,

$$P_{p,j}^\dagger = \left(P_p + \frac{\sigma^2}{h_j} \right) \left(1 - \frac{1}{2^{\left(\frac{b_j}{\tau B} + C^F\right)}} \right). \quad (5.26)$$

Note that, if $P_{p,j}^\dagger < P_p$, then user j 's b_j bits of private stream data can be completely transmitted and remaining power can be considered for user i 's private stream. Otherwise, there will be only user j 's private stream due to concavity. The second case is presented in (5.25), where the user i 's private stream is prioritized. Thus, $P_{p,j}^\dagger$ can be derived as,

$$P_{p,j}^\dagger = \frac{P_p + \frac{\sigma^2}{h_i}}{2^{\left(\frac{b_i}{\tau B} + C^F\right)}} - \frac{\sigma^2}{h_i} \quad (5.27)$$

Similarly, if $P_{p,j}^\dagger > 0$, then user i 's b_i bits of private stream data can be completely transmitted and remaining power can be considered for user j 's private stream. Otherwise, there will be only user i 's private stream due to concavity.

Let $\mathbb{P}_j^p$ be the set of user j private stream values as,

$$\mathbb{P}_j^p \in \left\{ \min \left(P_p, P_{p,j}^\dagger \right), \max \left(0, P_{p,j}^\dagger \right) \right\} \quad (5.28)$$

Note that, $\mathbb{P}_j^p$ may refer to $\mu_{k,(i,j)}^p < 1$ due to the FBL effect in (5.5). In such cases, the optimal decision is to set the respective private stream power as zero to avoid waste of resources, and the rest of the decision parameters $\{P_p, P_T\}$ must be updated according to the given common rates $\{a_i, a_j\}$.

Using the derivation for the optimal $P_p(a_i, a_j)$ and $P_c(a_i, a_j)$, we can define an equivalent problem formulation, that depends solely on $\{a_i, a_j, P_{p,j}\}$, as follows:

$$\min_{(a_i, a_j) \in \mathbb{A}, P_j \in \mathbb{P}_j} \mathcal{M}_{(i,j)}^R(a_i, a_j, P_{p,j}) \quad (5.29)$$

where

$$\begin{aligned}
\mathcal{M}_{(i,j)}^R(a_i, a_j, P_{p,j}) = & -\Gamma_j \left(\frac{a_j}{\tau\mathcal{B}} + \log_2 \left(\frac{h_j P_p(a_i, a_j) + \sigma^2}{h_j (P_p(a_i, a_j) - P_{p,j}) + \sigma^2} \right) \right) \\
& -\Gamma_i \left(\frac{a_i}{\tau\mathcal{B}} + \log_2 \left(\frac{h_i P_p(a_i, a_j) + \sigma^2}{h_i P_{p,j} + \sigma^2} \right) \right) \\
& + X_j \left(P_{p,j} + \left(\frac{a_j}{a_j + a_i} \right) P_c(a_i, a_j) \right) \\
& + X_i \left((P_p(a_i, a_j) - P_{p,j}) + \left(\frac{a_i}{a_i + a_j} \right) P_c(a_i, a_j) \right)
\end{aligned} \tag{5.30}$$

$$\mathbb{A} = \begin{cases} a_j + a_i \leq \min(q_j + q_i, R_c^{max}), \\ a_j \leq q_j, \\ a_i \leq q_i, \\ a_i, a_j \geq 0. \end{cases} \tag{5.31}$$

$$R_c^{max} = \tau\mathcal{B} \left(\log_2 \left(1 + \frac{P_0 \cdot h_i}{\sigma^2} \right) - C^F \right) \tag{5.32}$$

Note that, R_c^{max} represents the maximum achievable common stream rate, for which the total power budget P_0 is solely assigned to the common stream. However, the search space of $\mathbb{A}$ can be huge for the given q_i and q_j values, and the brute-force search for the optimal result becomes infeasible.

In the next subsection, we present an efficient search algorithm.

5.2.0.2 Reduced Search Space for Optimal Common Rate

In this part, we derive a significantly reduced search space compared to $\mathbb{A}$ for optimal common bit rates (a_i, a_j) . Let r_k^c be the ratio of common rate allocated for user k as,

$$r_k^c = \frac{a_k}{a_j + a_i} \tag{5.33}$$

Note that $0 \leq r_k^c \leq 1$. We rewrite the global minimizer $P_T^*(a_i, a_j)$ in (5.21) in terms of r_j^c as,

$$P_T^*(r_j^c) = \frac{1}{\log(2)} \left(\frac{\Gamma_j r_j^c + \Gamma_i (1 - r_j^c)}{X_j r_j^c + X_i (1 - r_j^c)} \right) - \frac{\sigma^2}{|h_i|^2} \tag{5.34}$$

Then, we can rewrite the objective function as $\mathcal{M}(r_j^c)$ by substituting r_j^c and $P_T^*(r_j^c)$ in (5.19) as follows:

$$\begin{aligned} \mathcal{M}_{(i,j)}^R(r_j^c) = & -\log_2(|h_i|^2 P_T^*(r_j^c) + \sigma^2) \sum_k r_k^c \Gamma_k \\ & + \sum_k X_k (r_k^c (P_T^*(r_j^c) - P_p) + C'') \end{aligned} \quad (5.35)$$

where C'' represents the parameters independent of r_j^c . We calculate the second derivative of the objective function as

$$\frac{d^2 \mathcal{M}(r_j^c)}{d(r_j^c)^2} = - \frac{(\Gamma_j X_i - \Gamma_i X_j)^2}{\log(2) (\Gamma_i \cdot (1 - r_j^c) + \Gamma_j \cdot r_j^c) \cdot (X_i \cdot (1 - r_j^c) + X_j \cdot r_j^c)^2} \quad (5.36)$$

Note that $\frac{d^2 \mathcal{M}(r_j^c)}{d(r_j^c)^2} < 0$ for $0 \leq r_j^c \leq 1$, and thus, $\mathcal{M}(r_j^c)$ is a concave function of r_j^c . Then we can conclude that optimal policy for assignment of the common stream power is based on favoring only one of the users until its whole packet data is completed. Then, the search space $\mathbb{A}$ can be reduced to $\bar{\mathbb{A}} \in \bar{\mathbb{A}}_1 \cup \bar{\mathbb{A}}_2 \cup \bar{\mathbb{A}}_3 \cup \bar{\mathbb{A}}_4$ as follows:

$$\bar{\mathbb{A}}_1 = \{a_i = 0, 0 \leq a_j \leq \min(q_j, R_c^{max})\} \quad (5.37)$$

$$\bar{\mathbb{A}}_2 = \{a_j = 0, 0 \leq a_i \leq \min(q_i, R_c^{max})\} \quad (5.38)$$

$$\bar{\mathbb{A}}_3 = \left\{ a_i = \min(q_i, R_c^{max}), \min(q_i, R_c^{max}) \leq \sum_k a_k \leq \min\left(\sum_k q_k, R_c^{max}\right) \right\} \quad (5.39)$$

$$\bar{\mathbb{A}}_4 = \left\{ a_j = \min(q_j, R_c^{max}), \min(q_j, R_c^{max}) \leq \sum_k a_k \leq \min\left(\sum_k q_k, R_c^{max}\right) \right\} \quad (5.40)$$

Thus, the reduced search space $\bar{\mathbb{A}}$ consists of the boundary elements of $\mathbb{A}$, which reduces the size of the search space from $\mathcal{O}(q_i q_j)$ to $\mathcal{O}(q_i + q_j)$.

In the following chapter, simulation results and discussions are presented for RSMA.

Chapter 6

Numerical Results and Discussions for RSMA

In this chapter, the performance of RSMA is compared with other MA schemes in terms of timely throughput. Letting α, γ, m, π be the system parameters such that $\alpha = \alpha_i(t)$, $\gamma = \gamma_i$, $m = m_i$, $\pi = \pi_i$ for all t and $i \in \mathcal{N}$, the timely throughput can be represented as $\pi - \overline{D}$. Simulations are repeated using 100 random seeds and their averages are reported. FI in (4.1) is considered for the analysis of fairness among users. In Table 6.1, default values for all parameters are presented. EPD is used for all algorithms. The simulations are implemented using MATLAB. The algorithms compared are:

- s-OMA: Optimal power assignment for single-user OMA only, which is presented in Section 3.4.1.
- H-NOMA: Optimal power assignment for Hybrid MA, which is presented in Section 3.4.
- RSMA: Dynamic power assignment algorithm presented in Sections 5.1-5.2 is considered. At each slot, RSMA occupies the best of optimal power assignments for single-user OMA in Section 3.4.1 and two-user RSMA in Section 5.2.

