ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

ANALYTICAL MODELS AND CROSS-LAYER DELAY OPTIMIZATION FOR RESOURCE ALLOCATION OF NOMA DOWNLINK SYSTEMS

Ph.D. THESIS

Ömer Faruk GEMİCİ

Electronics and Communications Engineering Department

Telecommunications Engineering Programme

SEPTEMBER 2020

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

ANALYTICAL MODELS AND CROSS-LAYER DELAY OPTIMIZATION FOR RESOURCE ALLOCATION OF NOMA DOWNLINK SYSTEMS

Ph.D. THESIS

Ömer Faruk GEMİCİ
(504142309)

Electronics and Communications Engineering Department

Telecommunications Engineering Programme

Thesis Advisor: Prof. Dr. Hakan Ali ÇIRPAN
Co-advisor: Dr. İbrahim HÖKELEK

SEPTEMBER 2020

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

AŞAĞI YÖNLÜ NOMA SİSTEMLERİNDE KAYNAK TAHSİSİ İÇİN ANALİTİK MODELLER VE KATMANLAR ARASI ETKİLEŞİMLİ GECİKME OPTİMİZASYONU

DOKTORA TEZİ

Ömer Faruk GEMİCİ
(504142309)

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Telekomünikasyon Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Hakan Ali ÇIRPAN
Eş Danışman: Dr. İbrahim HÖKELEK

EYLÜL 2020

Ömer Faruk GEMİCİ, a Ph.D. student of ITU Graduate School of Science Engineering and Technology 504142309 successfully defended the thesis entitled "ANALYTICAL MODELS AND CROSS-LAYER DELAY OPTIMIZATION FOR RESOURCE ALLOCATION OF NOMA DOWNLINK SYSTEMS", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Hakan Ali ÇIRPAN** Istanbul Technical University	.....
Co-advisor :	**Dr. İbrahim HÖKELEK** TÜBİTAK BİLGEM Istanbul Technical University	.....
Jury Members :	**Prof. Dr. Güneş Zeynep KARABULUT KURT** Istanbul Technical University	.....
	Prof Dr. Özgür ERÇETİN Sabancı University	.....
	Dr. Fatih KARA PAVOTEK	.....
	Prof. Dr. Ender Mete EKŞİOĞLU Istanbul Technical University	.....
	Dr. Selçuk CEVHER Karadeniz Technical University	.....

Date of Submission : **10 June 2020**
Date of Defense : **01 September 2020**

To my family,

FOREWORD

I would like to express my gratitude and appreciation to my advisor, Prof. Dr. Hakan Ali ÇIRPAN, for his continuous guidance, fruitful feedback, great support, and encouragement. I am immensely grateful to my esteemed co-advisor, Dr. İbrahim HÖKELEK who has always been more than a supervisor to me, a brother, a dependable friend, a life coach, someone who makes a difference in my life that probably will stay for the rest of it.

I would like to give my sincere appreciation to my committee members, Prof. Dr. Güneş Zeynep KARABULUT KURT, Prof. Dr. Özgür ERÇETİN, and Dr. Fatih KARA for their supportive and insightful comments to include new perspectives to the subject, which increase the quality of this thesis.

I would also like to thank my colleagues at TÜBİTAK BİLGEM, especially Dr. Serdar Özgür ATA and Muhammet Selim DEMİR for their friendship and colourful contributions on the thesis.

The greatest part of my gratitude belongs to my family definitely. I would like to thank my father, İzzet, whose personality inspired and motivated me, and my lovely mother, Fatma, for her endless love, support and belief all through my life. I would also like to thank my sisters, Aslıhan and Rabia, for their sincere encouragement and unconditional support. Last but not least, my deepest appreciation goes to my beloved wife, Ayşe Kübra, who was always on my side, supporting me, encouraging me, understanding me, giving me the strength to move forward at every moment of this journey. My little daughter Miray and son Ali, you have brightened our lives with your coming.

September 2020 Ömer Faruk GEMİCİ

TABLE OF CONTENTS

Page

ABBREVIATIONS

1G	The first generation
2G	The second generation
3G	The third generation
3GPP	3rd Generation Partnership Project
4G	The fourth generation
5G	The fifth generation
5G NR	5G new radio
AWGN	Additive white Gaussian noise
BS	Base station
CDF	Cumulative distribution function
CDMA	Code division multiple access
CSI	Channel state information
eMBB	Enhanced mobile broadband
FDMA	Frequency division multiple access
FIFO	First-in-first-out
GA	Genetic Algorithm
IGAM	Iterative gradient ascend method
LTE	Long-term evolution
MAQD	Maximum of the average queuing delays
MIMO	Multiple-input multiple-output
mMTC	Massive machine type communications
mmWave	Millimeter wave
NOMA	Non-orthogonal multiple access
OFDM	Orthogonal frequency division multiplexing
OFDMA	Orthogonal frequency division multiple access
OMA	Orthogonal multiple access
PDF	Probability density function
PD-NOMA	Power Domain non-orthogonal multiple access
PF	Proportional fairness
PUSF	Proportional user satisfaction fairness
PUSF-GA	Proportional user satisfaction fairness with genetic algorithm
SIC	Successive interference cancellation
SINR	Signal-to-interference-plus-noise ratio
SISO	Single-input single-output
SNR	Signal-to-noise ratio
TDMA	Time division multiple access
UDB-PF	User demand based proportional fairness
UDB-PF-GA	User demand based proportional fairness with genetic algorithm
UE	User equipment
URLLC	Ultra reliable and low latency communications

SYMBOLS

β	:	Path loss exponent
$C_k(X)$	:	Instantaneous channel capacity of the user k with the channel gain power of X
d_k	:	Distance between base station and user equipment k
γ_k	:	Outage condition of the user k
h_k	:	Channel coefficient of the user k
h_{km}	:	Channel coefficient of the user k at the resource block m
I	:	User group assignment of all resource blocks
I_m	:	User group assignment at the resource block m
K	:	Number of connected users in a cell
Λ_k	:	Arrival rate of the user k
$L_{k,m}$	:	Size of the the m^{th} packet to be served for the user k
μ_k	:	Service rate of the user k
M	:	Number of resource blocks
N_{max}	:	Maximum number of users sharing the same resource block
Ω_k	:	Traffic demand requirement of the user k
$\Omega_k[n]$	:	Traffic demand requirement of the user k at the time slot n
p_k	:	Power allocation coefficient of the user k
p_{km}	:	Power allocation coefficient of the user k at the resource block m
$PL(d_k)$	:	Path loss of the user k
P_t	:	Transmit power of the base station
$P(\gamma)$	:	Probability of satisfying the condition γ
P_{out}	:	Outage probability of the system
$\overline{P}$	:	Power allocation coefficients of all resource blocks
$\overline{P}_m$	:	Power allocation coefficients at the resource block m
Q_k	:	Queuing delay of user k
R_k	:	Throughput of the user k
R_{km}	:	Throughput of the user k at the resource block m
$R_k[n]$	:	Amount of served bits for the user k at the time slot n
$R_k^c[n]$	:	Average served bits for the user k over a time window at the time slot n
$R_k(X)$	:	Instantaneous amount of served bits of the user k with the channel gain power of X
$S_{k,m}$	:	Service time of the m^{th} packet to be served for the user k
t_c	:	Length of the averaging time window for PF-based schedulers
T_s	:	Time slot duration
$\Upsilon_k[n]$	:	Satisfaction level of the user k at the time slot n
$\Upsilon_k^c[n]$	:	Average satisfaction of the user k over a time window at the time slot n
$W_{0,k}$	:	Noise power of the user k

LIST OF TABLES

Page

LIST OF FIGURES

Page

ANALYTICAL MODELS AND CROSS-LAYER DELAY OPTIMIZATION FOR RESOURCE ALLOCATION OF NOMA DOWNLINK SYSTEMS

SUMMARY

5G is introduced by 3rd Generation Partnership Project (3GPP) to satisfy the stringent delay and reliability requirements of 5G services such as industrial automation, augmented and virtual reality, and intelligent transportation. Non-orthogonal multiple access (NOMA) is one of the promising technologies for low latency services of 5G, where the system capacity can be increased by allowing simultaneous transmission of multiple users at the same radio resource. The resource allocation in NOMA systems including user scheduling and power allocation determine the mapping of users to radio resource blocks and the transmission power levels of users at each resource block, respectively.

In this thesis, we first propose a genetic algorithm (GA) based multi-user radio resource allocation scheme for NOMA downlink systems. In our set-up, GA is used to determine the user groups to simultaneously transmit their signals at the same time and frequency resource while the optimal transmission power level is assigned to each user to maximize the geometric mean of user throughputs. The simulation results show that the GA based approach is a powerful heuristic to quickly converge to the target solution which balances the trade-off between total system throughput and fairness among users.

The most of the resource allocation studies for NOMA systems including our GA based approach assumes full buffer traffic model where the incoming traffic of each user is infinite while the traffic in real life scenarios is generally non-full buffer. As the second contribution, we propose User Demand Based Proportional Fairness (UDB-PF) and Proportional User Satisfaction Fairness (PUSF) algorithms for resource allocation in NOMA downlink systems when traffic demands of the users are rate limited and time-varying. UDB-PF extends the PF based scheduling by allocating optimum power levels towards satisfying the traffic demand constraints of user pair in each resource block. The objective of PUSF is to maximize the network-wide user satisfaction by allocating sufficient frequency and power resources according to traffic demands of the users. In both cases, user groups are selected first to simultaneously transmit their signals at the same frequency resource while the optimal transmission power level is assigned to each user to optimize the underlying objective function. In addition, the GA is employed for user group selection to reduce the computational complexity. When the user traffic rate requirements change rapidly over time, UDB-PF yields better sum-rate (throughput) while PUSF provides better network-wide user satisfaction results compared to the PF based user scheduling. We also observed that the GA based user group selection significantly reduced the computational load while achieving the comparable results of the exhaustive search.

The low latency objectives of URLLC services such as industrial control and automation, augmented and virtual reality, tactile Internet and intelligent transportation

requires delay analysis which cannot be possible using the rate limited traffic demands. The packet based traffic model with random inter-arrival times and packet sizes have to be utilized. New analytical models using packet based traffic model with random inter-arrival times and packet sizes are of paramount importance to develop high performance resource allocation strategies satisfying the challenging latency requirements of 5G services. As the third contribution, we propose an analytical model to characterize the average queuing delay for NOMA downlink systems by utilizing a discrete time M/G/1 queuing model under a Rayleigh fading channel. The packet arrival process is assumed to be Poisson distributed while the departure process depends on network settings and resource allocation. The average queuing delay results of the analytical model are validated through Monte Carlo simulation experiments. One of the main results is that the ergodic capacity region of NOMA is a superset of OMA indicating that the NOMA can support higher service rate and lower latency using the same resources such as transmission power and bandwidth. Furthermore, the proposed analytical model is applied for the performance evaluation of the 5G NR concept when the NOMA is utilized. The model accurately predicts that the average queuing delay decreases when wider bandwidth and shorter time slot duration are employed in 5G NR.

The outage probability becomes an important metric that should be minimized to address the reliability aspect of the URLLC services. We utilize the common outage condition such that the user fails either decoding its own signal or performing SIC for the signals of other users at the receiver when the SINR is lower than a predefined outage threshold. As the fourth contribution, the optimum power allocation for a single resource block that minimizes the system outage probability under Rayleigh fading channel, where a common signal to interference plus noise ratio (SINR) level is utilized as an outage condition, is provided as a closed form expression. The accuracy of the proposed optimum power allocation model is validated by the Monte Carlo simulations. The numerical results show that the outage probability of OMA with the fractional power allocation is lower than NOMA with the optimum power allocation. The results indicate that the trade-off between the outage and spectral efficiency in NOMA should be carefully controlled to meet higher throughput and lower latency objectives of 5G.

The last contribution considers the reliability and latency aspects jointly such that the discrete time M/G/1 queuing model of a NOMA downlink system is extended by taking the outage condition into account. The departure process of the queuing model is characterized by obtaining the first and second moment statistics of the service time that depends on the resource allocation strategy and the packet size distribution. The proposed model is utilized to obtain the optimum power allocation that minimizes the maximum of the average queuing delay (MAQD) for a two-user network scenario. The Monte Carlo simulation experiments are performed to numerically validate the model by providing MAQD results for both NOMA and orthogonal multiple access (OMA) schemes. The results demonstrate that the NOMA achieves lower latency for low SINR outage thresholds while its performance is degraded faster than OMA as the SINR outage threshold increases such that OMA outperforms NOMA beyond a certain threshold. Another important result is that the latency performance of NOMA is significantly degraded when the 5G NR frame types having wider bandwidth are utilized. The results provide powerful insights for 5G ultra-reliable low-latency communication (URLLC) services.

AŞAĞI YÖNLÜ NOMA SİSTEMLERİNDE KAYNAK TAHSİSİ İÇİN ANALİTİK MODELLER VE KATMANLAR ARASI ETKİLEŞİMLİ GECİKME OPTİMİZASYONU

ÖZET

3GPP tarafından tanımlanan 5G standartlarında endüstriyel kontrol ve otomasyon, artırılmış ve sanal gerçeklik ve akıllı ulaşım gibi düşük gecikme ve yüksek güvenilirlik gerektiren servisler öne çıkmaktadır. Dikgen olmayan çoklu erişim (NOMA) teknolojisi ile aynı radyo kaynağında birden fazla kullanıcının aynı anda iletilmesine olanak sağlayarak sistem kapasitesi artırılabilir. Böylece NOMA, 5G'nin düşük gecikmeli servislerini destekleyebilecek önemli teknolojilerden birisi olarak değerlendirilmektedir. Aynı radyo kaynağına atanacak kullanıcı gruplarının belirlenmesi ve her bir kullanıcı grubu içerisindeki güç tahsis seviyelerinin belirlenmesi ile tanımlanan NOMA sistemlerindeki radyo kaynak yönetimi ile, kullanıcıların veri çıktısı ve gecikme seviyeleri belirlenebilmektedir.

Bu tezde, ilk olarak aşağı yönlü NOMA sistemlerinde kaynak yönetimi için genetik algoritma (GA) tabanlı çok kullanıcılı radyo kaynağı tahsis şeması önerilmiştir. Önerilen yöntemde genetik algoritma, aynı zaman ve frekans kaynağını paylaşmak üzere seçilen kullanıcı gruplarının belirlenmesinde kullanılırken, her bir kullanıcı grubu içerisinde kullanıcı veri çıktılarının geometrik ortalamasını en üst düzeye çıkaran en uygun iletim gücü seviyesi atanmaktadır. Simülasyon sonuçları, GA tabanlı yaklaşımın, toplam veri çıktısı ile kullanıcılara tahsis edilen veri çıktıları arasındaki adaleti birlikte değerlendirerek hedef çözüme verimli bir şekilde ulaşmada kullanılanılabilecek güçlü bir sezgisel yöntem olduğunu göstermektedir.

Önerdiğimiz GA tabanlı yaklaşım da dahil olmak üzere, NOMA sistemleri için kaynak tahsisi için literatürde önerilen çalışmalarının çoğu, baz istasyonunda kullanıcılara iletilmek üzere sonsuz trafik olduğunu varsaymaktadır. Pratik uygulama alanlarında kullanılmak üzere radyo tahsis şemaları önerebilmek için kullanıcıların trafiğinin sonlu olduğu durumlar göz önüne alınmalıdır. Tez çalışmasındaki ikinci katkı olarak, kullanıcıların trafik talepleri sınırlı ve zaman içinde değiştiği durumda, aşağı yönlü NOMA sistemlerinde kaynak tahsisi için iki yeni kaynak tahsis algoritması önerilmiştir. Bunlar, kullanıcı talebine dayalı oransal adalet (UDB-PF) ve orantılı kullanıcı memnuniyeti adaleti (PUSF) algoritmaları olarak isimlendirilmiştir. UDB-PF, literatürde önerilmiş olan oransal adalet (PF) tabanlı tahsis algoritmasını her kaynak bloğundaki kullanıcı çiftinin trafik talebi kısıtlamalarını göz önüne alarak optimum güç seviyeleri tahsis edilmesi şeklinde tanımlanmaktadır. PUSF yönteminde ise, kullanıcıların trafik taleplerine göre radyo kaynakları tahsis edilerek, kullanıcıya atanan kapasitenin kullanıcı talebine oranı olarak tanımlanan kullanıcı memnuniyet parametresi, orantısal adaletli bir şekilde en üst düzeye çıkarılmaktadır. Her iki yöntem de, sinyallerini aynı radyo kaynağında iletecek kullanıcı gruplarını belirleme ve kullanıcı grubu içerisindeki güç tahsis oranlarını birlikte değerlendirerek en uygun atamayı gerçekleştirir. Ayrıca, işlem yoğunluğunu azaltmak amacıyla kullanıcı grubu seçimi için genetik algoritma (GA) yaklaşımı önerilmiştir. Simülasyon sonuçları

göstermektedir ki, kullanıcı trafik gereksinimleri zaman içinde hızla değiştiğinde, UDB-PF daha yüksek veri çıktısı oluştururken, PUSF, ağ genelinde en iyi kullanıcı memnuniyeti sonucu sağlamaktadır. Önerilen GA tabanlı kullanıcı grubu seçiminin, kapsamlı arama ile gerçekleştirilen grup seçimine göre benzer performans sonuçlarına ulaşırken hesaplama yükünü önemli ölçüde azalttığı gözlenmiştir.

Kullanıcı veri gereksinimlerinin sınırlı olduğu trafik modeli kullanılması, URLLC servislerinin gecikme dinamiklerinin araştırılması için yeterli olmadığından, paket tabanlı, rastgele varış süreleri ve farklı paket uzunluklarının göz önüne alındığı trafik modeli kullanılarak NOMA sistemleri incelenmelidir. 5G hücresel haberleşme sistemlerinde gecikme dinamiklerini karakterize edebilen yeni analitik modeller, 5G hizmetlerinin zorlu gereksinimlerini karşılayan yüksek performanslı kaynak tahsis stratejileri geliştirmek için büyük önem taşımaktadır. Bu nedenle, tez kapsamında üçüncü olarak, Rayleigh solma kanalı altında ayrık zaman ayrık durum M/G/1 kuyruklama modeli kullanarak aşağı yönlü NOMA sistemleri için ortalama kuyruk gecikmesini belirleyen analitik model önerilmiştir. Bu modelde, paket varış süreci Poisson dağılımlı varsayılırken, ayrılma süreci ağ koşullarına ve kullanılan kaynak tahsis yöntemine bağlı olarak belirlenmiştir. Analitik model ile elde edilen ortalama kuyruk gecikme sonuçları Monte Carlo simülasyon deneyleri örtüşmektedir. Sonuçlar, NOMA'nın dikgen çoklu erişimden (OMA) daha yüksek kapasite bölgesine sahip olması ile, daha yüksek servis hızını ve daha düşük gecikmeyi destekleyebileceğini göstermektedir. Ek olarak, önerilen analitik model, 5G yeni radyo (NR) parametreleri altında NOMA kullanıldığı durum için performans değerlendirmeleri sunulmuştur. Buradaki sonuçlara göre önerilen analitik model, 5G NR ile daha geniş bant genişliği ve daha kısa zaman aralığındaki çerçeve yapıları için, hem OMA hem de NOMA için ortalama kuyruk gecikmesinin azaldığını doğru bir şekilde tahmin etmektedir.

Hücresel şebekelerde güvenilirliğe duyarlı uygulamaların ve hizmetlerin çoğalmasıyla, gelişmiş kablosuz haberleşme sistemlerinde yüksek güvenilirlik gereksinimini desteklemek için kesinti olasılığı, en aza indirilmesi gereken önemli bir ölçüt haline gelmektedir. Tez kapsamında dördüncü olarak, ortak sinyal girişim artı gürültü oranı (SINR) seviyesi göz önüne alınarak Rayleigh solma kanalı altında aşağı yönlü NOMA sistemlerinde kesinti olasılığı analiz edilmektedir. Ayrıca, tek bir kaynak bloğu için sistem kesintisi olasılığını en aza indiren optimum güç tahsisi kapalı form şeklinde sunulmuştur. Önerilen analitik model Monte Carlo simülasyonları ile doğrulanırken, optimum güç dağılım yönteminin performans analizleri gerçekleştirilmiştir. Sayısal sonuçlarda kesinti olasılığının güç tahsisine göre değişimi raporlanırken, önerilen yöntemin diğer güç ataması algoritmalarına göre en iyi sonucu verdiği gösterilmektedir. Bunun yanında, NOMA için optimum güç tahsis algoritması kullanıldığında bile OMA'ya göre daha kötü performans verdiği gözlenmiştir. Sonuçlar, NOMA'daki kesinti ve spektral verimlilik arasındaki dengenin, 5G sistemlerde daha yüksek verim ve düşük gecikme hedeflerini karşılamak için dikkatle kontrol edilmesi gerektiğini göstermektedir.

Tez kapsamında son olarak, aşağı yönlü NOMA sistemlerinde, kullanıcının kendi sinyalini çözme veya ardışık girişim giderici (SIC) gerçekleştirebilmesi için gerekli olan SINR kesinti seviyesini göz önüne alan, genişletilmiş ayrık zaman ayrık durum M/G/1 kuyruk modeli önerilmiştir. Önerilen genişletilmiş kuyruk modelinin ayrılma süreci, kaynak tahsisi stratejisine ve paket büyüklüğü dağılımına bağlı olarak servis süresinin birinci ve ikinci moment istatistiklerinin elde edilmesiyle karakterize

edilmektedir. Ayrıca, iki kullanıcılı bir ağ senaryosu için önerilen model kullanılarak en büyük ortalama kuyruk gecikmesinin (MAQD) tek bir noktada en az olduğu ispatlanmıştır. Bununla birlikte, altın bölüm arama (golden section search) yöntemi kullanılarak en düşük MAQD değerini sağlayan optimum güç tahsisi elde edilmiştir. Monte Carlo simülasyon deneyleri ile hem NOMA hem de OMA için MAQD sonuçları elde edilerek önerilen genişletilmiş analitik yöntemin doğrulaması gerçekleştirilmiştir. Sonuçlar göstermektedir ki, düşük SINR kesinti eşikleri için NOMA ile daha düşük gecikme değerleri elde edilirken, SINR kesinti eşiği arttıkça gecikme süresindeki artış OMA ile kıyaslandığında fazla olmaktadır. Bunun sonucu olarak, belirli bir SINR eşik seviyesi üzerinde OMA ile elde edilen gecikme, NOMA kullanıldığı duruma göre daha düşük olmaktadır. Daha geniş bant genişliğine sahip 5G NR çerçeve tipleri kullanıldığında NOMA ile elde edilen gecikmenin önemli ölçüde artması bir başka önemli sonuç olarak raporlanmıştır. Tez kapsamında geliştirilen analitik modeller, 5G ultra güvenilir düşük gecikmeli iletişim (URLLC) hizmetleri için zorlu gecikme ve güvenilirlik ihtiyaçlarını karşılayan radyo kaynağı yönetimi için önemli bilgiler sağlamaktadır.

1. INTRODUCTION

The complexity of 5G network is expected to be significantly higher due to its inherent support for billions of Internet of Things (IoTs) devices enabling new services with stringent delay and reliability requirements. Three broad categories of 5G services considered by 3GPP are enhanced mobile broadband (eMBB), ultra reliable low latency communication (URLLC), and massive machine-type communications (mMTC). While the eMBB and mMTC services focus on the capacity and scalability aspects of 5G, respectively, URLLC is critical for enabling remote control of time and mission-critical services. Non-orthogonal multiple access (NOMA) is a promising technology for 5G systems due to its higher spectral efficiency potentially yielding lower latency and higher scalability results by allowing simultaneous transmission of multiple users at the same resource block. New analytical models which can characterize the latency dynamics of 5G are of paramount importance to develop high performance resource allocation strategies satisfying the challenging requirements of 5G services.

In NOMA systems, simultaneous transmission of multiple users at the same radio resource is allowed since signals of multiple users can be overlapped at the transmitter by assigning appropriate power levels. For example, assuming the total power of two users is fixed, a user with lower channel quality is assigned a higher power. The combined received signal of multiple users is separated at the receiver using successive interference cancellation (SIC). The performance of a NOMA downlink system with an SIC based receiver is reported to be around 30% higher than orthogonal multiple access (OMA) using power domain multiplexing [1–3].

User scheduling and power allocation in NOMA systems determine the mapping of users to radio resource blocks and the transmission power levels of users at each resource block, respectively. The procedures and algorithms used in the decision making process directly affect the performance of NOMA in terms of its spectral efficiency and computational power requirements. The objective of maximizing the

geometric mean of user throughputs in cell provide the optimal trade-off between total system throughput and fairness among users as stated in [4]. To achieve this objective, in this thesis, we first propose genetic algorithm (GA) based multi carrier NOMA downlink scheme which considers user grouping, user group to resource block matching and power allocation at each resource block. The simulation results demonstrate that the results our GA based approach can achieve the same performance results of exhaustive search method. Considering the computational load of exhaustive search, GA is an efficient way to converge to the best solution.

The proportional-fairness (PF) based approaches [5–7] have been widely used for resource and power allocation in NOMA systems. The objective of the PF based scheduler is to assign radio resources to users in such a way that the PF metric, which is the product of average user throughputs over a time window, is maximized. This objective provides a good compromise between the sum-rate of all users (i.e., network-wide throughput) and the fairness among users. In [5], the user pairing corresponding to the highest PF metric is selected among all possible user pairing combinations. Since this approach requires prohibitively expensive computational power, [6] and [7] propose simplified PF based algorithms which require significantly lower computational power while yielding comparable results to the optimum solution. Another approach [8, 9] to resource allocation for NOMA systems aims to maximize the sum-rate of all users at each time epoch after satisfying certain constraints such as the minimum power allocation and throughput of each user. All of the above PF based resource allocation studies for NOMA systems assumes full buffer traffic model which does not correspond to real life traffic scenario. The traffic model in a real network setting is generally non-full buffer where the traffic demand for each user is limited to a certain application rate.

In this thesis, we secondly propose two user scheduling and power allocation methods employing PF based objective functions for NOMA downlink systems under non-full buffer traffic models. Although the existing PF based user scheduling in NOMA systems has been demonstrated to significantly improve the system capacity when the user traffic model is full buffer, it does not perform well when user traffic rates are limited and time-varying. In User Demand Based Proportional Fairness (UDB-PF) algorithm, the PF based scheduling is extended to take time varying user

traffic demands into account in addition to allocating optimum power levels towards satisfying the traffic demand constraints of user pair in each resource block. The main contribution in UDB-PF is to provide the optimum power allocation under user rate constraints. In other words, when the optimum power level of a user provides higher rate than its rate constraint, the excessive power is reallocated to the other user(s) in the same NOMA group. The objective of Proportional User Satisfaction Fairness (PUSF) algorithm is to maximize the network-wide user satisfaction which is the product of average satisfaction values of all users for a given time window. Note that the highest network-wide user satisfaction is achieved when the resources are sufficient to carry traffic demands of all users. In the PUSF approach, the user satisfaction objective for the user grouping and power allocation optimization is defined by us for the first time. However, the maximization of the product of average user satisfaction is similar to PF based methods. As in the UDB-PF approach, the PUSF can also reallocate the excessive power to the other users in the same NOMA group. In both UDB-PF and PUSF algorithms, user groups are selected first to simultaneously transmit their signals at the same frequency resource while the optimal transmission power level is assigned to each user to optimize the underlying objective function. These proposed algorithms evaluate all user group possibilities to select the best user group allocation at each resource block. However, the computational complexity becomes an important issue when the number of users gets higher, especially to meet the real time requirements of the scheduling decisions. We also present a Genetic Algorithm (GA) heuristic to find the user group at each resource block with a relatively low computational load. The UDB-PF and PUSF algorithms with the GA extensions are named as UDB-PF-GA and PUSF-GA, respectively.

One of the main objective of 5G is to satisfy the lower latency requirements of URLLC services such as industrial control and automation, augmented and virtual reality, tactile Internet and intelligent transportation [10, 11]. The latency contribution of the user plane end-to-end (E2E) delay of a packet transmission in 5G can be divided into three main parts: radio access, mobile core, and cloud. The radio access latency between a base station and user equipment includes over-the-air transmission and propagation, queuing, processing, and re-transmission delays [12]. A cross-layer resource allocation approach considering not only wireless channel characteristics

in the physical layer but also traffic arrival and queue occupancy information at the link layer should be employed to achieve the challenging latency objectives of 5G [13]. An opportunistic NOMA downlink approach is presented in [14] such that they propose two queues with different priority levels at the base station for all users. The performance limitations of NOMA in short packet communication for URLLC services is studied by analytically deriving the effective-bandwidth in [15]. The performance of NOMA in short-packet communications is studied in [16] and the optimal power allocation scheme is presented to provide fairness among users' throughput while satisfying QoS requirements of URLLC. The effective capacity of NOMA under statistical delay guarantees has been studied in [17, 18]. In another study [19], a cross-layer approach using integer linear programming is proposed to minimize the average delay for NOMA applications of delay sensitive communication.