Table 6.1: Simulation parameters and default values

Λ (Packet size)	$(160 \cdot w)_{w=1}^{20}$ bits
γ (Average power constraint for all users)	0.6 W
π (Arrival rate for all users)	0.3
R (Cell radius)	50 m
$\#Slots$ (Simulation slot count)	10^4
τ (Slot duration)	0.1 ms
$\mathcal{B}$ (Bandwidth)	1 MHz
P_0 (Power budget)	3 W
N (User count)	5
m (Deadline slot count)	5
Path loss	$35.3 + 37.6 \log_{10}(r_i)$ dB
Fast fading component	$CN(0, 1)$
N_0 (Noise power spectral density)	-174 dBm/Hz
ϵ_O (OMA Target BLER)	10^{-5}
ϵ_N (NOMA/RSMA Target BLER [35])	$5 \cdot 10^{-6}$

6.1 Dynamic Power Allocation with EPD

In Figs. 6.1, 6.2, 6.3, $\bar{D}$ with different values of α and V under $\pi \in \{0.1, 0.3, 0.5\}$ are presented. In each plot, the results for $V \in \{10, 100, 1000\}$ are presented to investigate the convergence with respect to V . In all of the plots, there is almost no difference between $V = 100$ and $V = 1000$, which shows us that $V = 100$ is suitable. In Fig 6.1, RSMA performs slightly better than s-OMA and H-NOMA under $\pi = 0.1$, since traffic is low. As traffic load increases towards $\pi = 0.3$ and $\pi = 0.5$, RSMA and H-NOMA perform significantly better than s-OMA. This observation reveals that high connectivity based advantage of RSMA can be achieved with high number of candidate user pairs. Under all traffic loads, RSMA performs the best by leveraging the spectrum better than s-OMA and H-NOMA, showing that RSMA is robust under varying traffic loads. As best results are obtained under $\alpha = 0.1$ in Figs 6.1-6.3, this values is used for the rest of this paper.

In Fig. 6.5, $\bar{P}$ under $\alpha = 0.1$ and $\pi = 0.3$ with different values of γ are presented. RSMA based algorithms consume more power due to better spectral utilization, thus they are more sensitive to the selected weight parameter V as

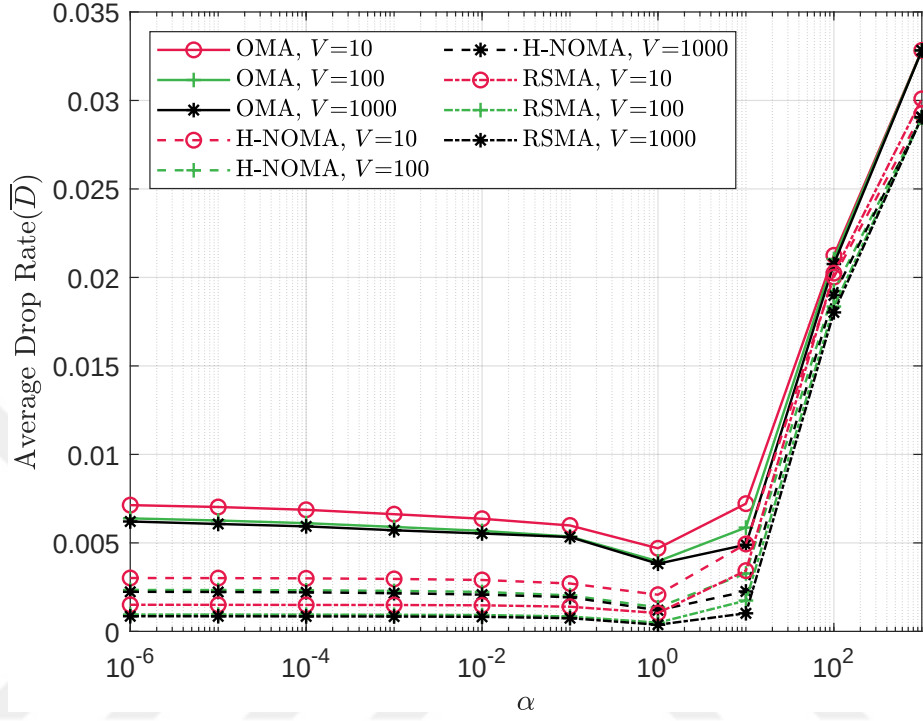


Figure 6.1: Average drop rate $\bar{D}$ as a function of α with $\pi = 0.1$.

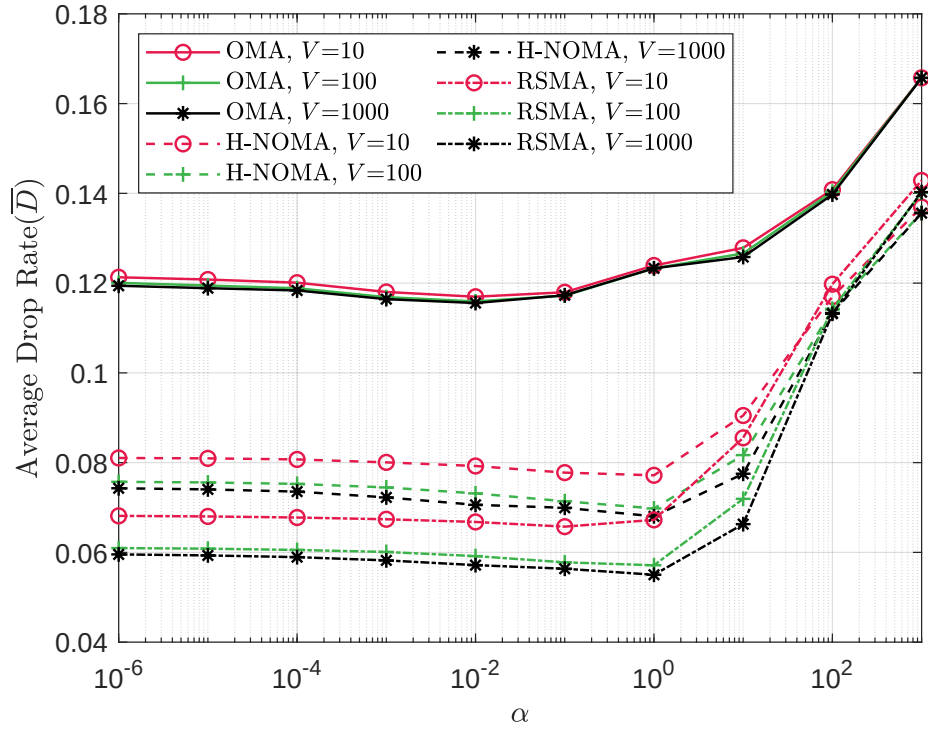


Figure 6.2: Average drop rate $\bar{D}$ as a function of α with $\pi = 0.3$.

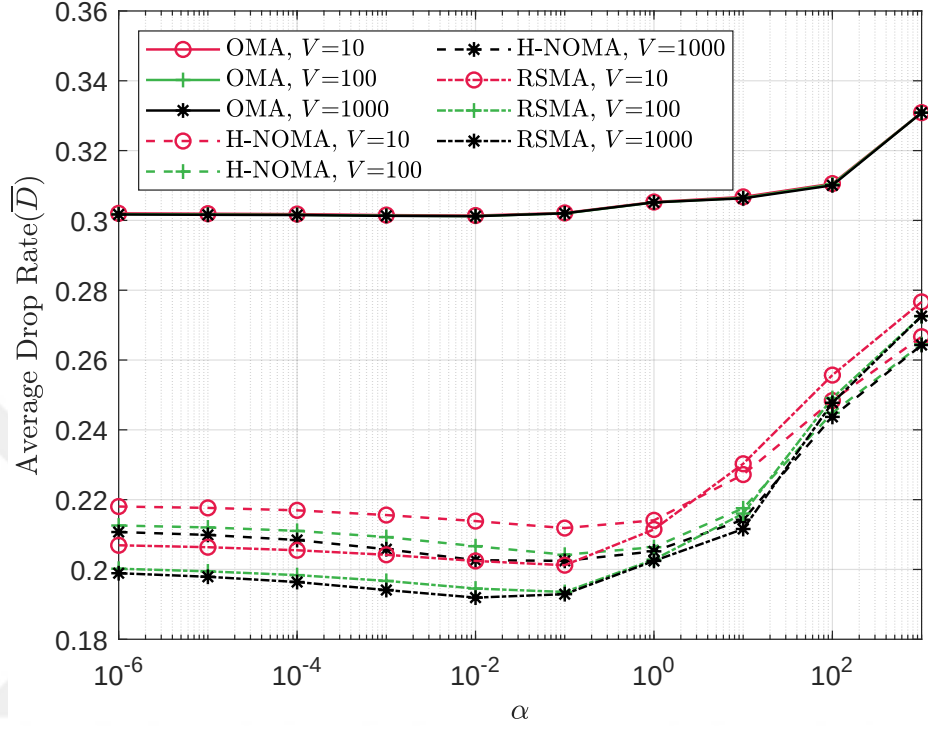


Figure 6.3: Average drop rate $\bar{D}$ as a function of α with $\pi = 0.5$.

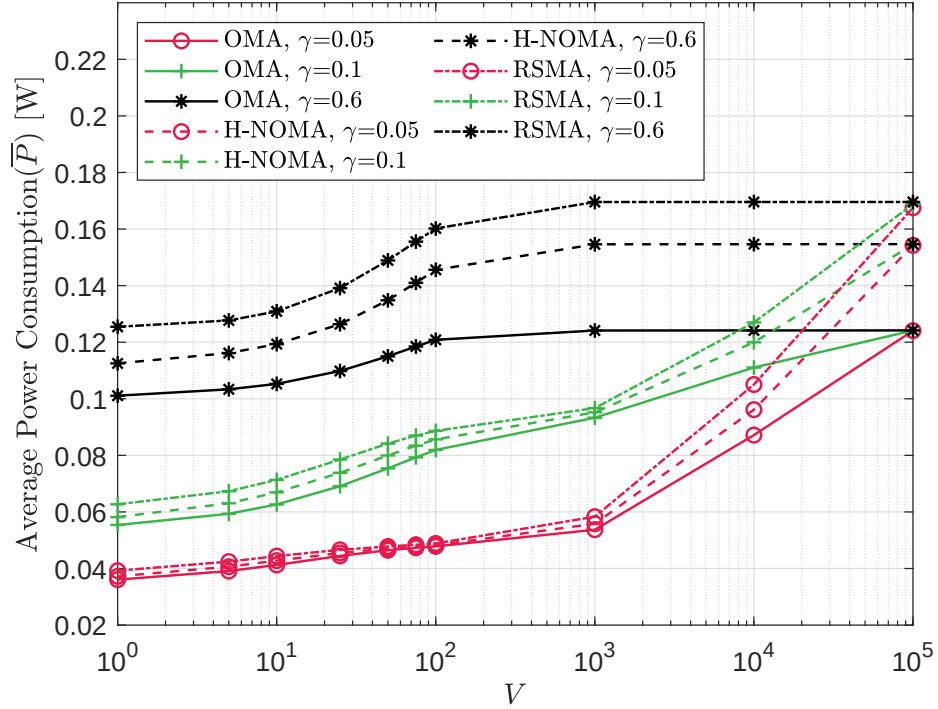


Figure 6.4: Average power $\bar{P}$ as a function of γ with $\pi = 0.1$ and $\alpha = 0.1$.

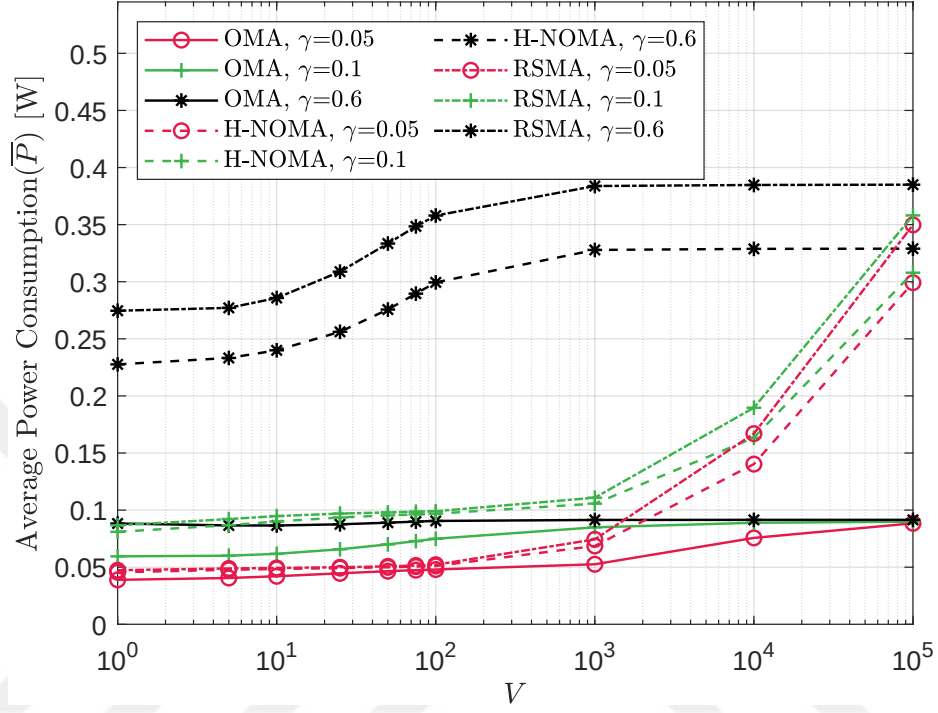


Figure 6.5: Average power $\bar{P}$ as a function of γ with $\pi = 0.3$ and $\alpha = 0.1$.

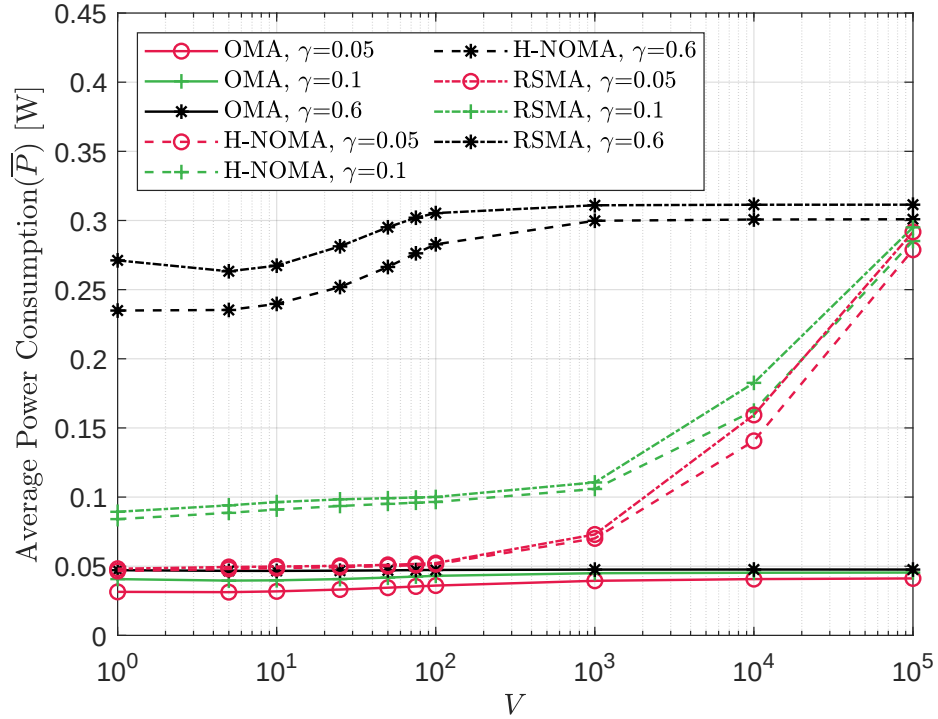


Figure 6.6: Average power $\bar{P}$ as a function of γ with $\pi = 0.1$ and $\alpha = 0.5$.

Table 6.2: In RSMA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	5.25%	22.59%	42.02%
0.1	4.95%	25.54%	46.87%
0.6	4.77%	32.51%	54.75%

Table 6.3: In H-NOMA and s-OMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	3.70%	17.13%	38.48%
0.1	3.55%	19.97%	43.13%
0.6	3.51%	25.06%	49.39%

presented. The weight parameter V must be selected as the highest value that satisfies the average power constraint γ . In Fig. 6.5, $V = 100$ is observed to be the best case. As expected, the RSMA algorithm is highly affected by the selected weight parameter V under the given the average power constraint γ .

In Tables 6.2-6.4, different values of $\gamma \in \{0.05, 0.1, 0.6\}$ and $\pi \in \{0.1, 0.3, 0.5\}$ under $\alpha = 0.01$ and $V = 100$ are considered, and the percentage increase in timely throughput with RSMA compared to s-OMA and H-NOMA are reported. In Table 6.2, as the traffic load increases from $\pi = 0.1$ to $\pi = 0.5$, the timely throughput improvement of RSMA over s-OMA increases from 5.25% to 54.75%. This result shows that as the number of candidate user pairs increases, the benefit of using RSMA increases over single-user OMA. In Table 6.3, H-NOMA behaves

Table 6.4: In RSMA and H-NOMA comparison, percentage of increase in timely throughput for different π and γ values under $\alpha = 0.1$ and $V = 100$.