In our second contribution, the rate limited traffic demands are considered instead of packet based traffic model with random inter-arrival times and packet sizes so that the delay dynamics can not be studied. New analytical models which can characterize the latency dynamics of 5G are of paramount importance to develop high performance resource allocation strategies satisfying the challenging requirements of 5G services. As the third contribution of the thesis, we propose an analytical model to characterize the average queuing delay for NOMA downlink systems by utilizing a discrete time M/G/1 queuing model under a Rayleigh fading channel. The packet arrival process is assumed to be Poisson distributed while the departure process depends on network settings (e.g., transmit power, bandwidth, and channel model) and resource allocation (e.g., power allocation). We provide an approximation for the service time statistics under a certain packet size distribution by utilizing the random sums of independent and identically distributed (i.i.d.) random variables. Pollaczek Khintchine formula and Little's Law are applied to obtain the queuing dynamics such as the average queuing delay. Extensive simulations are carried out to validate the accuracy of the proposed analytical model for both NOMA and OMA under different network settings including bandwidth, traffic arrival rate, and packet size distribution. The results show that the ergodic capacity region of NOMA is a superset of OMA and the NOMA supports higher arrival rates. The numerical results of the analytical model are close to the results of the simulation experiments indicating that the proposed analytical

model provides a tight approximation for the average queuing delay. Furthermore, the proposed analytical model is applied to evaluate the performance improvements of the 5G NR concept when the NOMA is utilized with the 5G NR frame types.

The outage probability analysis has been taken considerable attention to study the reliability of wireless cellular networks. The outage event can be defined for cellular systems using various performance metrics such as maximum delay, minimum throughput, minimum BER, and minimum SINR levels. As the forth contribution, we investigate the outage probability of NOMA downlink systems under the Rayleigh fading channel model by taking the SINR outage constraint into account to successfully perform both decoding and SIC processes at the receiver. The system outage probability is provided as a closed form expression. Then, we derive the the optimum power allocation that minimizes the system outage probability for two-user scenario. The accuracy of the theoretical derivations are validated with the Monte Carlo simulations. In addition, the proposed power allocation method is compared with fixed NOMA and fractional NOMA and OMA power allocation methods [20]. The results demonstrate that the proposed optimum power allocation method yields the minimum outage probability among all the power allocation schemes of NOMA. However, the outage probability of OMA with the fractional power allocation is lower than NOMA with the optimum power allocation. Note that the spectral efficiency of NOMA is higher since the bandwidth can be utilized by multiple users. These results indicate that the trade-off between the outage and spectral efficiency in NOMA should be carefully controlled to meet higher throughput and lower latency objectives of 5G.

As the fifth contribution, we extend the proposed analytical model by taking the outage event into account such that the user fails either decoding its own signal or performing SIC for the signals of other users at the receiver when the SINR is lower than a predefined outage threshold. The first and second moment statistics of users' service rate under the SINR outage constraint are derived for a NOMA downlink system simultaneously serving K users sharing a single resource block. We propose the optimum power allocation framework by utilizing the extended analytical model such that the maximum of average user queuing delays (MAQD) are minimized for a single resource block simultaneously serving two users.

The extended analytical model including the approximation of the second moment of the service time is validated by performing the Monte Carlo simulation results. In the first set of experiments, the ergodic capacity region of NOMA and OMA is reported for all possible power allocations. The results show that the ergodic capacity region of NOMA is a superset of OMA for lower SINR outage thresholds while OMA yields higher ergodic capacity when the SINR outage thresholds is high. For the second set of experiments, the delay performance of NOMA and OMA is reported using the proposed delay optimization method under various network settings such as SINR outage threshold, user arrival rates and distances. The results show that when the SINR outage threshold is disabled the maximum average queuing delay (MAQD) of NOMA is lower than OMA due to its higher spectral efficiency. The effect of white noise is higher for NOMA compared to OMA since each NOMA user has larger bandwidth. As the SINR outage threshold increases, the MAQD increases for both NOMA and OMA; however, the rate of increase for NOMA is higher than OMA due to the white noise effect. The MAQD of NOMA becomes higher than OMA beyond a certain SINR threshold which depends on the network settings.

In addition, the optimization framework using the extended analytical model is applied for the performance evaluations of the 5G NR concept when the NOMA is utilized. The numerical results show that when the SINR outage thresold is disabled the MAQD performance of NOMA outperform OMA for all 5G NR frame types. However, for the 5G frame types having wider bandwidth, when the SINR constraint enabled and set to higher levels, OMA becomes more preferable than NOMA due to higher noise effect over wider bandwidth.

1.1 Main Contributions

The main contributions of this thesis, which is summarized in Figure 1.1, can be described as follows:

1. We investigate the radio resource allocation problem for NOMA downlink systems and propose the joint user grouping and power allocation mechanism under full buffer traffic and perfect CSI at the base station in Chapter 3. The objective of the multi-user resource allocation scheme is to maximize the geometric mean of the

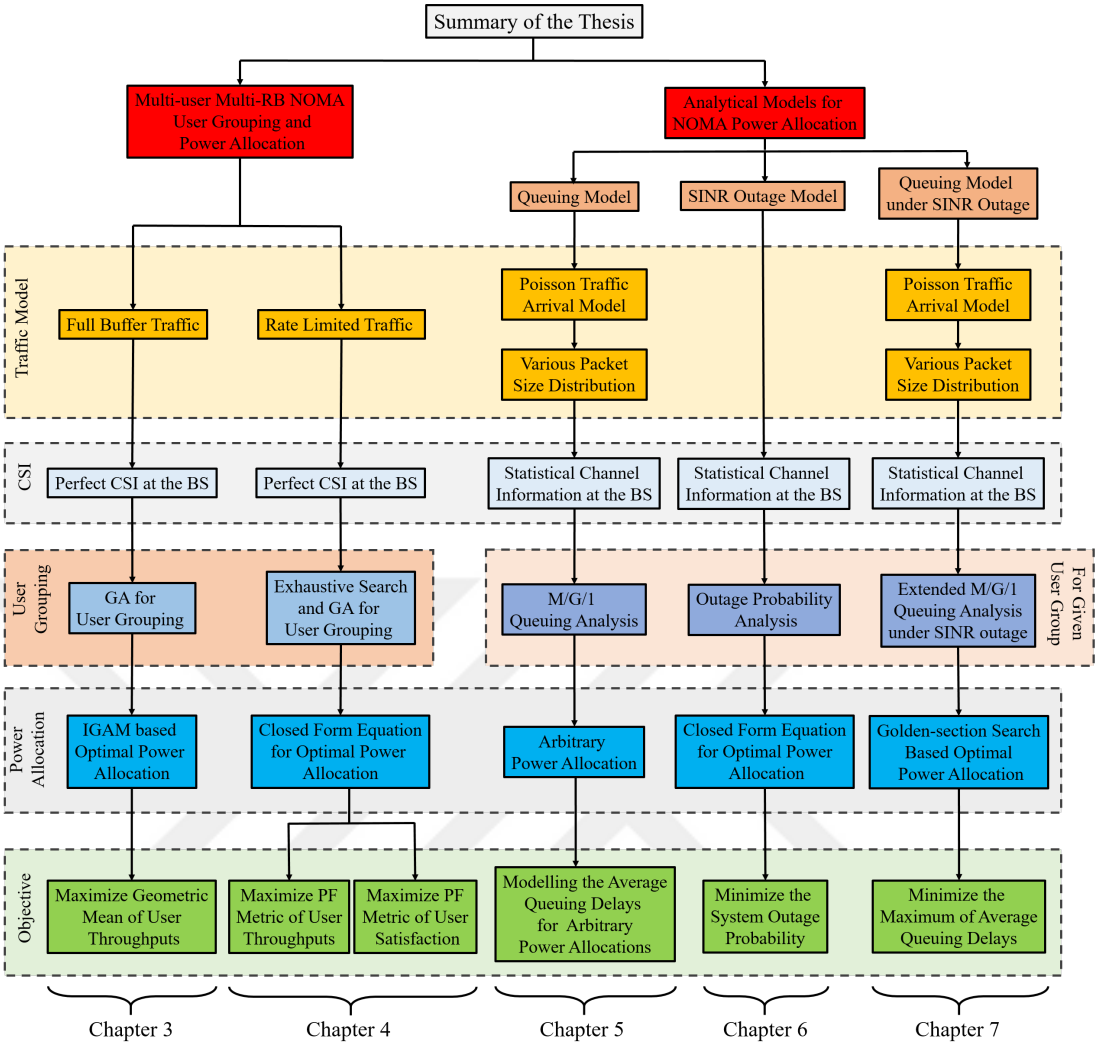


Figure 1.1 : The summary of the thesis.

user throughputs to achieve the optimal trade-off between total system throughput and fairness among users within a single metric as stated in [4].

(i) Genetic algorithm (GA) based user grouping heuristic is proposed while providing the optimal transmission power levels at each user group.

(ii) For each candidate user grouping, an iterative gradient ascent-based power allocation method (IGAM) is utilized to maximize the geometric mean of user throughputs.

2. The new proportional fairness (PF) based multi-user grouping and power allocation schemes are proposed for NOMA downlink systems under perfect CSI and non-full buffer traffic model where the traffic demand for each user is limited to a certain application rate in Chapter 4.

(i) User Demand Based Proportional Fairness (UDB-PF) resource allocation scheme is proposed as an extension of a PF based scheduling to take time varying user traffic demands into account. In addition, the optimum power levels are allocated towards satisfying the traffic demand constraints of user pair in each resource block.

(ii) Proportional User Satisfaction Fairness (PUSF) resource allocation scheme is proposed to maximize the network-wide user satisfaction which is the product of average satisfaction values of all users for a given time window. The user satisfaction objective for the resource allocation optimization is defined by us for the first time.

(iii) The simulation results show that UDB-PF yields higher sum-rate (throughput) while PUSF provides higher network-wide user satisfaction results compared to the conventional PF based user scheduling. The performance gains of the proposed methods increase as the variation of user traffic demands increases over time.

(iv) Due to the high computational load of the proposed PF-based algorithms (i.e., UDB-PF and PUSF), GA heuristic is utilized to find the user group at each resource block with a relatively low computational load. The complexity analysis of PF based algorithms and their GA accelerated extensions reported. When the number of users in the network gets higher, the GA heuristics provide the performance gain on the computational load while the throughput and user satisfaction results are only slightly degraded.

3. An analytical model to characterize the average queuing delay for NOMA downlink systems is proposed in Chapter 5 by utilizing a discrete time M/G/1 queuing model when the statistical channel information is known at the base station. The packet arrival process is assumed to be Poisson distributed with various packet size distributions while the departure process depends on power allocation for a given network setting.

(i) The first and second moment statistics of the user service capacities are formulated for a single resource block using transmit power, bandwidth, and channel model and power allocation coefficients.

(ii) We utilize the random sums of independent and identically distributed (i.i.d.) random variables approach to provide an approximation for the service time statistics under a certain packet size distribution.

(iii) Pollaczek Khintchine formula and Little's Law are applied to obtain the average queuing delay using the derived service time statistics.

(iv) Extensive simulations are carried out to validate the accuracy of the proposed analytical model under different network settings including bandwidth, traffic arrival rate, and packet size distribution.

(v) The proposed analytical model is applied to evaluate the performance improvements of the 5G NR concept when the NOMA is utilized with the 5G NR frame types. The results confirm that the 5G NR significantly improves the delay performance as the frame type having wider bandwidth and shorter duration is employed.

4. The optimum power allocation that minimizes the system outage probability in NOMA downlink systems is proposed in Chapter 6 when the statistical channel information is known at the base station and without considering any traffic information.

   (i) The system outage probability is derived under the Rayleigh fading channel model when the common SINR outage threshold is used as the outage condition.

   (ii) The optimum power allocation that minimizes the system outage probability is provided as a closed form expression for two-user NOMA downlink systems.

   (iii) The accuracy of the theoretical derivations are validated with the Monte Carlo simulations. The results show that the proposed optimum power allocation yields the minimum system outage probability among all the power allocation schemes of NOMA. However, the outage probability of OMA with the fractional power allocation is lower than NOMA with the optimum power allocation. These results indicate that the trade-off between the outage and spectral efficiency in NOMA and OMA should be carefully controlled to meet higher throughput and lower latency objectives of 5G.

5. The proposed analytical model in Chapter 5 is extended to characterize the queuing delay of NOMA downlink systems under the SINR outage constraint and presented in Chapter 7 with the following contributions:

   (i) The first and second moment statistics of the users service rates are are derived for a NOMA downlink system simultaneously serving K users sharing a single resource block under a common SINR outage threshold which is the minimum required level to successfully perform both of the SIC and decoding processes. Similar to the queuing model proposed in Chapter 5, for a given probability distribution of the packet size, a fairly close analytical approximation of the first and second moment statistics for the users' service time is obtained. The underlying queuing system with Poisson traffic arrivals becomes M/G/1, where the Pollaczek Khintchine formula of the residual service approach together with the Little's Law are utilized to obtain the average queuing delay.

   (ii) Utilizing the analytical model, we prove that the maximum of average queuing delays for two-user NOMA and OMA systems is a unimodel function with a single minimum point for the power allocation yielding stable queues. The optimum power allocation framework is proposed by using the M/G/1 queuing model such that the maximum of average queuing delays is minimized for a single resource block simultaneously serving two users.

   (iii) The delay performance of NOMA and OMA is reported using the proposed delay optimization method under various network settings such as SINR outage threshold, user arrival rates and distances. The numerical results show that without considering the SINR outage constraint, the ergodic capacity region of NOMA is always a superset of OMA due to its higher spectral efficiency as demonstrated in Chapter 5. As the SINR outage threshold increases, the average queuing delay increases for both NOMA and OMA; however, the rate of increase for NOMA is higher than OMA due to the white noise effect over larger bandwidth. The proposed model in this paper show that NOMA can yield higher delay when the SINR outage threshold is set to higher levels.

   (iv) The delay optimization framework is applied for the 5G NR concept when the NOMA is utilized. The results demonstrate that OMA becomes more

preferable than NOMA due to higher noise effect over the 5G NR frame types having wider bandwidth for higher outage thresholds. For a given network scenario including the SINR outage threshold that satisfy reliability requirement of 5G URLLC services, our proposed model is capable of determining the frame type that achieves the lowest delay performance for both NOMA and OMA.

1.2 Organization of the Thesis

Following this introductory chapter, Chapter 2 provides an overview of NOMA resource allocation concept for both user grouping and power allocation in addition to summarizing the related studies. The genetic algorithm based NOMA downlink resource allocation scheme is presented in Chapter 3, where the user traffic requirements are not taken into account. In Chapter 4, the rate limited user traffic demands are introduced and two novel resource allocation mechanisms are proposed to maximize either the proportional fairness among user service rates or proportional fairness among user satisfactions. Chapter 5 is devoted to the presentation of the analytical model which characterizes the average queuing delay of NOMA downlink systems when the statistical channel state information is known at the base station. The outage probability analysis and the optimum power level assignment minimizing the system outage probability is presented in Chapter 6, where a common SINR outage threshold is utilized as the outage condition. In Chapter 7, the extended analytical model by taking the SINR outage condition into account is presented to evaluate the queuing delay dynamics of NOMA downlink systems. In addition, the optimum power allocation framework is proposed by using the derived analytical models such that the maximum of average queuing delays is minimized. Finally, Chapter 8 concludes the thesis and includes suggestions for future work.