$\gamma \backslash \pi$	0.1	0.3	0.5
0.05	1.50%	4.66%	2.55%
0.1	1.35%	4.64%	2.61%
0.6	1.22%	5.95%	3.58%

Table 6.5: Mean and variance of FI for H-NOMA, RSMA algorithms and γ values under EPD mechanism, $\pi = 0.1$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
H-NOMA	0.78548	0.80558	0.83928	0.01876	0.019083	0.016854
RSMA	0.59397	0.60091	0.6151	0.01832	0.018857	0.019466

similar to RSMA regarding s-OMA with increasing timely throughput on the average by 22.65%, showing the significant effect of increased connectivity on timely throughput. In Table 6.4, RSMA increases timely throughput on the average by 3.11% compared to H-NOMA. Note that, the traffic level is low in $\pi = 0.1$, and it is mostly handled by all the algorithms in Fig 6.1. Moreover, the traffic level is high in $\pi = 0.5$, so that the existence probability of a suitable user pair for RSMA or NOMA is highest. Thus, the spectrum efficiency is the most critical for $\pi = 0.3$ case. The largest performance increase with RSMA over H-NOMA is observed, as up to 5.95%, under $\pi = 0.3$ in Table 6.4. RSMA utilizes the spectrum much better than s-OMA and H-NOMA under varying traffic and power conditions.

In Tables 6.5-6.6-6.7, the mean and variance of FI for H-NOMA, RSMA algorithms and different γ values are presented under traffic levels $\pi \in \{0.1, 0.3, 0.5\}$, respectively. In Table 6.5, FI of RSMA with different γ values are lower than H-NOMA under low traffic level of $\pi = 0.1$. As the traffic level increases towards $\pi = 0.3$ and $\pi = 0.5$, the gap between FI of RSMA and H-NOMA further reduces, as presented in Tables 6.6-6.7. This difference is caused by the way per-user power consumption is formulated for H-NOMA and RSMA algorithms. In (5.8), the RSMA's common stream power is divided between users proportional to their respective common stream transmission rates. Then, they are accumulated with each user's private stream power to calculate per-user power consumption for RSMA. However, per-user power consumption for NOMA is the power assigned for non-orthogonal stream of a user. This basic difference between the formulation of the user power consumption for RSMA and NOMA leads to alteration within the resultant timely throughput under a given average power constraint.

Table 6.6: Mean and variance of FI for H-NOMA, RSMA algorithms and γ values under EPD mechanism, $\pi = 0.3$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
H-NOMA	0.86126	0.86952	0.89436	0.011474	0.011491	0.010656
RSMA	0.80843	0.80362	0.81076	0.009411	0.009496	0.008874

Table 6.7: Mean and variance of FI for H-NOMA, RSMA algorithms and γ values under EPD mechanism, $\pi = 0.5$, $\alpha = 0.1$ and $V = 100$.

	Mean of Fairness Index			Variance of Fairness Index		
	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.6$
H-NOMA	0.90849	0.91176	0.92646	0.0064252	0.0062855	0.0056765
RSMA	0.90577	0.90417	0.9125	0.0053039	0.0053668	0.004726

6.2 Complexity Reduction Heuristic

In this section, the performance of computational load reduction heuristic *BEST* is presented for H-NOMA and RSMA algorithms in comparison with *RAND*. Simulations are performed with $\alpha = 0.1$ and $V = 100$, which are the values optimized in Section 4.1. EPD is utilized in all simulations. Simulations are performed for 100 seeds and their average performance is presented. We formulated the computation time of the proposed algorithm as a function of Y as presented in (4.2). We prepared plots to combine normalized timely throughput and normalized computation time for different values of Y , as considered in Section 4.3. Normalized timely throughput and normalized computation time are presented on the left and right y-axis, respectively, of the plots. The normalized computation time is derived according to the formulation presented in (4.2). The user count of the NOMA/RSMA subset Y is illustrated on the x-axis of the plots.

In Figs 6.7-6.8-6.9, normalized timely throughput obtained with H-NOMA and RSMA using *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 5$ and the traffic levels $\pi \in \{0.1, 0.3, 0.5\}$, respectively. In all simulations, *BEST* performs better compared to *RAND*. When user count is low with $N = 5$ and traffic is up to a

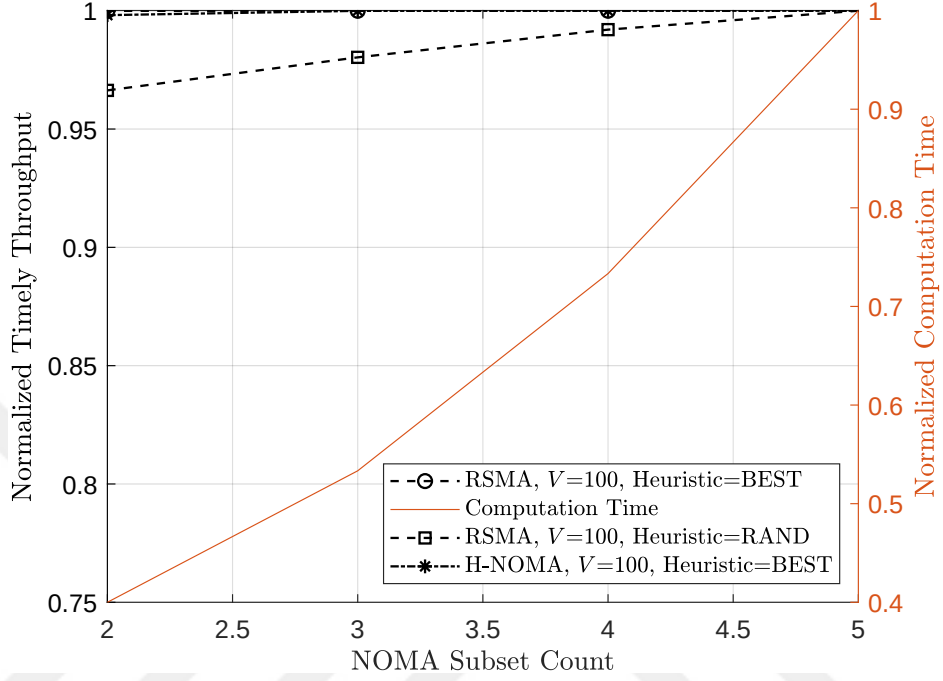


Figure 6.7: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 5$ and $\pi = 0.1$.

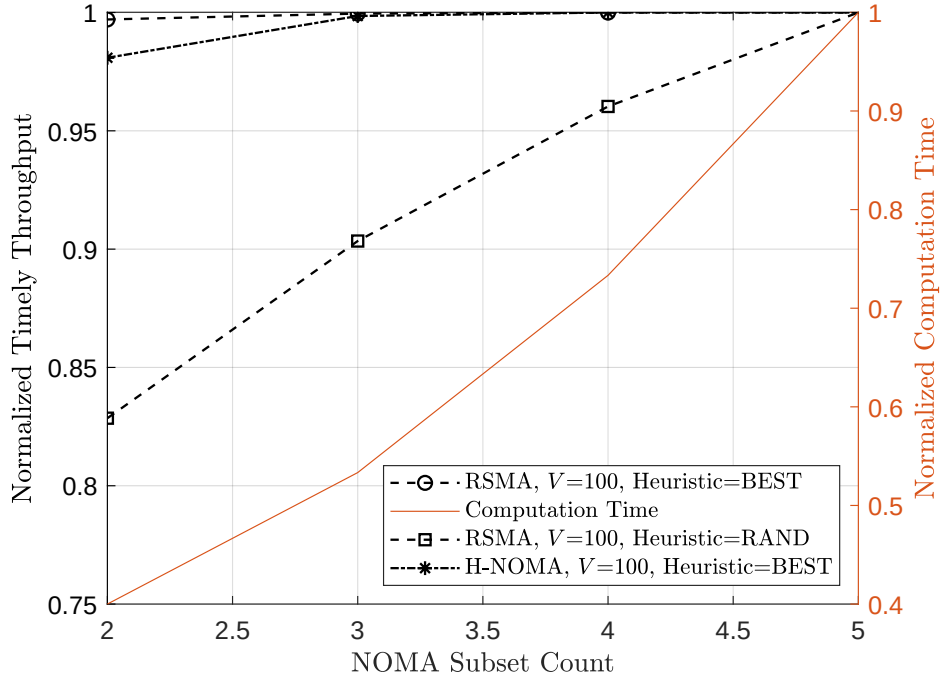


Figure 6.8: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 5$ and $\pi = 0.3$.