2. RELATED WORK FOR NOMA RESOURCE ALLOCATION

Radio access technologies for cellular communications are characterized by multiple access schemes to share a finite amount of radio resources to multiple users simultaneously. As cellular technology has advanced different multiple access schemes have been standardized and used to satisfy the technological requirements. The different multiple access technologies employed in the cellular systems are Frequency-division Multiple Access (FDMA) for the first generation (1G), Time-division Multiple Access (TDMA) for the second generation (2G), Code-division Multiple Access (CDMA) used by both 2G and the third generation (3G), and Orthogonal Frequency division Multiple Access (OFDMA) for 4G. These conventional multiple access schemes share the radio resources to the multiple users providing the orthogonality in either time, frequency, or code domain to prevent the multiple access interference. Therefore, they categorized as orthogonal multiple access (OMA) technologies [21]. Illustrative examples of conventional multiple access schemes are given in Figure 2.1.

Three broad categories of 5G services considered by 3GPP are enhanced mobile broadband (eMBB), ultra reliable low latency communication (URLLC), and massive machine-type communications (mMTC). Non-orthogonal multiple access (NOMA) is a promising technology for 5G systems due to its higher spectral efficiency potentially yielding lower latency and higher scalability results by allowing simultaneous transmission of multiple users at the same resource block.

2.1 Conventional Multiple Access Technologies

Frequency-division Multiple Access (FDMA) technology is used to support multiple analog voice calls at the same base station for the system named Advanced Mobile Phone Service (AMPS) [22]. It was the first cellular concept to reuse the frequency spectrum between cells. At each base station, calls are assigned to different frequency

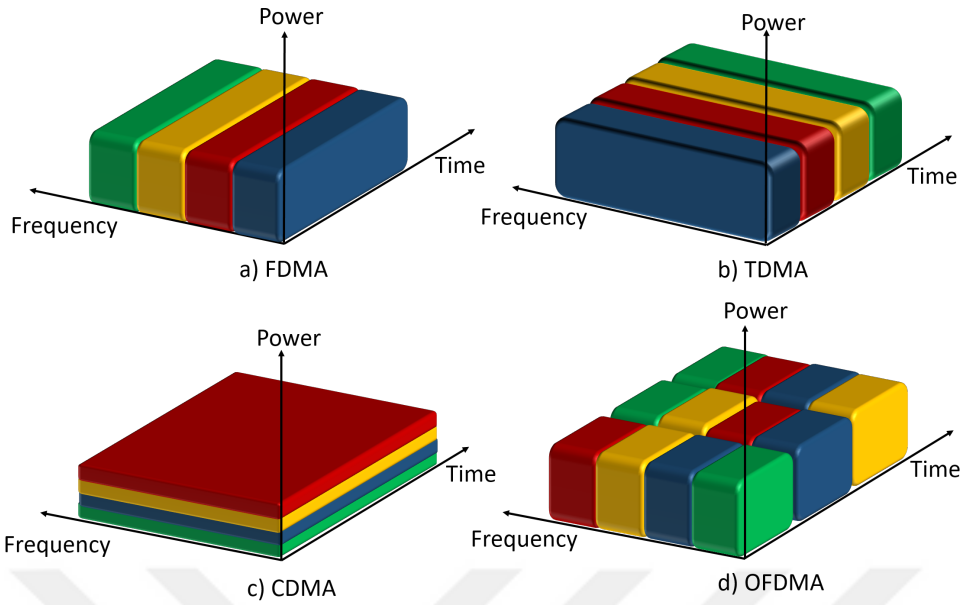


Figure 2.1 : Conventional multiple access schemes.

channels. In FDMA, the spectrum is partitioned into non-overlapping channels which can accommodate traffic of a particular user.

In Time-division Multiple Access (TDMA) systems, radio resources are partitioned in time domain into multiple time slots which can be independently assigned to users. At each time slot the frequency and power resources are only available for the assigned users and time domain orthogonality is provided. The analog system is replaced by Global System for Mobile Communication (GSM) [23] where Time-division Multiple Access (TDMA) is used as a multiple access scheme. The digitized voice packets are transmitted over TDMA's allocated slots in addition the earliest form of mobile data services.

To satisfy the larger mobile data service requirements Code-division Multiple Access (CDMA) based systems were developed to be deployed in cellular systems. In 3G systems, CDMA is introduced to obtain more degrees of freedom in terms of resource partitioning such that multiple users can utilize the same resource both in time and frequency domain by user specific orthogonal code signatures. Each mobile terminal has an unique spreading codes to isolate corresponding signal received from base station. The major advantage of CDMA over TDMA systems was the superior statistical multiplexing at the radio signal level. Since every conversation alternates periods of speaking and listening with interspersed periods of silence, in TDMA

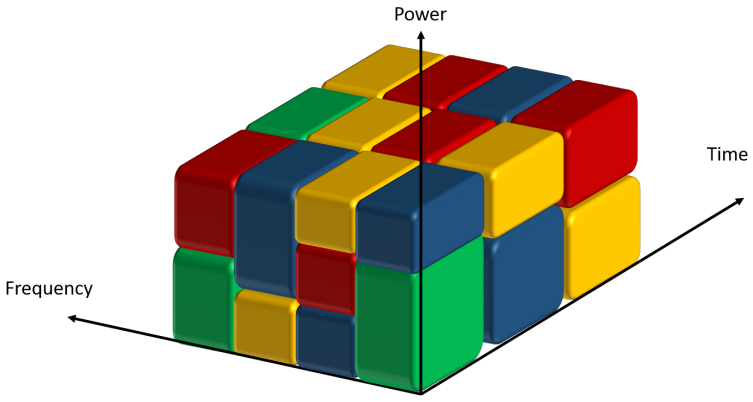


Figure 2.2 : An illustration of NOMA.

systems these periods of silence result in wasted resources since time slots are assigned for entire session [21]. Universal Mobile Telecommunication System (UMTS) and CDMA 2000 are the standards of the 3G era utilizing the CDMA as a multiple access technology [24].

Orthogonal Frequency-division Multiple Access (OFDMA) is the multi carrier multiple access scheme employed in 4G communication systems. The orthogonal subcarriers in the frequency domain and the time slots in the time domain are the resource allocation units to be assigned to users. OFDMA is more flexible than the other conventional multiple access schemes as it allows transmitter to divide radio resources into both the time and the frequency domain. Orthogonal Frequency Division Multiplexing (OFDM) is the one of the key difference between 4G and 3G systems to satisfy the high data rate requirements. OFDM is the technology that emerged to achive high data rates. Long Term Evolution (LTE) which is designed to increase the capacity and speed of cellular networks, brings the mobile communication to 4G era [25]. LTE utilizes OFDMA at the downlink transmission to allow that multiple users can receive their traffic over OFDM.

2.2 Non-orthogonal Multiple Access (NOMA)

The conventional multiple access schemes mentioned in the previous section are categorized as orthogonal multiple access (OMA) technologies, where different users are allocated to orthogonal resources in either time, frequency, or code domain. Although these multiple access schemes provide enough capacity to the users until the 4G era, the use of non-orthogonal multiple access (NOMA) techniques can be

instrumental in meeting 5G requirements such as, high system throughput, low latency and massive connectivity using the readily available frequencies. Sharing the same time and frequency resource with different power levels is the main idea behind the NOMA technology which is a candidate multiple access scheme for 5G standards [1,2]. An illustration of the NOMA scheme is given in Figure 2.2, where the same time and frequency resources are assigned to multiple users. Power domain multiplexing at the transmitter and signal separation at the receiver using successive interference cancellation (SIC) are key elements of NOMA.

A 5G new radio (5G NR) access technology is introduced with shorter frame duration and wider bandwidth to satisfy the lower latency requirements of URLLC services such as industrial control and automation, augmented and virtual reality, tactile Internet and intelligent transportation [10, 11]. The Multi-User Superposed Transmission (MUST) has been proposed by the pioneering technology companies such as Huawei, Qualcomm, NTT DOCOMO, Nokia, Intel, Alcatel Lucent to 3GPP in order to standardize the NOMA, which has been studied in 3GPP Releases 13 and 14 and is under consideration at the standardization activities for 5G NR, can be another instrument to potentially decrease the radio access latency for URLLC services [26–29]. For example, the application of 5G URLLC services for autonomous vehicular networks and road safety applications has been an active research topic [13, 30, 31]. Non-orthogonal multiple access (NOMA) may be a candidate technology for 5G NR due to its higher spectral efficiency potentially yielding lower radio access latency by allowing simultaneous transmission of multiple users at the same resource block [2, 26, 27, 32]. The system-level performance evaluation of NOMA shows that, it has been a promising multiple access technology due to its high spectral efficiency, massive connectivity, low latency, and high user fairness [3].

The recent NOMA solutions can be classified into two main categories such that code domain and power domain NOMA techniques. The code domain NOMA schemes including Low Density Spreading (LDS) [33], Sparse Code Multiple Access (SCMA) [34], Pattern Division Multiple Access (PDMA) [35], Multiuser Shared Access (MUSA) [36], etc, utilizes user-specific spreading sequences to provide separation at the receiver [37]. On the other hand, power domain NOMA (PD-NOMA) exploits the power domain to serve multiple users in the same time and frequency resources,

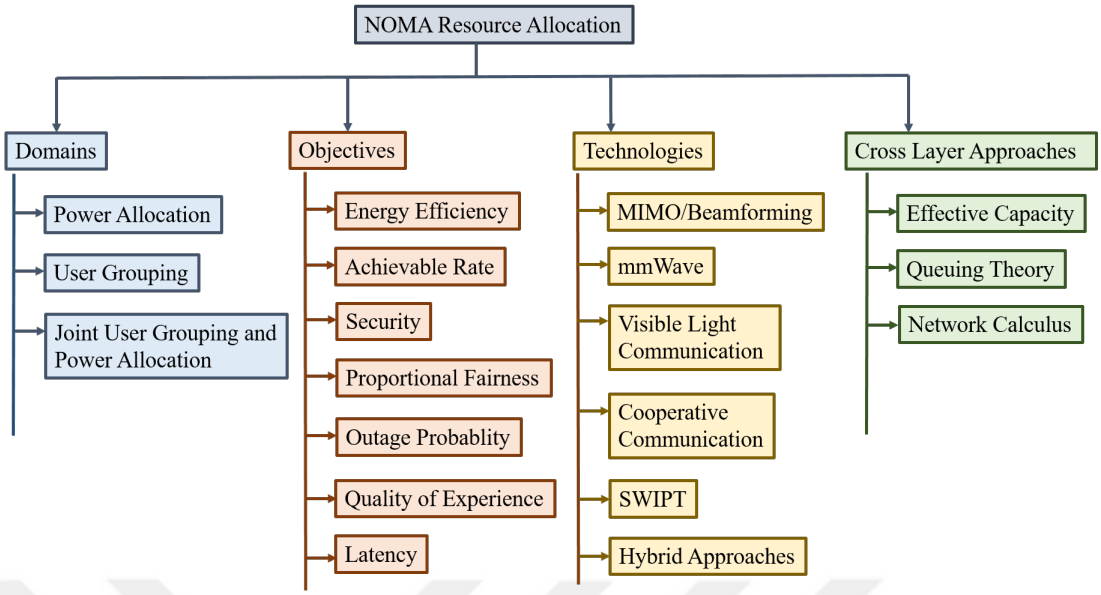


Figure 2.3 : The summary of the related studies.

and performs successive interference cancellation (SIC) at users with better channel conditions [3]. In this thesis, PD-NOMA is utilized as the NOMA technology while investigating the resource allocation mechanisms in addition to the analytical models.

The studies in [1] and [2] present the concept and practical consideration about NOMA with an SIC based receiver, where the system level performance of NOMA can be 30% higher than orthogonal multiple access (OMA). In [38], it is demonstrated that NOMA can achieve better outage performance compared to OMA if the user rates and power levels are appropriately selected but NOMA does not provide much gain when channel qualities of users are low. The comparison of NOMA with other candidate multiple access schemes for 5G standards show that NOMA can improve downlink throughput by more than 30% using power domain multiplexing at the expense of higher complexity at the receiver [3].

2.3 Resource Allocation for NOMA

User scheduling and power allocation in NOMA systems determine the mapping of users to radio resource blocks and the transmission power levels of users at each resource block, respectively. The procedures and algorithms used in the decision making process directly affect the performance of NOMA in terms of its spectral efficiency and computational power requirements. The literature review presented in

this thesis is classified into four categories as resource allocation domain, underlying objective function, utilized technologies, and cross-layer approaches. Figure 2.3 gives the summary of the related studies for NOMA resource allocation.

2.3.1 Domains

The power allocation is one of key factors that affect the performance of NOMA and it is generally considered as an independent problem and solved either after user scheduling or as a sub-task of the user scheduling algorithm. The fixed power allocation method uses pre-defined constant power allocation ratios among users with low and high quality channels [5, 20]. The fractional power allocation method distributes power among users as inversely proportional to their channel qualities [5]. Iterative and non-iterative water filling algorithms [5, 39] are proposed to allocate the power among NOMA users sharing the same resource block. In our methods, we use a modified version of the optimum power allocation strategy presented in [40] to take non-full buffer user traffic model into account.

In [41], a power allocation scheme is proposed to maximize the sum rate under the constraints of total power and minimum rates for the NOMA system with one transmitter and two receivers. [42] presents two sub-optimal power allocation schemes which target to maximize the sum rate while guaranteeing the minimum rate condition with a low computational load. Power allocation can also be used to provide proportional fairness among users by adjusting the transmission power levels of users in NOMA [43].

For the user grouping of multi-user NOMA system, if the members of a group occupying the same time and frequency resource are not carefully selected, the system performance may be severely degraded due to the interference effect. The performance gain of NOMA over OMA can be significantly improved if users whose channel conditions are more distinctive are selected as a group [20, 44, 45]. For example, in [44], it is shown that the performance gain of NOMA over OMA can be significantly improved if users whose channel conditions are more distinctive are selected as a group. In another user grouping strategy proposed in [46], two far users having similar gains with a single near user are grouped together to maximize the spectral efficiency. In addition, matching theory-based algorithms [39, 47], are widely used to determine

the user group selection for NOMA systems due to its lower complexity. In another user scheduling strategy [48], an optimal user grouping algorithm is proposed to maximize the sum-rate after guaranteeing a certain minimum rate for a particular user. These studies focus on multi-user scheduling and user grouping without optimizing the power allocation while our proposed schedulers optimize the power allocation within each group after the users are grouped and mapped to resources.

Some studies in the literature consider joint optimization of multiuser scheduling and power allocation for NOMA systems [9, 49]. In [49], multi-user scheduling and power allocation for the NOMA system are considered jointly to satisfy proportional fairness among users. In order to maximize both fairness and throughput of users in a cell, [50, 51] presents both multi-user scheduling and power allocation per frequency block by maximizing the product of the average user throughput. In another study [52], a game-theory based joint power allocation and user grouping for NOMA downlink system is proposed which maximize sum rate and reduce the outage probability of the mmWave-NOMA.

Energy efficient dynamic resource and power allocation for NOMA networks is investigated in [53] under the QoS constraints, the queue stability, and the power limits for non-full buffer traffic. These objectives are jointly optimized during the resource and power allocation by utilizing the Lyapunov optimization method for stochastic traffic arrivals. Similarly, we consider non-full buffer stochastic traffic arrivals for NOMA downlink systems. However, we propose to PF based algorithms to consider both fairness and throughput and a GA based heuristic to reduce the computational load. An optimal resource allocation method jointly considering power allocation, user scheduling, and rate allocation is proposed in [54] with the use of imperfect channel state information and QoS requirements informations of user terminals. In [55], quality of experience based NOMA system is proposed using a cross-layer approach under finite buffer traffic model. The system level performance of NOMA downlink system is investigated using non-full buffer traffic models in [56] reporting that NOMA provides higher performance gain over OMA especially for small packet sizes. Similar to [56], our PF-based schedulers are designed for non-full buffer traffic. However, we have systematically investigated the effect of traffic demand variations on the performance of NOMA in Chapter 4 when traffic demand rates of users are independently selected

and change over time. We also employ optimum power allocation within each NOMA group while pre-defined power allocation levels are used in [56].

NOMA can achieve better outage performance compared to OMA if the user grouping together with their power levels are appropriately selected [38]. However, NOMA does not provide much gain when channel qualities of users grouped in the same resource block are low. Cooperative NOMA can decrease the outage probability of users having bad channel conditions since users with higher quality channels can be used as relay nodes to forward other users' data [57]. The usage of NOMA supporting MIMO systems is reported in [58]. NOMA is integrated with beamforming systems with multiple users in [59] such that users are classified into multiple groups and users in each group share a beamforming vector. The usage of NOMA in energy harvesting systems transferring wireless power provides advantages in terms of the transmitted data and the harvested energy [60, 61]. In another set of applications [62–64], NOMA is used for improving the physical layer security. In [64], secure resource allocation has been studied for NOMA two-way relay wireless networks and novel matching and power allocation methods are presented to improve the secrecy energy efficiency.

In a cognitive radio based power allocation method [65], the user with lower channel quality is considered as the primary user and the other user is allocated a power only if the requested service quality of the primary user is not adversely affected. The power allocation method proposed in [66] guarantees the total data rate of all users sharing the same resource in NOMA to be at least the data rate of OMA. In [41], a power allocation scheme is proposed to maximize the sum rate under the constraints of total power and minimum rates for the NOMA system with one transmitter and two receivers. [42] presents two sub-optimal power allocation schemes which target to maximize the sum rate while guaranteeing the minimum rate condition with a low computational load. Power allocation can also be used to provide proportional fairness among users by adjusting the transmission power levels of users in NOMA [43]. In [49], multiuser scheduling and power allocation for a NOMA system are considered jointly to satisfy proportional fairness among users. The proposed PF-based methods in Chapter 4 not only provide the proportional fairness by utilizing the similar approach but also take the user traffic variations into account using non-full buffer traffic model.

2.3.2 Objectives

The total power consumption can be minimized for NOMA systems under QoS requirements to improve energy efficiency the due to its spectral efficiency. The subchannel assignment and power allocation are jointly optimized in [67] to maximize the energy efficiency where the CSI is perfectly known at the base station. Since it a challenging task to obtain the perfect CSI, [68] presents the joint resource allocation scheme which maximize the energy efficiency with imperfect CSI. In [69], an optimum power allocation that minimizes the transmit power under the throughput outage constraint is proposed. In another study [70], an optimum power allocation for minimizing the transmit power (i.e., energy efficiency) under the outage constraint is proposed.

The achievable rate regions of the NOMA downlink systems under the outage constraints are presented in [71] using the channel statistics based SIC ordering. In [66], the expressions of the average user throughput is provided for both NOMA downlink and uplink systems under a Rayleigh fading channel model by considering target data rates as constraints. The outage probability and ergodic capacity expressions are derived for a two-user NOMA uplink system such that the same spectrum is shared by multiple device-to-device (D2D) user pairs [72]. In addition, the average user throughput for a Rayleigh fading channel model is formulated to analyze the secrecy capacity for NOMA downlink systems [73]. Since these studies focus on only modelling of the throughput, they do not provide higher order statistics of the service rates. In this thesis at Chapters 5 and 7, the first and second moments of the service rate statistics are derived for the objective of characterizing the latency dynamics of both NOMA and OMA users under a Rayleigh fading channel model.

The proportional-fairness (PF) based approaches [5–7] have been widely used for resource and power allocation in NOMA systems. The objective of the PF based scheduler is to assign radio resources to users in such a way that the PF metric, which is the product of average user throughputs over a time window, is maximized. This objective provides a good compromise between the sum-rate of all users (i.e., network-wide throughput) and the fairness among users. In [5], the user pairing corresponding to the highest PF metric is selected among all possible user pairing combinations. Since this approach requires prohibitively expensive computational

power, [6] and [7] propose simplified PF based algorithms which require significantly lower computational power while yielding comparable results to the optimum solution. Another approach [8, 9] to resource allocation for NOMA systems aims to maximize the sum-rate of all users at each time epoch after satisfying certain constraints such as the minimum power allocation and throughput of each user. All of the above PF based resource allocation studies for NOMA systems assumes full buffer traffic model which does not correspond to real life traffic scenario. In Chapter 4, the non-full buffer traffic model is utilized, where the traffic demand for each user is limited to a certain application rate.

The outage probability is an important metric that can be used to characterize the reliability and latency of wireless networks. For example, the hybrid automatic repeat request (HARQ) is heavily utilized to re-transmit lost data in outage causing additional overhead and latency [74]. The outage event can be defined for cellular systems using various performance metrics such as maximum delay, minimum throughput, minimum BER, and minimum SINR levels. The outage analysis is provided in [38, 66, 71, 75] when each user has a different rate constraint. A closed-form formulation of individual user's outage probability under the Nakagami-m channel together with the optimum power allocation in terms of power efficiency is presented in [76]. The outage probability and ergodic capacity expressions are derived for two-user NOMA uplink system, where the same spectrum is reused by the device to device NOMA user pairs [72]. In [77] and [71], a system outage occurs if any or both of the users are in the outage state. We have utilized this definition in the same Chapter 6 and provided the optimum power allocation coefficients as a closed form expression which minimize the system outage probability.

The study in [78] investigates the outage probability of OMA downlink transmission, in which the transmitter knows the probability distributions of the fading. In [66], the expressions of the average user throughput is provided for both NOMA downlink and uplink systems under the Rayleigh fading channel model by considering target data rates as a QoS constraints. A closed-form formulation of individual user's outage probability under the Nakagami-m channel considering energy constraint is presented in [79]. The SINR outage constraint is considered in [73] to analyse the individual and system outage probabilities in addition to the secrecy capacity of the NOMA

system under Rayleigh fading channel. These studies assume that the transmitter has the probability distributions of the fading coefficients instead of their realizations. By following a similar approach, the proposed models in Chapters 6 and 7 takes the SINR outage constraint into account for both of the decoding and SIC processes at the receiver. On the other hand, the aforementioned studies focus on only modelling of the throughput, they do not provide higher order statistics of the service rates under the outage constraint. In Chapter 7, the first and second moment statistics of the service rate are derived by considering the SINR outage constraint to characterize the latency dynamics of individual users for both NOMA and OMA systems under the Rayleigh fading channel.

In [66], the expressions of the average user throughput is provided for both NOMA downlink and uplink systems under the Rayleigh fading channel model by considering target data rates as constraints. A closed-form formulation of individual user's outage probability under the Nakagami-m channel considering energy constraint is presented in [79]. The study in [78] investigates the outage probability of OMA downlink transmission, in which the transmitter knows the probability distributions of the fading. The SINR outage constraint is considered in [73] to analyse the individual and system outage probabilities in addition to the secrecy capacity of the NOMA system under Rayleigh fading channel. The study in [80] investigates the outage probability of OMA downlink transmission, in which the transmitter knows the probability distributions of the fading. By utilizing the similar approach, in Chapter 6 we analyzed the outage probability of the NOMA downlink transmission under the Rayleigh fading channel model. Further, we present the optimum power allocation that minimizes the NOMA system outage probability under the assumption that the transmitter knows only the probability distributions of the fading coefficients.

In [81], the power control policy for NOMA is studied to meet the delay objectives deriving the effective capacity formulation when the channel state information (CSI) is known. Their delay results are obtained only using the simulations while in this thesis we analytically expressed the average queuing delay for NOMA downlink systems when the probability distribution of the underlying channel is known. In general, NOMA resource allocation studies focused on achieving maximal sum-rate capacity, fairness or minimizing latency. Quality of experience (QoE) is one of the key metric

that provide perceptual quality of service (QoS) from the user's perspective [55, 82]. In [55], upper-layer impact of NOMA on the user side is investigated and cross-layer NOMA frameworks for QoE provisioning is presented. They also provide an optimal user clustering scheme to minimize the average QoE loss with dynamic scheduling technique. In another study [83], both sub-channel assignment and power allocation scheme is presented improve the user QoE for multi-cell NOMA networks.

2.3.3 Technologies

NOMA can be applied to different communication scenarios such as device-to-device (D2D), multiple-input multiple-output (MIMO), and cooperative communication. Several relay nodes are used to support a source while transmitting the information to the receiver in cooperative communications. Therefore, the integration of cooperative communications with NOMA can further improve system efficiency in terms of capacity and reliability [84]. NOMA has been enabled for a two-way relay network and its superiority over conventional time division multiple access (TDMA) based scheme is presented by providing the closed form expression of ergodic capacity [85]. The impact on relay selection for NOMA is investigated in [86]. They propose the two-stage relay selection algorithm to maximize the diversity gain in addition to minimize the outage probability. In [87], the secrecy capacity of a cooperative NOMA system is analyzed and the condition of NOMA to outperform OMA is presented. They also provide the optimal power allocation method to maximize the secrecy rate. Besides, beamforming and cooperative relaying systems are considered with NOMA to provide low latency transmission by maximizing the spectral efficiency [88].

Beamforming is the signal processing technique that allows directional communication of wireless communication systems. Enabling beamforming with multi-user cellular systems, the same radio resources can be reused at each beams to increase the system level capacity. Multiple-input multiple-output (MIMO) communications with multi-user beamforming have been widely investigated while NOMA is used as a multiple access scheme. The users are grouped as clusters in MIMO-NOMA and a single beam is utilized by all the receivers of a cluster. In [89], the users with different receive antennas are grouped as a clusters and each cluster of MIMO-NOMA is served by a single MIMO beam. They have also provide a power allocation

method for both inter-cluster and intra-cluster power level assignment in addition to the user clustering algorithm which maximize the total system throughput. A joint user pairing and power allocation scheme is proposed to maximize the energy efficiency multiple-input multiple-output (MIMO) NOMA downlink system in [90]. Their proposed approach utilize the median and the euclidean norm of the MIMO channels to reach the corresponding objective. The co-existence of NOMA and mmWave solutions are investigated in [91] relying on random beamforming. They evaluate the potential line-of-sight (LOS) blockages of mmWave systems by utilizing the stochastic geometry model.

Simultaneous wireless information and power transfer (SWIPT) can also be integrated into NOMA systems. For example, [92] investigate the SWIPT concept in cooperative NOMA systems and provide a framework to increase energy efficiency. In [93], power allocation framework is utilized for power transfer assisted cooperative NOMA systems and the capacity results are provided as closed form expressions. Furthermore, a novel communication scheme including user clustering algorithm is proposed in [94] such that beamforming, energy harvesting, and cooperative NOMA technologies are combined. They also provide the outage probability analysis where the required SINR level is selected as the outage condition.

The application of NOMA in Visible Light Communication (VLC) has been an active research topic as it is a potential candidate for next-generation wireless communications. The comparison between NOMA and OMA for VLC and the superiority of NOMA in terms of the achievable capacity is emphasized in [95]. In [96], the power allocation for NOMA-VLC is studied to maximize the sum of data rates while satisfying the required SINR levels of users. The user grouping and the power allocation for NOMA VLC networks are proposed in [97], where the network-wide sum rate is maximized while the minimum throughput is utilized as a QoS constraint.

Recent studies in the literature present the hybrid multiple access schemes, where NOMA and OMA are jointly utilized according to the network conditions [98–101]. For example, a hybrid NOMA and OMA scheme is proposed at the uplink to enhance the fairness among users by utilizing the Jain's index as a performance criterion [98]. In another study of uplink [99], NOMA is integrated into OMA to provide energy efficiency. To achieve this objective, a joint user grouping and power allocation scheme

is proposed while guaranteeing the minimum rate requirement of each user. The buffer-aided relay selection method is presented to main the sum-rate of the network for hybrid NOMA and OMA scheme [100]. They also investigate the power level assignment of the hybrid scheme in terms of the outage probability according to the user rate requirements. In [101], dynamic power allocation scheme is proposed for hybrid downlink NOMA systems such that when the strong user's channel gain is lower than a threshold, which is determined by the weak user's predefined rate requirement, OMA is utilized.

2.3.4 Cross-layer approaches

Non-orthogonal multiple access (NOMA) may be a candidate technology for 5G NR due to its higher spectral efficiency potentially yielding lower radio access latency by allowing simultaneous transmission of multiple users at the same resource block [2, 26, 27, 32]. The user plane end-to-end delay of packet transmission can be divided into three main parts: radio access, mobile core, and cloud, where the radio access latency between a base station and user equipment includes over-the-air transmission and propagation, queuing, processing, and re-transmission delays [12]. The outage probability analysis has also been taken considerable attention to study the reliability of wireless networks. New analytical models, which can characterize the radio access latency dynamics by taking the outage event into account, are of paramount importance to evaluate the NOMA suitability for URLLC services of 5G NR.