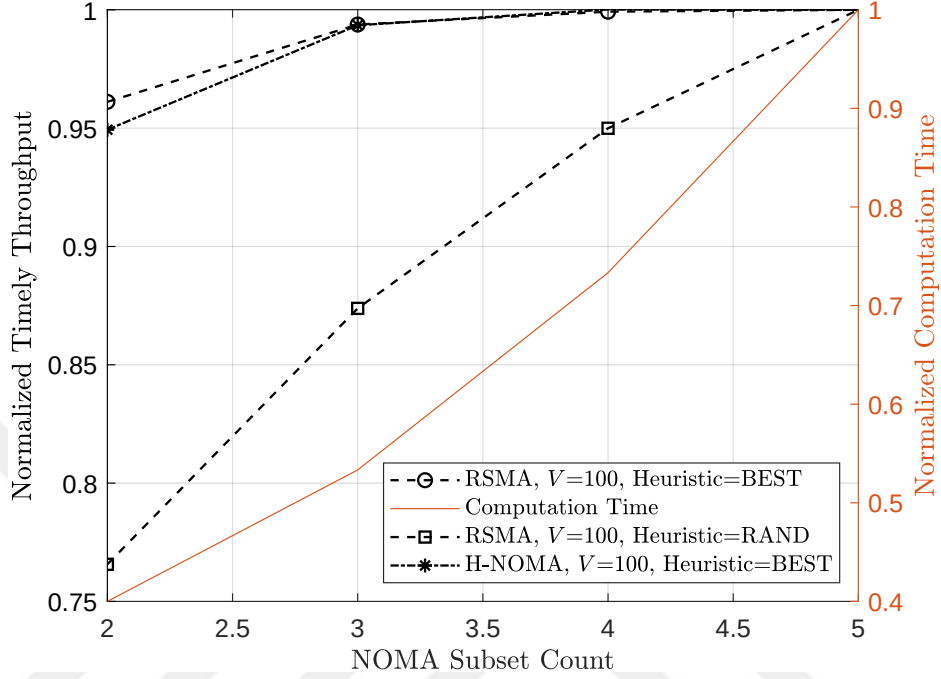


Figure 6.9: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 5$ and $\pi = 0.5$.

medium level with $\pi \in \{0.1, 0.3\}$, the RSMA algorithm with *BEST* mostly preserves timely throughput down to $Y = 2$, which refers to about 60% reduction in the computation time. Whereas, the H-NOMA algorithm with *BEST* mostly preserves timely throughput down to $Y = 3$ in this scenario. The cause of this difference is the performance increase achieved with RSMA over H-NOMA. As the traffic level increases towards $\pi = 0.5$, the timely throughput performance of RSMA with *BEST* becomes slightly better than H-NOMA with *BEST*. For $Y = 2$ case, normalized timely throughput for RSMA with *BEST* gets the lowest value of 0.96 at the traffic level 0.5. In all traffic levels, the timely throughput performance of *RAND* continuously deteriorates as NOMA/RSMA subset count decreases from $Y = 5$ to $Y = 2$. These results show that RSMA with *BEST* significantly preserves timely throughput compared to *RAND* for small number of users such as $N = 5$.

In Figs 6.10-6.11-6.12, normalized timely throughput obtained with H-NOMA and RSMA using *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 10$ and the traffic level $\pi \in$

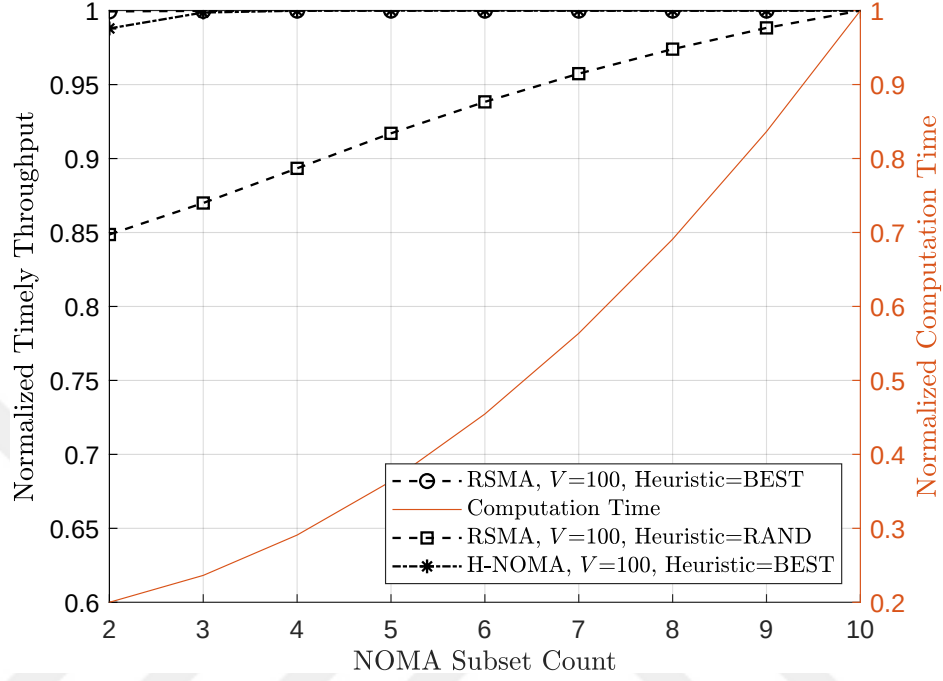


Figure 6.10: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 10$ and $\pi = 0.1$.

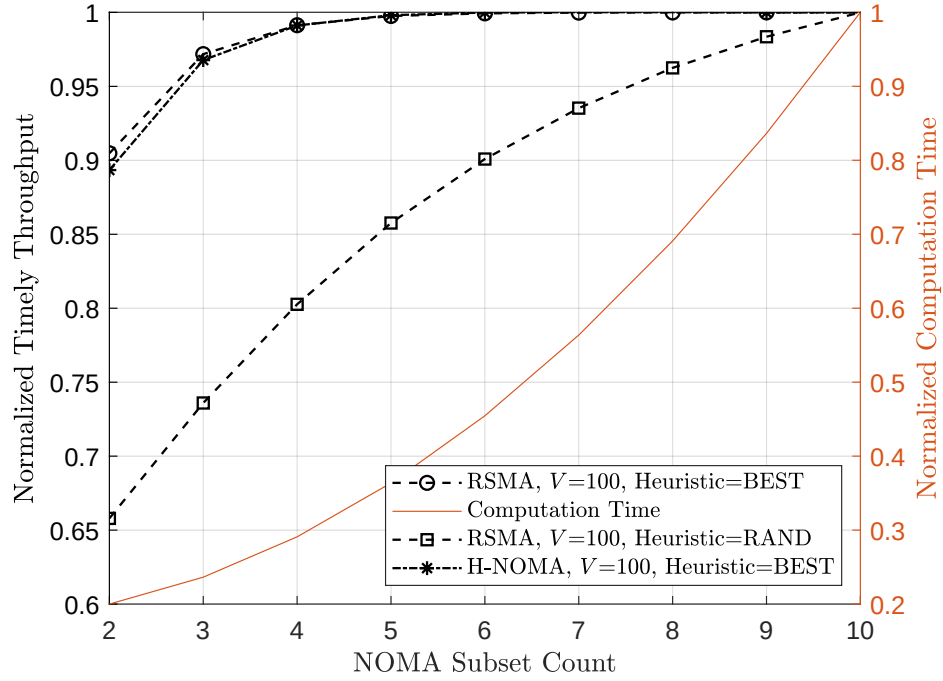


Figure 6.11: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 10$ and $\pi = 0.3$.

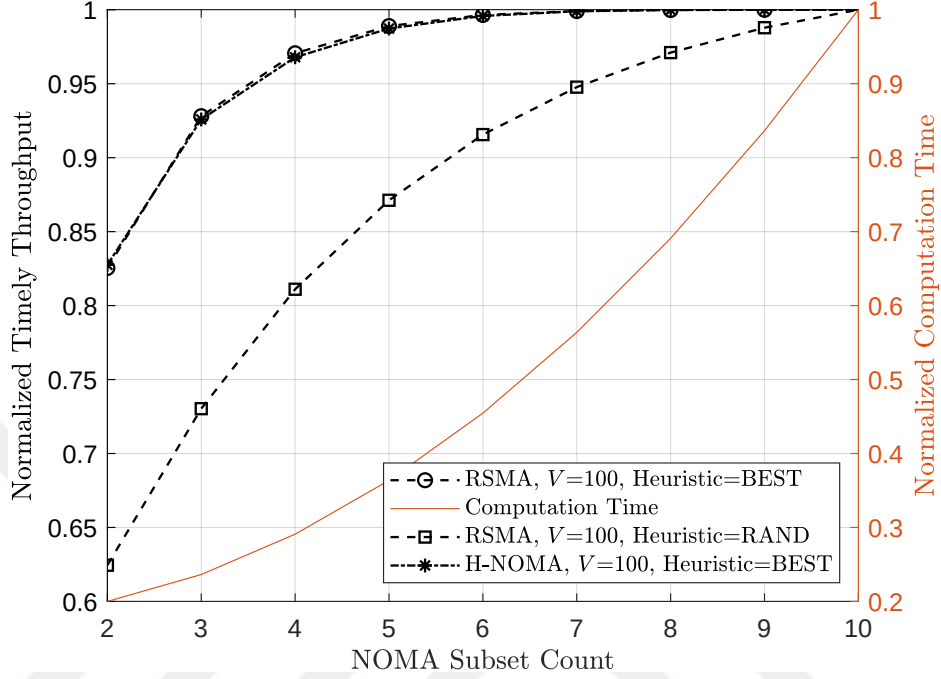


Figure 6.12: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 10$ and $\pi = 0.5$.