A cross-layer resource allocation approach considering not only wireless channel characteristics in the physical layer but also traffic arrival and queue occupancy information at the link layer should be employed to achieve the challenging latency objectives of 5G [13]. The effective capacity approach in [102] is used to accurately predict several link-level QoS metrics such as delay bounds for admission control and resource reservation in wireless communication systems. In [81], the sub-optimal power allocation policy for NOMA is studied to meet the delay objectives deriving the effective capacity formulation when the channel state information (CSI) is known at the transmitter.. The performance limitations of NOMA in short packet communication for URLLC services are studied by analytically deriving the effective-bandwidth in [15]. The effective capacity of NOMA guaranteeing the statistical delay requirements under

fading channels has been studied in [17, 18, 103]. The bisection-based cross-layer power allocation scheme is proposed in [103], where the max-min effective capacity of NOMA is selected as the optimization objective. Another NOMA downlink study considering short packet transmission for the IoT applications is presented in [104] such that they employ particle swarm optimization technique to minimize the system energy consumption while maintaining a certain level of effective-throughput.

An opportunistic NOMA downlink approach is presented in [14] such that they propose two queues with different priority levels at the base station for all users. The performance limitations of NOMA in short packet communication for URLLC services is studied by analytically deriving the effective-bandwidth in [15]. The performance of NOMA in short-packet communications is studied in [16] and the optimal power allocation scheme is presented to provide fairness among users' throughput while satisfying QoS requirements of URLLC. Another NOMA downlink study considering short packet transmission for the IoT applications is presented in [104] such that they employ particle swarm optimization technique to minimize the system energy consumption while maintaining a certain level of user throughput. The effective capacity of NOMA under statistical delay guarantees has been studied in [17, 18]. In another study [19], a cross-layer approach using integer linear programming is proposed to minimize the average delay for NOMA applications of delay sensitive communication. These studies consider the outage condition as a delay violation constraint while this thesis presents an analytical model in Chapter 5 to characterize the average queuing delay. Furthermore, Chapter 7 presents the extended analytical model for average queuing delay by taking the SINR outage constraint into account.

In [105], the authors utilize the stochastic network calculus approach to study the resource allocation problem for uplink NOMA systems by minimizing the delay violation probability. They stated that NOMA with the SIC decoding may not be suitable for low latency system under realistic system effects such as imperfect CSI. We have also achieved a similar result for two-user NOMA downlink systems when the SINR outage constraint is set to higher levels in Chapter 7.

Stable throughput regions for uplink NOMA systems under unsaturated traffic are investigated using the queuing theory approach, where traffic arrival for each user is

assumed to be independent Bernoulli process [106]. Similarly, the delay analysis of NOMA is studied using the queuing theory approach in this thesis; however, we focus on downlink channels. [107] investigates the average delay minimization problem for two-user OMA networks and show that the optimal resource allocation policy needs to equalize the queue lengths of both users. We present the optimum cross-layer power allocation framework minimizing the maximum of average queuing delays in two-user NOMA downlink system in Chapter 7. Consistent with the proposal in [107], we have analytically shown that the optimal power allocation method yields the minimum average queuing delay by minimizing the difference between the average queuing delays of both users. The queuing analysis of block Rayleigh fading channels for conventional OMA system is presented in [108] by utilizing the discrete time discrete state D/G/1 queuing model. They drive the probability distribution of packet service time by taking advantage of the channel distribution of the low SNR regime. In another study [109], a general state space Markov chain model is proposed to calculate the throughput regions of OFDMA users under a Rayleigh fading channel by taking the scheduling algorithms into account. The buffer overflow probability providing insights for buffer dimensioning problems is obtained assuming that each user has finite traffic arrival and queue capacity. In Chapters 5 and 7, we adopt a similar system model for the NOMA downlink such that each user has a dedicated queue with the packet based random traffic arrival model and the departure process of each queue is determined by the NOMA resource allocation parameters in addition to the Rayleigh fading channel. Since we focus on the latency analysis for the URLLC services, we utilize a discrete-time M/G/1 queuing model to obtain the average queuing delays of both NOMA and OMA downlink systems by taking both arrival and departure models into account. In [110], theoretical queuing analysis and system-level simulations are performed to study the system design principles of 5G NR. They emphasize that the queuing effect has an important contribution on the URLLC latency. They emphasize that the queuing delay has an important contribution on the URLLC latency. Although they study both uplink and downlink models for 5G NR, the NOMA technology is not considered in their model. The average queuing delay of 5G NR frame types for both NOMA and OMA downlink systems has been evaluated using our discrete-time M/G/1 queuing model proposed in this thesis.

3. GENETIC ALGORITHM APPROACH FOR NOMA RESOURCE ALLOCATION

In this chapter, genetic algorithm (GA) based multi carrier NOMA downlink radio resource allocation scheme is proposed to reach a target solution which balances the trade-off between total system throughput and fairness among users. In [4], the authors state that the criterion of a single metric (i.e., maximizing the geometric mean of user throughputs) can achieve the optimal trade-off between total system throughput and fairness among users. Inspired from this study, our objective function is set to maximize the geometric mean of user throughputs in a cell. Our GA based allocation scheme considers not only the transmission power levels of users at the same radio resource but also user grouping to satisfy the maximum geometric mean of user throughputs in a cell. The GA approach is used to determine the user groups for the NOMA downlink system while the optimal transmission power level assignment is applied for each user group. In other words, power allocation and user grouping are jointly considered in our GA based resource allocation approach.

3.1 NOMA System Model

The main principle of NOMA with SIC is illustrated in Figure 3.1, where two users are sharing the same bandwidth with distinct power levels. BS represents a base station transmitting the signals of UE_1 and UE_2 simultaneously. In this example, the signal of UE_2 has more power level than the signal of UE_1. It is assumed that UE_1 as a near user can first decode the signal of UE_2 from the received combined signal by applying SIC process. Then, UE_1 can decode its own signal by removing the decoded signal of UE_2. Although the signal of UE_1 interferes with the signal of UE_2, UE_2 can still decode its own signal since the signal of UE_2 has more power level than the signal of UE_1.

For the NOMA example in Figure 3.1, the transmitted signal x is the sum of $x_1 = \sqrt{p_1}.s_1$ and $x_2 = \sqrt{p_2}.s_2$, where s_1 and s_2 represent signals to be transmitted to UE_1

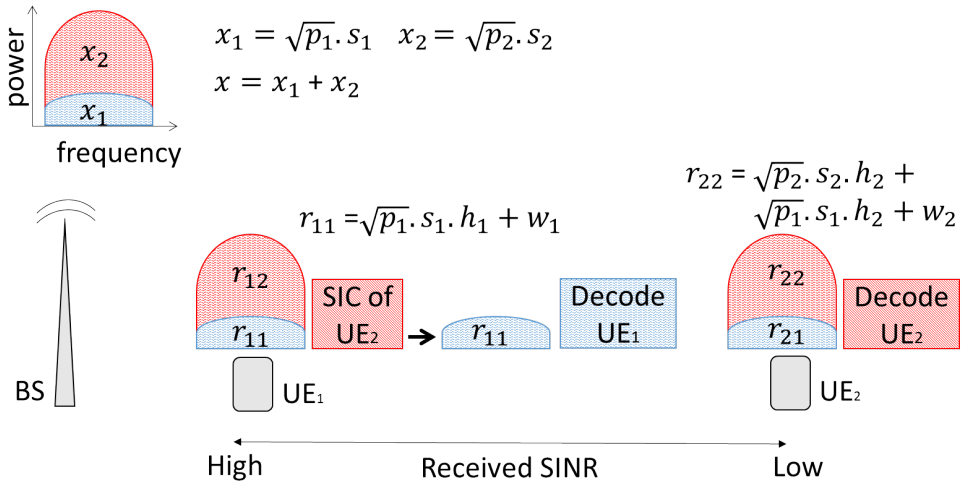


Figure 3.1 : NOMA with SIC concept for two UE receivers in downlink.

and UE_2 at the BS while p_1 and p_2 are power levels of s_1 and s_2, respectively. At the receiver side, $r_1 = r_{11} + r_{12}$ and $r_2 = r_{21} + r_{22}$ represent the received signals while h_1 and h_2 are the channel response for UE_1 and UE_2, respectively. Finally, w_1 and w_2 represent the channel noise for UE_1 and UE_2, respectively. UE_1 first employs an SIC process to remove r_{12} from r_1 so that r_{11} can be successfully decoded. For the error-free SIC process and the transmission bandwidth of 1 Hz, R_{1m} and R_{2m} represent throughputs of UE_1 and UE_2 at the subcarrier m, respectively:

$$\begin{aligned} R_{1m} &= log_2\left(1 + \frac{p_{1k}\,|h_{1m}|^2}{W_{0,1m}}\right), \\ R_{2m} &= log_2\left(1 + \frac{p_{2m}\,|h_{2m}|^2}{p_{1m}\,|h_{2m}|^2 + W_{0,2m}}\right). \end{aligned} \tag{3.1}$$

The generalized form of the throughput for an arbitrary user k at the subcarrier m is:

$$\begin{aligned} R_{km} &= log_2\left(1 + \frac{p_{km}\,|h_{km}|^2}{\sum_{i=1}^{n-1} p_{im}\,|h_{km}|^2 + W_{0,km}}\right) \\ \text{s. t.} \quad & \frac{|h_{km}|^2}{W_{0,km}} > \frac{|h_{0,(k+1)m}|^2}{W_{0,(k+1)m}}. \end{aligned} \tag{3.2}$$

3.2 Multi-user Resource Allocation for NOMA

In this chapter, we consider OFDM based NOMA downlink systems, where each OFDM subcarrier can be assigned to multiple users with distinct power levels. Let N_{max} and M be the maximum number of user per subcarrier and the number of subcarriers, respectively. A resource allocation map represents the assigment of user groups and their corresponding subcarriers. Then, the number of allocation blocks in the resource allocation map is $N_{max} \times M$. In other words, the resource allocation map shows the assignment of the users to $N_{max} \times M$ allocation blocks in a cell. When the number of connected users in a cell is represented as K, the number of all possible resource allocation maps is:

$$\prod_{1}^{M} \sum_{i=1}^{N_{max}} \binom{K}{i} . \tag{3.3}$$

Our objective is to maximize the geometric mean of user throughputs in a cell when the NOMA technology is employed. Towards reaching this objective, one should decide both resource allocation map and the power assignment of users in each subcarrier which require an exhaustive search within a prohibitively huge search space. The optimal user group assignment map (I_{opt}) and power ($\overline{P}_{opt}$) assignment yielding the maximum geometric mean of user throughputs can be defined as follows:

$$(\overline{P}, I) = \arg\max \left\{ \left[\prod_{n=1}^{K} R_k \right]^{1/K} \right\} \tag{3.4}$$

s. t. $R_k = \sum_{m=1}^{M} R_{km}$, $I = \bigcup_{m=1}^{M} I_m$, $\overline{P} = \bigcup_{m=1}^{M} \overline{P}_m$,
$\overline{P}_m = \bigcup_{i \in I_m} p_{im}$ and $1 = \sum_{i \in I_m} p_{im}$

where $\overline{P}$, p_{im}, I_m, and I represent a power assignment vector, the power level of ith member of user group at the subcarrier m, user group at the mth subcarrier, and the resource allocation map including the information of user groups for all subcarriers, respectively.

3.3 Genetic Algorithm Based Allocation Scheme

In this section, we present a multi-user resource allocation scheme to assign available radio resources to the users assuming that the NOMA technology is utilized. The objective is to find the resource allocation which maximizes the geometric mean of user throughputs, where there is a prohibitively huge search space. Genetic Algorithm

(GA) is a population based powerful meta-heuristic for exploring a huge search space in complex optimization problems [111].

In this chapter, the definition of the NOMA resource allocation problem is given for the generic case, where there is no constraint for the number of users, the number of allocation blocks and each user can be assigned to multiple allocation blocks. However, we propose a GA based NOMA resource allocation scheme inspired from the traveling salesman problem (TSP). In the TSP concept, a traveling salesman must plan his trip with a minimum cost assuming that he visits every city in his trip exactly once and the costs of traveling between each pair of cities are known [112]. The TSP version of the GA can be directly applied to NOMA resource allocation problem when the number of users (K) is equal to the number of allocation blocks ($N_{max} \times M$) and each user is assigned to only one allocation block. The search space includes $(N_{max} \times M)^K$ solutions for the generic case and $(N_{max} \times M)!$ solutions for the special condition studied in this chapter.

The work flow of the proposed GA is depicted in Figure 3.2. A certain number of chromosomes corresponding to the number of individuals in the population is initially created. The crossover operation on the current population periodically updates this solution set by generating new offspring, and hence, explores the search space of all possible solutions. After the crossover operation, the selection process is performed to determine surviving individuals according to their fitness values. This process is continued until the stopping criteria is met [111]. Using the same procedure, we propose a GA based NOMA-OFDM multi user allocation scheme to maximize the geometric mean of user throughputs. The details of the proposed scheme are given in the following sections.

3.3.1 Chromosome structure and initialization

Methods for the chromosome structure of TSP such as adjacency, ordinal, and path representations are presented in [112]. We consider that the path representation is the most suitable method for NOMA radio resource allocation because other representations increase the complexity of the problem. Each path in TSP can correspond to the user groups and their corresponding subcarriers in the NOMA resource allocation case. Figure 3.3 shows an example chromosome representing

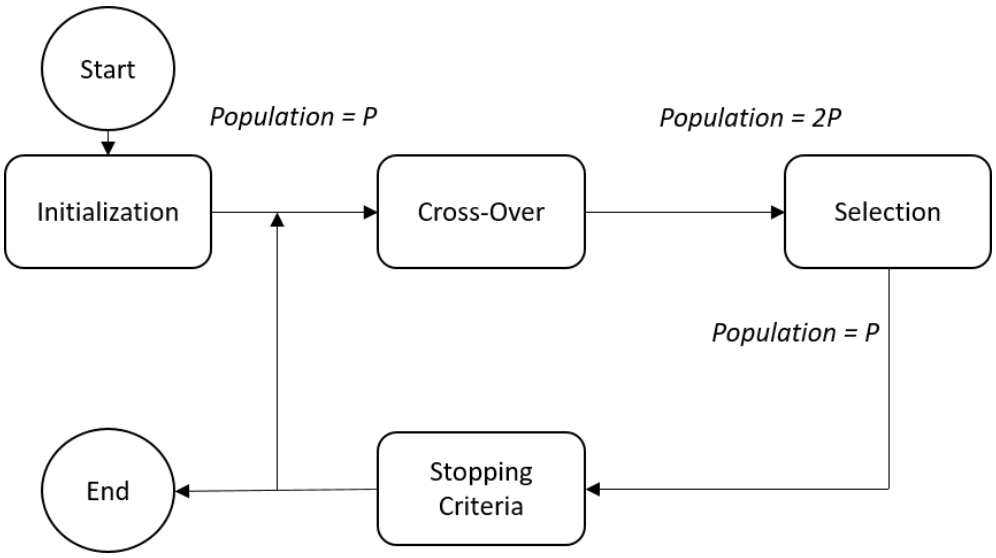


Figure 3.2 : Flow chart of genetic algorithm.

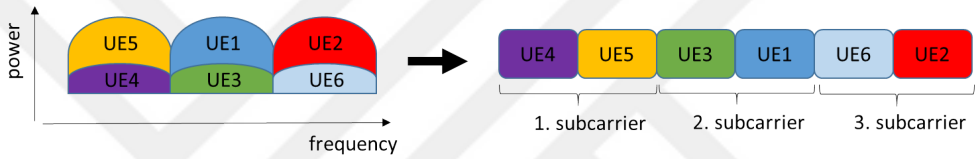


Figure 3.3 : Chromosome structure.

an allocation map solution for the NOMA-OFDM system having 6 allocation blocks when the maximum number of user per subcarrier, N_{max}, the number of subcarriers, M, and the number of users, K are set to 2, 3, and 6, respectively. Each chromosome is composed of 6 genes and each gene corresponds to an allocation block. The assignment of an allocation block to users is shown with a user id (e.g., UE1,UE2,...,UE6), where each id uniquely defines the corresponding user. P individuals are randomly created at the initialization phase, then update and selection processes are repeated at the main loop until the stopping criteria is reached.

3.3.2 Cross-over process

In GA applications, generating new individuals by the cross-over process ensures the search diversity. The order-based cross-over presented in [113] is employed as the cross over process of the GA algoritm. This operation randomly selects several positions in a chromosome vector and the users in the selected positions in the first parent is imposed on the corresponding allocation blocks in the second parent. Figure 3.4 shows an example of the cross-over process. The randomly generated integer numbers 4 and 2 represent the locations of the users in the allocation blocks. The

users corresponding to these positions in the second parent are UE2 and UE6. These users are ordered as UE6 and UE2 and this order is imposed to the the first parent in order to generate a new offspring. Imposing process is defined as changing the value of the allocation blocks with new ordered users in the first parent. The same process is applied for the second parent so that a solution diversity is provided by generating two child chromosomes.

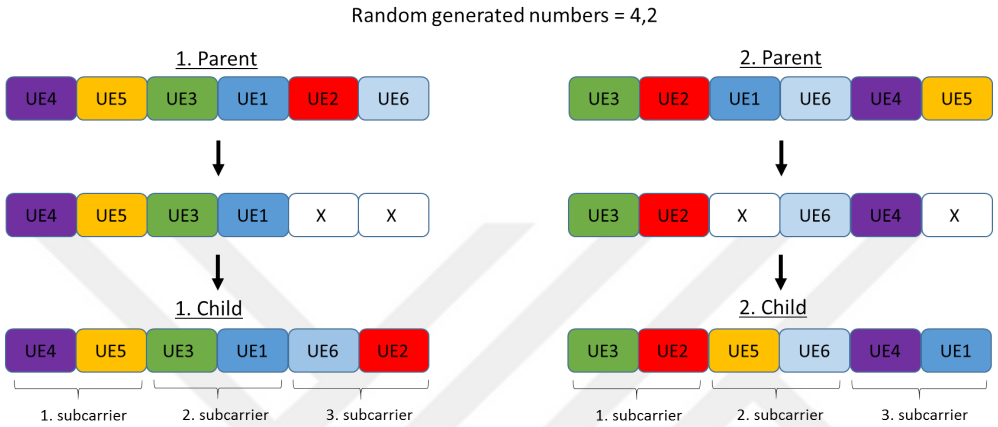


Figure 3.4 : Cross-over process.

3.3.3 Selection and fitness function

The selection operation is performed for $2 \times P$ chromosomes consisting of P initial parents and P offspring. The first P individuals with the highest fitness values are selected to form the new population. The fitness of each chromosome is evaluated using a fitness function presented in Table 3.1, in which the geometric mean of user throughputs for a candidate allocation map (i.e., chromosome) is calculated. Note that the power levels of the users at each subcarrier is determined using power optimization described in the next section.

3.4 Power Optimization

Power allocation at the NOMA process directly affects a user's SINR value and hence is a key parameter for determining the throughput value of each user. Flexibility of controlling the power allocation level at the base station can be used as an effective tool to determine the total throughput level of cell and fairness among users. Maximizing the geometric mean of user throughputs, which is based on proportional fairness, can achieve optimal trade-off between total user throughput and user fairness with a single

Table 3.1 : Fitness function of GA.

input: allocation map
output: fitness value
for each subcarrier (m) in allocation map
I_m : get user group in the subcarrier m
$\overline{g}_m$: get normalized channel gains for I_m
$\overline{P}_m$: find optimal power levels for $\overline{g}_m$
$\overline{R}_m$: calculate user throughputs ($\overline{P}_m$,$\overline{g}_m$)
end
fitness value = calculate geometric mean of all users

metric [4]. Equation (3.5) provides optimal power levels for users at the subcarrier m while N_{max}, R_{km}, and $\overline{P}m = (p_{1m}, p_{2m}, \ldots, p_{N_{max}m})$ representing the maximum number of users per subcarrier, user throughput value, and optimal power levels, respectively.

$$\overline{P}_m = \arg\max \left\{ \prod_{i=1}^{N_{max}} R_{im} \right\} . \tag{3.5}$$

The processed channel capacities can be calculated using the normalized channel gains $g_{km} = \frac{|h_{km}|^2}{W_{0,km}}$ and power levels p_{km}. Therefore throughput formulations can be reconstructed using normalized channel gains and power level values for each user. It is assumed that the user channel conditions are ordered as $g_{km} > g_{(k+1)m}$, so, the inequality of power level values satisfy the condition of $p_{km} < p_{(k+1)m}$ where the total power of each subcarrier m is equal to one ($1 = \sum_{i=1}^{N_{max}} p_{im}$). For $N_{max} = 2$, user throughputs are:

$$\begin{aligned} &R_{1m} = log_2\left(1 + p_{1m} \cdot g_{1m}\right), \\ &R_{2m} = log_2\left(\frac{1 + g_{2m}}{1 + p_{1m} g_{2m}}\right) \\ &\text{s. t.} \quad p_{1m} \in (0, 1/2) \; and \; p_{2m} = 1 - p_{1m} . \end{aligned} \tag{3.6}$$

For $N_{max} = 3$, user throughputs are:

$$
\begin{aligned}
R_{1m} &= log_2\left(1 + p_{1m}.g_{1m}\right), \\
R_{2m} &= log_2\left(\frac{1 + (p_{1m} + p_{2m}).g_{2m}}{1 + p_{1m}.g_{2m}}\right), \\
R_{3m} &= log2\left(\frac{1 + g_{3m}}{1 + (p_{1m} + p_{2m}).g_{3m}}\right) \\
\text{s. t.} \quad & p_{1m} \in (0, 1/3), \quad p_{2m} < (1 - p_{1m})/2, \\
& p_{3m} = 1 - p_{1m} - p_{2m}\ .
\end{aligned}
\tag{3.7}
$$

3.5 Simulation Results

When the geometric mean of user throughputs is numerically calculated for N_{max} is 2 and 3, its distribution forms a convex curve and has only one maximum value under given power constraints. Iterative gradient ascend method (IGAM) is suitable to solve this type of convex optimization problem. Therefore, in this chapter, we employ IGAM to find the optimum power allocations for each user group. There are several power allocation methods proposed in the literature such as fixed power allocation (FPA) [5] and fractional transmission power control (FTPC) methods [1]. The FPA does not take channel qualities into account and determines power levels of users using a pre-defined parameter. On the other hand, FTPC is inspired from LTE uplink transmission power control to equalize the user channel variations.

The optimized power allocation which satisfies the maximum geometric mean of user throughputs for different user channel qualities is evaluated and compared with FPA and FTPC. The simulation setup is configured for one subcarrier and two users for all experiments. While one of the user's SNR value is constant and set to 0 dB, the second user's SNR value varies from 0 dB to 50 dB. These SNR values are calculated before the power allocation process by using normalized channel gains. Figure 3.5 represents the behavior of geometric mean of user throughput values formulated as $\sqrt{R_1.R_2}$ where R_1 and R_2 represent the throughput value of first and second users. The proposed *NOMA power optimized* method satisfies the best geometric mean of user throughput for all SNR variations. For all power allocation methods, the performance improvement of NOMA over OMA increases as the difference between user channel gains gets higher.

The performance of GA based NOMA-OFDM multi user allocation scheme is evaluated and compared with *OMA*, *Random NOMA* and *Exhaustive Search* allocation

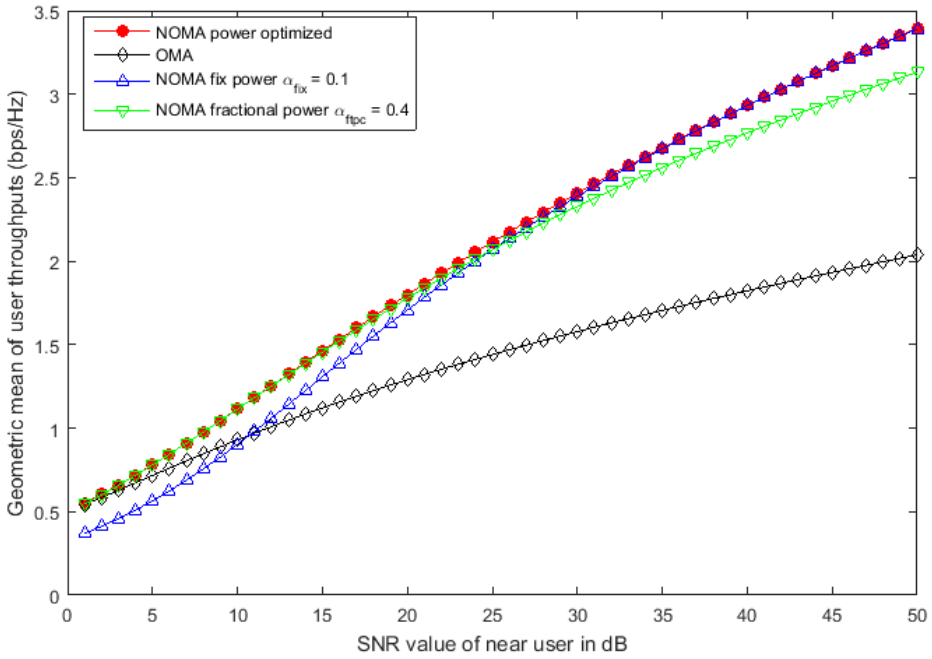


Figure 3.5 : Geometric mean values for two-user case.

schemes. The *OMA* scheme satisfies that each subcarrier is assigned to only one user. Therefore, when the number of subcarriers is equal to the number of users then each user has one subcarrier to satisfy the maximum cell geometric mean. *Random NOMA* allocation scheme randomly selects user groups to be assigned to subcarriers while the power level optimization is employed for each user group. *Exhaustive search* provides the best allocation map among all possible allocation map solutions. This scheme satisfies the best allocation map because of comparing all possible solutions with each other. However, huge computational power requirement makes this allocation scheme prohibitively expensive for practical implementations.

For the simulations, one base station and uniformly distributed user locations in a cell is used. The number of individuals of GA is set to 8 and the maximum number of iterations of 50 is selected as the stopping criteria. The rest of the simulation parameters are given in Table 3.2. For all NOMA multi user allocation algorithms, power optimization among user groups is applied for both $N_{max} = 2$ and $N_{max} = 3$. The simulation results are reported as the average of 1000 experiments for each algorithm.

Figure 3.6 shows the geometric mean variations while the number of users in a cell increases. The first result from this figure is that, even if user grouping in NOMA

Table 3.2 : Simulation parameters for GA based resource allocation.

Parameter	**Value**	
Transmission Bandwidth	1 MHz	
Path Loss Exponent	3	
Transmit Power	10 dBm	
Receiver Noise Density	-169 dBm/Hz	
Shadowing standard deviation	Lognormal with 8 dB	
Fading Model	Rayleigh Fading	
Cell Radius	1 km	
N_{max}	2,3	
Number of users	6,12,18,24,30	
Number of subcarriers	*OMA*	6,12,18,24,30
	$N_{max} = 2$	3,6,9,12,15
	$N_{max} = 3$	2,4,6,8,10

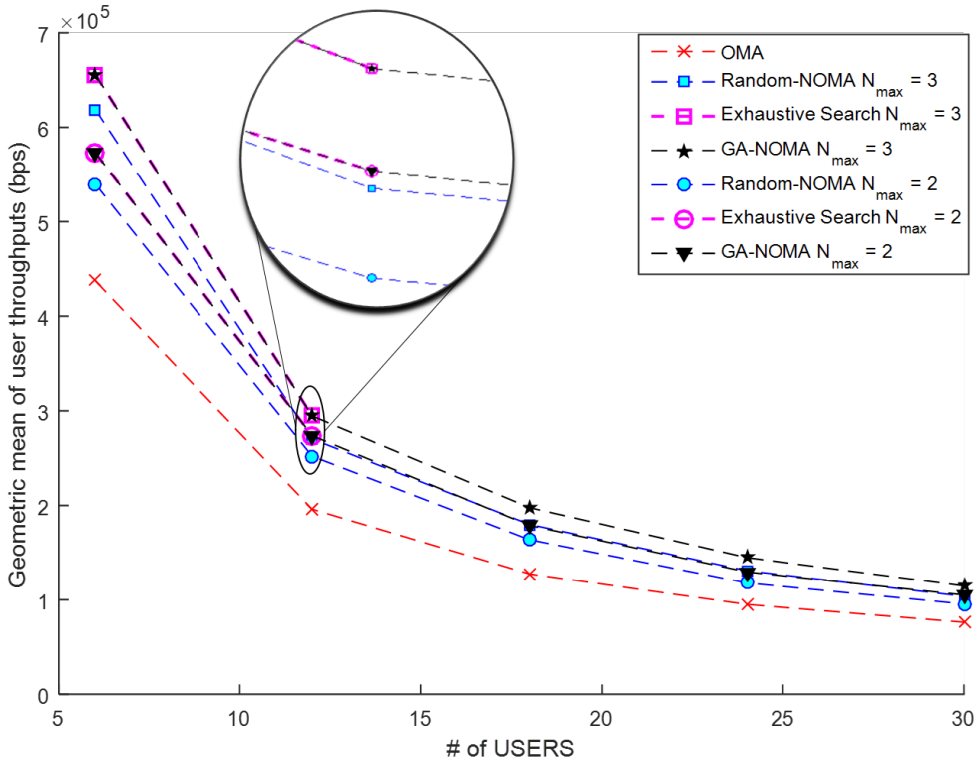


Figure 3.6 : Geometric mean of user throughput versus the number of users.

executed randomly, NOMA has better performance than OMA. Secondly, although it is clear that exhaustive search NOMA has the best performance, GA based NOMA achieves the same performance of exhaustive search allocation scheme. Considering the computational load of exhaustive search, GA is an efficient way to converge to the best solution. Note that, simulations for exhaustive search algorithm are only applied when the number of user is set to 6 and 12 because of higher computational load when the number of users is higher. Finally, it can be said that when the maximum number of users per subcarrier N_{max} increases, the geometric mean of user throughputs increases if the optimized power levels are in use. These results are obtained under the assumption that receivers perfectly execute successive interference cancellation (SIC) process.