$\{0.1, 0.3, 0.5\}$, respectively. Similarly, *BEST* performs better than *RAND* in all simulations. For low traffic level of $\pi = 0.1$, RSMA with *BEST* mostly preserves timely throughput performance at $Y = 2$, which refers to about 80% reduction in computation time. However, H-NOMA with *BEST* achieves the same performance at $Y = 3$. This observation is due to the performance gain achieved with RSMA compared to H-NOMA. As the traffic level increases towards $\pi = 0.3$ and $\pi = 0.5$, both RSMA and H-NOMA with *BEST* can preserve timely throughput at $Y = 4$ and $Y = 6$, respectively. These cases refer to 70% and 55% reduction in computation time, respectively. Note that, *BEST* only reduces computation load. Therefore, as the number of users increases in the system, the performance of *BEST* starts to proportionally alter with respect to the load of the system, both for RSMA and H-NOMA.

In Figs 6.13-6.14-6.15, normalized timely throughput for H-NOMA and RSMA using *BEST* and *RAND*, and normalized computation time as a function of Y are presented under user count $N = 20$ and the traffic level $\pi \in \{0.1, 0.3, 0.5\}$, respectively. Similarly, *BEST* performs better than *RAND* in all cases. For low

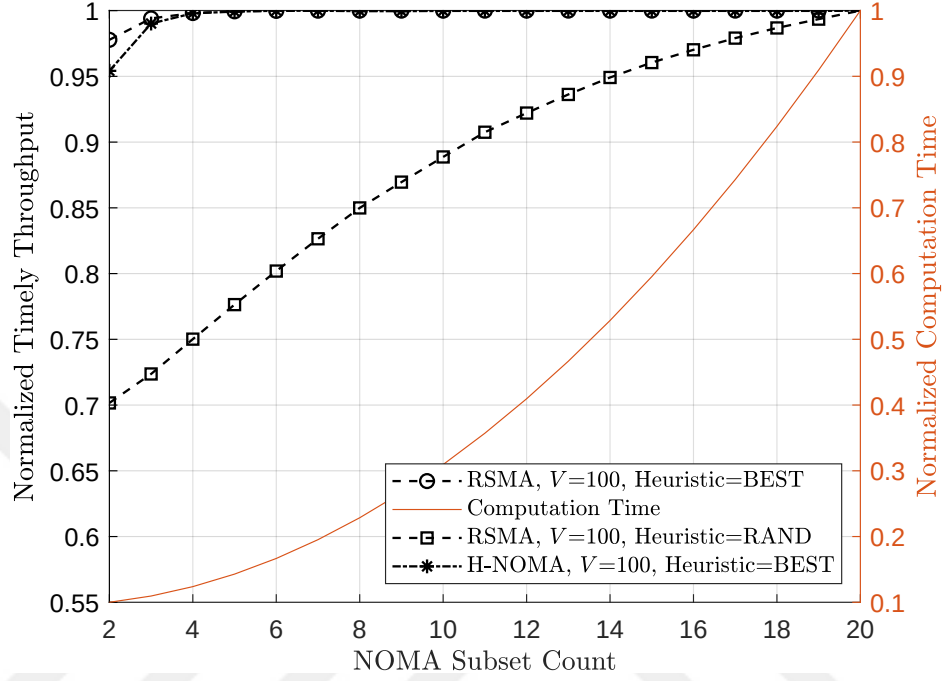


Figure 6.13: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 20$ and $\pi = 0.1$.

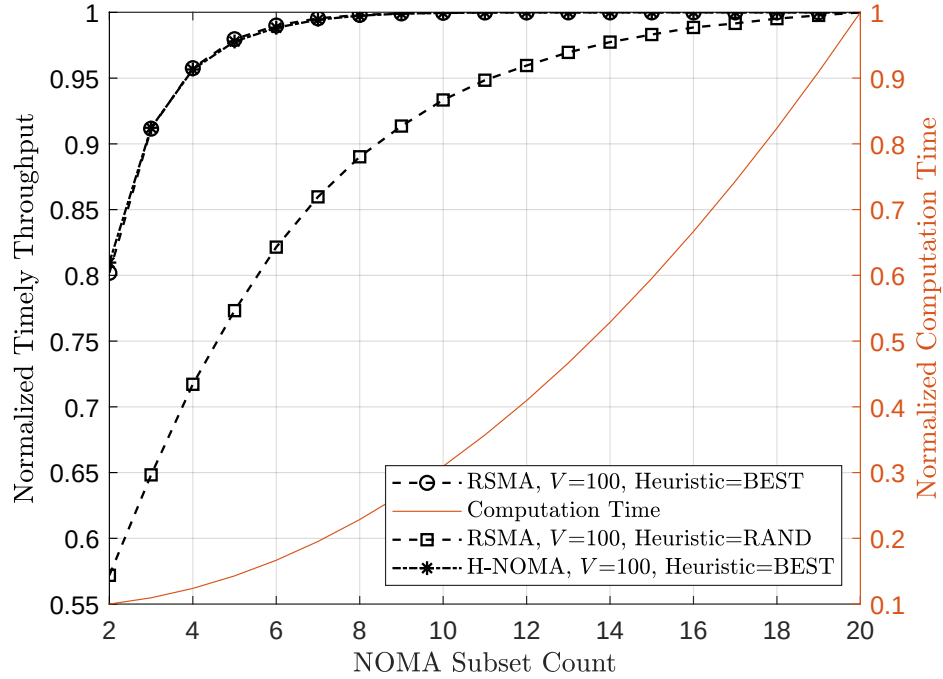


Figure 6.14: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 20$ and $\pi = 0.3$.

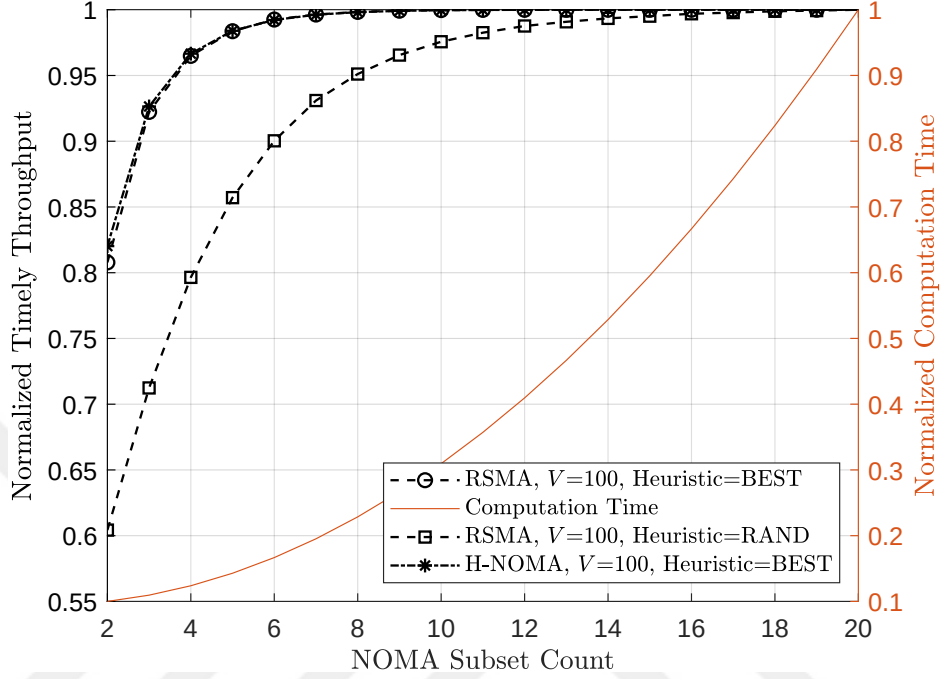


Figure 6.15: Normalized timely throughput and normalized computation time as a function of RSMA Subset Count Y under $N = 20$ and $\pi = 0.5$.

traffic level of $\pi = 0.1$, both H-NOMA and RSMA with *BEST* preserves timely throughput performance at $Y = 4$, which refers to about 90% reduction in computation time. Under this low traffic level $\pi = 0.1$, RSMA with *BEST* performs slightly better than H-NOMA for $Y \in \{2, 3\}$. As the traffic level increases towards $\pi = 0.3$ and $\pi = 0.5$, *BEST* can preserve timely throughput at $Y = 7$ for both cases, and it refers to about 80% reduction in computation time. Therefore, as the number of users further increases in the system, the performance of *BEST* starts to saturate both for RSMA and H-NOMA due to the increasing number of small packets in the system. These results clearly show that *BEST* is a simple and effective heuristic to reduce computational load both for the H-NOMA and RSMA algorithms.

Chapter 7

Conclusions and Future Work

Increasing demand for time-critical applications relying on URLLC highlights research on latency constrained communications with strict deadlines. In this thesis, we addressed the problem of latency constrained communications with strict deadlines under average power constraint using Hybrid-MA, which consists of both OMA and NOMA as transmission scheme options, and RSMA. We aimed to maximize the timely throughput, which represents the average number of successfully transmitted packets before deadline expiration. Since expired packets waiting in the system are dropped, we proposed a flexible packet dropping mechanism called EPD, which detects to-be-expired packets before deadline violation and drops them proactively. We extended the stochastic network optimization framework for packets with deadlines by Fountoulakis et al. to a much more realistic scale in terms of data representation, MA model and wireless channel model. We used Lyapunov stochastic optimization methods to develop a dynamic power assignment algorithm for minimizing packet drop rate while satisfying time average power constraints. The performance of the proposed dynamic algorithm can be adjusted by properly selecting the α parameter for prioritizing the user packets according to the channel states and remaining deadlines. Finally, we proposed a computational load reduction heuristic, which reduces computation time at the cost of a small reduction in the timely throughput performance.

Numerical results indicate that Hybrid MA increases the timely throughput compared to the OMA-only case by up to 46% and on average by more than 21% while satisfying average power constraints. The increase in the timely throughput with Hybrid MA is further improved up to 53% and on average by 24.5% using the proposed flexible packet dropping mechanism EPD. Moreover, timely throughput increases on the average by 3.12% and up to 5.95% with RSMA using EPD compared to Hybrid-MA using EPD. RSMA is robust under varying traffic and power conditions by utilizing the spectrum efficiently. RDPPA introduces a novel degree of freedom for the packet drop rate minimization by prioritizing packets in the system considering both their remaining deadlines as well as channel states. EPD is a simple but effective example for illustrating the potential of flexible packet dropping mechanisms on the timely throughput. Moreover, the EPD mechanism preserves fairness among users while increasing the timely throughput performance. Finally, the proposed computational load reduction heuristic *BEST* is a simple method that performs effectively both for the Hybrid-MA and RSMA algorithms. *BEST* heuristic reduces computational load by 50%–90% depending on the number of users in the system and the traffic level.

As a future work, advanced flexible packet dropping mechanisms to reduce incomplete transmissions can be studied in order to further increase the timely throughput of the proposed MA transmission schemes. Moreover, the optimal power allocation with exact channel dispersion $\mathcal{V}(\eta)$ in the FBL regime is subject to further research. Furthermore, development of packet drop rate minimization heuristics for RSMA with lower computation times while achieving near-optimal results is another significant topic for future work. Another RSMA-oriented future work would be to increase fairness among users while preserving the achieved timely throughput performance. Finally, consideration of MIMO is another important task for future work.

Bibliography

- [1] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, “Wireless access in ultra-reliable low-latency communication (URLLC),” *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.
- [2] N. Larson, D. Baltrunas, A. Kvalbein, A. Dhamdhere, K. Claffy, and A. Elmokashfi, “Investigating excessive delays in mobile broadband networks,” in *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, AllThingsCellular ’15, (New York, NY, USA), p. 51–56, Association for Computing Machinery, 2015.
- [3] 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.
- [4] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis, “Extreme ultra-reliable and low-latency communication,” *Nature Electronics*, vol. 5, no. 3, pp. 133–141, 2022.
- [5] S. Baek, D. Kim, M. Teseanovic, and A. Agiwal, “3GPP new radio release 16: Evolution of 5G for industrial internet of things,” *IEEE Communications Magazine*, vol. 59, no. 1, pp. 41–47, 2021.
- [6] T.-K. Le, U. Salim, and F. Kaltenberger, “An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17,” *IEEE Access*, vol. 9, pp. 433–444, 2021.

- [7] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, “Ultra-reliable low latency cellular networks: Use cases, challenges and approaches,” *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, 2018.
- [8] O. Holland, E. Steinbach, R. V. Prasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. S. Rao, S. Oteafy, M. Eid, M. Luden, A. Bhardwaj, X. Liu, J. Sachs, and J. Araújo, “The IEEE 1918.1 “tactile internet” standards working group and its standards,” *Proceedings of the IEEE*, vol. 107, no. 2, pp. 256–279, 2019.
- [9] S. S. Sefati and S. Halunga, “Ultra-reliability and low-latency communications on the internet of things based on 5G network: Literature review, classification, and future research view,” *Transactions on Emerging Telecommunications Technologies*, vol. 34, no. 6, p. e4770, 2023.
- [10] G. Maciel Ferreira Silva and T. Abrão, “Multipower-level Q-learning algorithm for random access in nonorthogonal multiple access massive machine-type communications systems,” *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 9, p. e4509, 2022.
- [11] L. D. Xu, W. He, and S. Li, “Internet of things in industries: A survey,” *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.
- [12] U. Cisco, “Cisco annual internet report (2018–2023) white paper,” *Cisco: San Jose, CA, USA*, vol. 10, no. 1, pp. 1–35, 2020.
- [13] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, “Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G,” *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1117–1174, 2022.
- [14] I. U. Din, M. Guizani, S. Hassan, B.-S. Kim, M. K. Khan, M. Atiquzzaman, and S. H. Ahmed, “The internet of things: A review of enabled technologies and future challenges,” *IEEE Access*, vol. 7, pp. 7606–7640, 2018.

- [15] M. El Tanab and W. Hamouda, “Efficient resource allocation in fast-uplink grant for machine-type communications with NOMA,” *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 18113–18129, 2022.
- [16] P. Popovski, F. Chiarioti, K. Huang, A. E. Kalør, M. Kountouris, N. Pappas, and B. Soret, “A perspective on time toward wireless 6G,” *Proceedings of the IEEE*, vol. 110, no. 8, pp. 1116–1146, 2022.
- [17] J. Yao, Q. Zhang, and J. Qin, “Joint decoding in downlink NOMA systems with finite blocklength transmissions for ultrareliable low-latency tasks,” *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17705–17713, 2022.
- [18] C. Zhang, X. Sun, J. Zhang, X. Wang, S. Jin, and H. Zhu, “Throughput optimization with delay guarantee for massive random access of M2M communications in industrial IoT,” *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10077–10092, 2019.
- [19] K. T. Phan, P. Huynh, D. N. Nguyen, D. T. Ngo, Y. Hong, and T. Le-Ngoc, “Energy-efficient dual-hop internet of things communications network with delay-outage constraints,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4892–4903, 2020.
- [20] I.-H. Hou and P. R. Kumar, “Packets with deadlines: A framework for real-time wireless networks,” *Synthesis Lectures on Communication Networks*, vol. 6, no. 1, pp. 1–116, 2013.
- [21] I.-H. Hou, V. Borkar, and P. R. Kumar, “A theory of QoS for wireless,” in *IEEE INFOCOM 2009*, pp. 486–494, 2009.
- [22] S. Lashgari and A. S. Avestimehr, “Timely throughput of heterogeneous wireless networks: Fundamental limits and algorithms,” *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8414–8433, 2013.
- [23] 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.

- [24] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.
- [25] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, “Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations,” *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.
- [26] 3GPP, “Study on downlink multiuser superposition transmission (MUST) for LTE,” Jan 2016. Version 13.0.0.
- [27] Y. Yuan, Z. Yuan, G. Yu, C.-h. Hwang, P.-k. Liao, A. Li, and K. Takeda, “Non-orthogonal transmission technology in LTE evolution,” *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, 2016.
- [28] H. Chamkhia, A. Erbad, A. Al-Ali, A. Mohamed, A. Refaey, and M. Guizani, “PLS performance analysis of a hybrid NOMA-OMA based IoT system with mobile sensors,” pp. 1419–1424, IEEE Wireless Communications and Networking Conference (WCNC), 2022.
- [29] U. Ghafoor, M. Ali, H. Z. Khan, A. M. Siddiqui, and M. Naeem, “Efficient resource allocation for hybrid nonorthogonal multiple access based heterogeneous networks beyond fifth-generation,” *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 12, p. e4630, 2022.
- [30] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, “Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis,” *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 349–353, 2020.
- [31] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, “Rate-splitting multiple access: Fundamentals, survey, and future research trends,” *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.
- [32] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, “Rate-splitting multiple access for downlink multi-antenna communications: Physical layer design and link-level simulations,” *IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–6, 2020.

- [33] O. Dizdar, Y. Mao, Y. Xu, P. Zhu, and B. Clerckx, “Rate-splitting multiple access for enhanced URLLC and eMBB in 6G: Invited paper,” in *2021 17th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 1–6, 2021.
- [34] X. Ou, X. Xie, H. Lu, and H. Yang, “Resource allocation in MU-MISO rate-splitting multiple access with sic errors for URLLC services,” *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 229–243, 2023.
- [35] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, “Rate-splitting multiple access with finite blocklength for short-packet and low-latency downlink communications,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 12333–12337, 2022.
- [36] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, “Max-min fairness of rate-splitting multiple access with finite blocklength communications,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 6816–6821, 2023.
- [37] Y. Wang, V. W. S. Wong, and J. Wang, “Flexible rate-splitting multiple access with finite blocklength,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1398–1412, 2023.
- [38] E. Fountoulakis, N. Pappas, Q. Liao, A. Ephremides, and V. Angelakis, “Dynamic power control for packets with deadlines,” pp. 1–6, IEEE Global Communications Conference (GLOBECOM), 2018.
- [39] L. Georgiadis, R. Guérin, and A. Parekh, “Optimal multiplexing on a single link: delay and buffer requirements,” *IEEE Transactions on Information Theory*, vol. 43, no. 5, pp. 1518–1535, 1997.
- [40] 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.
- [41] M. Choi, J. Kim, and J. Moon, “Dynamic power allocation and user scheduling for power-efficient and delay-constrained multiple access networks,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4846–4858, 2019.

- [42] M. J. Neely, “Stochastic network optimization with application to communication and queueing systems,” *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.
- [43] J. Zeng, C. Xiao, Z. Li, W. Ni, and R. P. Liu, “Dynamic power allocation for uplink NOMA with statistical delay QoS guarantee,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 8191–8203, 2021.
- [44] M. Zafer and E. Modiano, “Optimal rate control for delay-constrained data transmission over a wireless channel,” *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 4020–4039, 2008.
- [45] L. Lei, L. You, Q. He, T. X. Vu, S. Chatzinotas, D. Yuan, and B. Ottersten, “Learning-assisted optimization for energy-efficient scheduling in deadline-aware NOMA systems,” *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 615–627, 2019.
- [46] Y. Xu, C. Shen, D. Cai, and G. Zhu, “Latency constrained non-orthogonal packets scheduling with finite blocklength codes,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12312–12316, 2020.
- [47] L. Lyu, C. Chen, N. Cheng, X. Guan, and X. Shen, “NOMA-assisted small-packet transmissions in mission-critical MTCs for industrial automation,” pp. 1–6, IEEE Global Communications Conference (GLOBECOM), 2018.
- [48] Y. Zhu, X. Yuan, Y. Hu, T. Wang, M. Cenk Gursoy, and A. Schmeink, “Low-latency hybrid NOMA-TDMA: QoS-driven design framework,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.
- [49] J. Choi, “Opportunistic noma for uplink short-message delivery with a delay constraint,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3727–3737, 2020.
- [50] E. Fountoulakis, N. Pappas, and A. Ephremides, “Dynamic power control for time-critical networking with heterogeneous traffic,” *ITU Journal on Future and Evolving Technologies*, vol. 2, no. 1, 2021.

- [51] O. B. Gamgam and E. Karasan, “Satisfying strict deadlines for cellular internet of things through hybrid multiple access,” *Transactions on Emerging Telecommunications Technologies*, vol. 34, no. 8, p. e4814, 2023.
- [52] Q. Abbas, S. A. Hassan, H. K. Qureshi, K. Dev, and H. Jung, “A comprehensive survey on age of information in massive IoT networks,” *Computer Communications*, 2022.
- [53] Q. Wang, H. Chen, C. Zhao, Y. Li, P. Popovski, and B. Vucetic, “Optimizing information freshness via multiuser scheduling with adaptive NOMA/OMA,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1766–1778, 2021.
- [54] J. Wang, X. Jia, Z. Chen, and X. Zhang, “Optimization on information freshness for multi-access users with energy harvesting cognitive radio networks,” *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 11, p. e4591, 2022.
- [55] H. Pan, J. Liang, S. C. Liew, V. C. Leung, and J. Li, “Timely information update with nonorthogonal multiple access,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4096–4106, 2020.
- [56] Q. Wang, H. Chen, Y. Li, and B. Vucetic, “Minimizing age of information via hybrid noma/oma,” pp. 1753–1758, IEEE International Symposium on Information Theory (ISIT), 2020.
- [57] 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,” IEEE International Conference on Communications (ICC), 2021.
- [58] C. Wang, R. Zhang, J. Tan, and B. Jiao, “Hybrid NOMA user grouping for short packet communications in IoT network with different types of devices,” pp. 228–234, IEEE International Conference on Communications (ICC) Workshops, 2022.
- [59] A. Maatouk, M. Assaad, and A. Ephremides, “Minimizing the age of information: NOMA or OMA?,” *IEEE Conference on Computer Communications (INFOCOM) Workshops*, pp. 102–108, 2019.

- [60] 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.
- [61] Y. Xu, C. Shen, T.-H. Chang, S.-C. Lin, Y. Zhao, and G. Zhu, “Transmission energy minimization for heterogeneous low-latency NOMA downlink,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1054–1069, 2019.
- [62] R.-J. Reifert, S. Roth, A. A. Ahmad, and A. Sezgin, “Energy efficiency in rate-splitting multiple access with mixed criticality,” *IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 681–686, 2022.
- [63] Y. Karacora and A. Sezgin, “Rate-splitting enabled multi-connectivity in mixed-criticality systems,” 2022.
- [64] F. Xiao, P. Chen, H. Wu, Y. Mao, and H. Liu, “Delay minimization using hybrid RSMA-TDMA for mobile edge computing,” *Electronics*, vol. 12, no. 11, 2023.
- [65] Y. Hua, Y. Fu, and Q. Zhu, “Latency minimization for advanced RSMA-enabled wireless caching networks,” *IEEE Wireless Communications Letters*, vol. 12, no. 8, pp. 1329–1333, 2023.
- [66] S. Zhang, S. Zhang, W. Yuan, Y. Li, and L. Hanzo, “Efficient rate-splitting multiple access for the internet of vehicles: Federated edge learning and latency minimization,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1468–1483, 2023.
- [67] S. Akyürek, “Age aware power allocation for energy-efficient wireless networks using RSMA,” Master’s thesis, Bilkent University, 2023.
- [68] J. Xu, O. Dizdar, and B. Clerckx, “Rate-splitting multiple access for short-packet uplink communications: A finite blocklength analysis,” *IEEE Communications Letters*, vol. 27, no. 2, pp. 517–521, 2023.

- [69] M. Amjad, L. Musavian, and S. Aissa, “NOMA versus OMA in finite block-length regime: Link-layer rate performance,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16253–16257, 2020.
- [70] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, “On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2238–2252, 2017.
- [71] L. Georgiadis, M. J. Neely, L. Tassiulas, *et al.*, “Resource allocation and cross-layer control in wireless networks,” *Foundations and Trends® in Networking*, vol. 1, no. 1, pp. 1–144, 2006.
- [72] C. Li, C. She, N. Yang, and T. Q. Quek, “Secure transmission rate of short packets with queueing delay requirement,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 203–218, 2021.
- [73] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, “Non-orthogonal multiple access: Common myths and critical questions,” *IEEE Wireless Communications*, vol. 26, no. 5, pp. 174–180, 2019.
- [74] E. Erturk, O. Yildiz, S. Shahsavari, and N. Akar, “Power allocation and temporal fair user group scheduling for downlink NOMA,” *Telecommunication Systems: Modelling, Analysis, Design and Management*, vol. 77, pp. 753–766, August 2021.
- [75] F. Liu and M. Petrova, “Performance of proportional fair scheduling for downlink PD-NOMA networks,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 7027–7039, 2018.
- [76] M. Yang, J. Chen, Z. Ding, L. Lv, N. Al-Dhahir, and L. Yang, “Rate-aware user pair scheduling with joint power allocation and decoding order selection in NOMA systems,” *IEEE Transactions on Communications*, vol. 71, no. 9, pp. 5303–5319, 2023.
- [77] S. Fu, F. Fang, L. Zhao, Z. Ding, and X. Jian, “Joint transmission scheduling and power allocation in non-orthogonal multiple access,” *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 8137–8150, 2019.

- [78] C. She and C. Yang, “Ensuring the quality-of-service of tactile internet,” pp. 1–5, IEEE 83rd Vehicular Technology Conference (VTC Spring), 2016.
- [79] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, “A quantitative measure of fairness and discrimination,” *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, 1984.