3.6 Summary

In this chapter, a downlink radio resource management for wireless NOMA system is studied from the multi-user scheduling perspective towards maximizing the geometric mean of user throughputs. Genetic algorithm (GA) approach is proposed to reach the best resource allocation solution, where the power level optimization is employed for each candidate user group. Simulation experiments show that the proposed method quickly converges to the target solution that balances the trade-off between total system throughput and fairness among users.

Another performance criterion which provides a good compromise between the sum-rate of all users (i.e., network-wide throughput) and the fairness among users is Proportional Fairness (PF) metric. The proposed resource allocation approach in this chapter assumes full buffer traffic model which does not correspond to real life traffic scenario. The traffic model in a real network setting is generally non-full buffer where the traffic demand for each user is limited to a certain application rate. In the next chapter, we have investigated the NOMA downlink resource allocation to maximize the PF metric by taking the rate limited user traffic demands in to account.

4. PROPORTIONAL FAIR NOMA RESOURCE ALLOCATION UNDER RATE LIMITED TRAFFIC

The most of the resource allocation studies for NOMA systems including our Genetic Algorithm approach presented in the previous chapter assumes full buffer traffic model where the incoming traffic of each user is infinite while the traffic in real life scenarios is generally non-full buffer The objective of the PF based scheduler is to assign radio resources to users in such a way that the PF metric, which is the product of average user throughputs over a time window, is maximized. This objective provides a good compromise between the sum-rate of all users (i.e., network-wide throughput) and the fairness among users. This objective is utilized for NOMA downlink systems without considering the user traffic demands [5–8]. In [5], the user pairing corresponding to the highest PF metric is selected among all possible user pairing combinations. Since this approach requires prohibitively expensive computational power, [6] and [7] propose simplified PF based algorithms which require significantly lower computational power while yielding comparable results to the optimum solution. All of the above PF based resource allocation studies for NOMA systems assumes full buffer traffic model which does not correspond to real life traffic scenario. The traffic model in a real network setting is generally non-full buffer where the traffic demand for each user is limited to a certain application rate.

In this chapter, two user scheduling and power allocation methods employing PF based objective functions for NOMA downlink systems under non-full buffer traffic models are proposed. Although the existing PF based user scheduling in NOMA systems has been demonstrated to significantly improve the system capacity when the user traffic model is full buffer, it does not perform well when user traffic rates are limited and time-varying. In User Demand Based Proportional Fairness (UDB-PF) algorithm, the PF based scheduling is extended to take time varying user traffic demands into account in addition to allocating optimum power levels towards satisfying the traffic demand constraints of user pair in each resource block. The main contribution in UDB-PF is to provide the optimum power allocation under user rate constraints. In

other words, when the optimum power level of a user provides higher rate than its rate constraint, the excessive power is reallocated to the other user(s) in the same NOMA group. The objective of Proportional User Satisfaction Fairness (PUSF) algorithm is to maximize the network-wide user satisfaction which is the product of average satisfaction values of all users for a given time window. Note that the highest network-wide user satisfaction is achieved when the resources are sufficient to carry traffic demands of all users. In the PUSF approach, the user satisfaction objective for the user grouping and power allocation optimization is defined by us for the first time. However, the maximization of the product of average user satisfaction is similar to PF based methods. As in the UDB-PF approach, the PUSF can also reallocate the excessive power to the other users in the same NOMA group. In both UDB-PF and PUSF algorithms, user groups are selected first to simultaneously transmit their signals at the same frequency resource while the optimal transmission power level is assigned to each user to optimize the underlying objective function. These proposed algorithms evaluate all user group possibilities to select the best user group allocation at each resource block. However, the computational complexity becomes an important issue when the number of users gets higher, especially to meet the real time requirements of the scheduling decisions. We also present a Genetic Algorithm (GA) heuristic to find the user group at each resource block with a relatively low computational load. The UDB-PF and PUSF algorithms with the GA extensions are named as UDB-PF-GA and PUSF-GA, respectively.

We performed simulation experiments by assuming a single input single output antenna for multi-carrier systems and the base station has a perfect knowledge of channel state information of users. The performance evaluation has been done by varying the number of users and traffic characteristics of each user and sum-rate (throughput) and user satisfaction results are reported. The results show that both algorithms consistently perform better than the PF based user scheduling and converge to the same results if user traffic demands remain constant with time. When user traffic demands change rapidly over time, UDB-PF yields better sum-rate while PUSF provides better network-wide user satisfaction results compared to the PF based user scheduling. UDB-PF-GA and PUSF-GA provide the same performance results with the UDB-PF and PUSF algorithms, respectively when the number of users is relatively

low. When the number of users in the network increases, the performance gain on the computational load significantly increases compared to the UDB-PF and PUSF, while the throughput and user satisfaction results are only slightly degraded.

4.1 System Model

In this section, firstly a NOMA downlink system model and ergodic capacities for user terminals are introduced for a single resource block. Then, the multi user multi resource block NOMA downlink resource allocation system model is presented under the constraint that user traffic rate requirements are limited with a certain data rate. Let us consider a downlink NOMA system consisting of a single base station and multiple users. Compared to the case of a single user allocation to each radio resource block (i.e., frequency subband) in OMA, multiple users can be assigned to each resource block at the same time epoch in NOMA. Assuming that K users are allocated to each frequency subband, the combined signal transmitted from the base station to users can be represented as follow:

$$x = \sqrt{p_1}s_1 + \sqrt{p_2}s_2 + \cdots + \sqrt{p_K}s_K \tag{4.1}$$

where s_k and p_k represent the data symbol per unit energy of kth user and the amount of power allocated for this user, respectively ($1 \leq k \leq K$). The total power allocated to all users should be equal to the amount of power allocated for this subband ($p_1 + \ldots + p_K = P_{sb}$). Assuming that the system is single-input single-output (SISO), the signal of the kth user is:

$$y_k = h_k\sqrt{p_k}s_k + h_k\sum_{i=1}^{k-1}\sqrt{p_i}s_i + h_k\sum_{i=k+1}^{N}\sqrt{p_i}s_i + w_k \tag{4.2}$$

where h_k is the channel gain between the base station and the kth user, w_k is additive white Gaussian noise including both inter-cell and inter-channel interferences, and $W_{0,k}$ represents the spectral power density of the noise. The users should apply SIC procedure to decode their corresponding signals. Since the users with lower channel qualities are allocated more power in the NOMA system, the users with high channel qualities first decode the signals of the users with lower channel qualities and then the decoded signal is subtracted from the combined received signal to obtain their own signal.

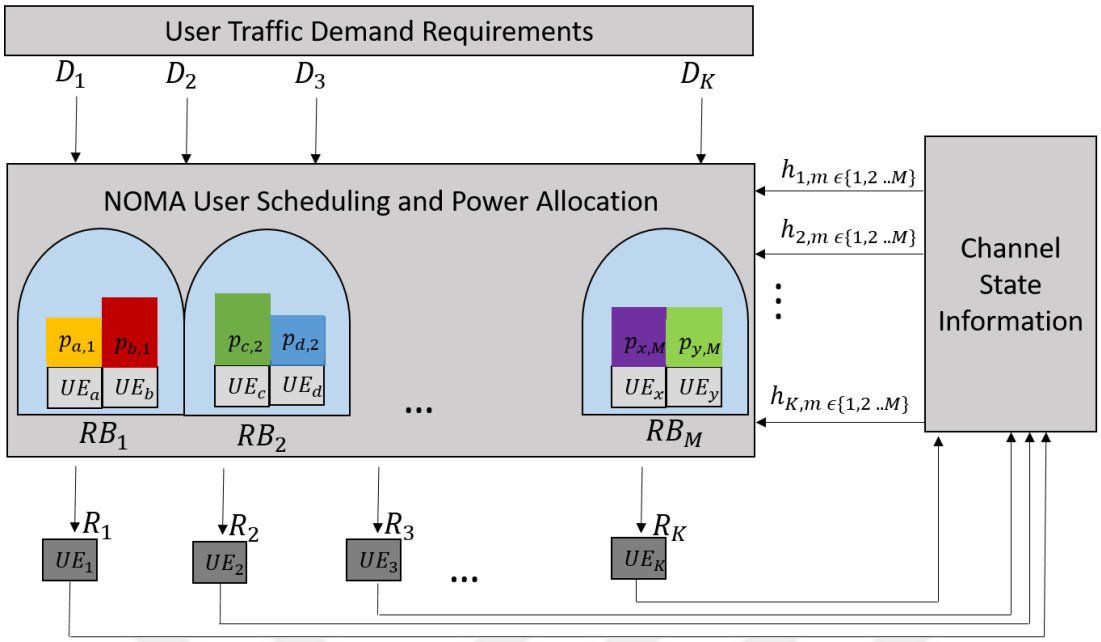


Figure 4.1 : User demand based NOMA resource allocation concept.

Assume that the channel qualities of the users are arranged in descending order. In this case, the kth user will remove the signals of the users having k+1 or higher index (i.e., the third term in equation (4.2) after the equality sign) using the SIC method.

Assuming that the channel information of all users are known at the base station, the role of the base station is to decide the allocation of users to radio resources and their transmission power levels. The detailed expressions of the NOMA concept is proposed in [111].

Figure 4.1 shows the concept of user demand based NOMA resource allocation scheme at the base station. The system model is designed for a single cell NOMA downlink system with M resource blocks and K user terminals. Each resource block $m \in \{1,2...M\}$ is composed of one subcarrier at the frequency domain and one time slot at the time domain. The proposed NOMA resource allocation scheme takes the user traffic demand requirements (Ω_1, Ω_2, ..., Ω_K) and the channel state information ($h_{1,m}$, $h_{2,m}$, ..., $h_{K,m}$, where $m \in \{1,2...M\}$) into account during the allocation of resources to users such as user grouping and user power assignment are jointly considered to optimize the underlying objective. Assume that the channel state information of UEs are perfectly known at the base station.

Assuming that perfect SIC is employed at the user terminals, the throughput of kth user is represented as R_k:

$$\begin{aligned} R_k &= \sum_{m=1}^{M} R_{km}, \\ s.t. \quad R_{km} &= B \times log_2(1 + \frac{|h_{km}|^2 p_{km}}{|h_{km}|^2 \sum_{i=1}^{k-1} p_{im} + W_{0,k}}), \\ R_k &\leq \Omega_k, \quad and \quad \sum_{i=1}^{K} p_{im} = 1 \ . \end{aligned} \tag{4.3}$$

B represents the bandwidth of a resource block and the maximum throughput, the channel gain and the amount of power allocated for the user k at the resource block m are represented as R_{km}, h_{km}, and p_{km}, respectively. Note that kth user throughput R_k can not be greater than kth user traffic demand Ω_k.

4.2 User Demand Based Resource Allocation for NOMA Systems

In this section, we present two PF based user scheduling and power allocation algorithms, namely UDB-PF and PUSF, for the NOMA downlink systems to decide which user(s) shall be assigned to a particular resource block together with power allocation coefficients. These algorithms provide resource allocations based on rate limited user traffic and channel state information in each time frame. We also propose the GA based heuristics to speed up UDB-PF and PUSF for the NOMA user scheduling.

4.2.1 User Demand-Based Proportional Fairness (UDB-PF)

The UDB-PF user scheduling is based on the well-studied proportional-fairness (PF) algorithm [5] which provides a good compromise between user fairness and total throughput (sum-rate) [114]. The algorithm provides the optimum power allocation under user rate constraints while user user groping approach is similar to other PF based solutions apart from the rate constraint. It essentially aims to maximize the product of average user throughputs calculated over a time window by determining the user groups and their power level assignments:

$$\begin{aligned} &\arg\max_{I} \prod_{k=1}^{K} R_k^c[n] \\ &s.t. \quad R_k[n] \leq \Omega_k[n] \end{aligned} \tag{4.4}$$

I is the set of users scheduled at time frame n. $\Omega_k[n]$ and $R_k[n]$ are the kth user traffic demand and throughput at the time frame n, respectively. R_k in the ergodic capacity form without time frame is given in equation (4.3) including the power constraint. $R_k^c[n]$ is the average throughput of user k over a time window, which is characterized by the time constant t_c, and is defined as:

$$R_k^c[n] = \begin{cases} \frac{t_c-1}{t_c} R_k^c[n-1] + \frac{1}{t_c} R_k[n], & k \in I \\ \frac{t_c-1}{t_c} R_k^c[n-1], & k \notin I \end{cases} \tag{4.5}$$

$R_k[n]$ is the kth throughput (data rate) at time frame n, and calculated using equation (4.3) at this time instant. Within the context of this chapter, the term time window should not be confused with time frame, such as a time window is usually one or a few hundred times larger than a time frame. Note that the solution to equation (4.4) can be obtained by solving the following optimization problem [115]:

$$\begin{aligned} I_{opt} = &\arg\max_{I} \sum_{k \in I} R_k[n]/R_k^c[n] \\ &s.t. \quad R_k[n] \leq \Omega_k[n] \end{aligned} \tag{4.6}$$

In the brute-force approach, for all resource blocks, the user set that yields the maximum PF metric is selected among all user allocation setups (i.e., $M \times \left(\binom{K}{1} + ... + \binom{K}{N_{max}}\right)$ different possibilities) as shown in equation (4.6). K is the total number of users, and N_{max} is the maximum number of users that can be assigned to a single resource block. The power allocation between the selected users can either be fixed for all combinations [5], or selected optimally to maximize the PF metric for a given user set [40].

In our proposed method, the scheduling algorithm is the same as the PF-based methods, however, the power allocation method in [40] is slightly modified to take non-full buffer traffic (user traffic demands) into consideration. In [40], assuming N_{max}=2, the optimum power allocation coefficients are calculated using Karush-Kuhn-Tucker (KKT) conditions:

$$p_{xm} = (w_{ym}g_{ym} - w_{xm}g_{xm})/(g_{xm}g_{ym}(w_{xm} - w_{ym})), \qquad p_{ym} = 1 - p_{xm} \tag{4.7}$$

where users x and y are grouped for the current resource block m, x is the near user while y is the far user. w_{xm} and w_{ym} are the weights associated with users x and y, and they are given by $w_{xm} = 1/R_x^c[n]$ and $w_{ym} = 1/R_y^c[n]$. $g_{xm} = |h_{xm}|^2/W_{0,x}$ and $g_{ym} = |h_{ym}|^2/W_{0,y}$ are the normalized channel gains for users x and y. When the first-order derivative of the PF-metric calculated at both p_{xm}=0 and p_{xm}=1 are negative, then p_{xm} is set to 0 and p_{ym} is set to 1. Similarly, when the first-order derivative of the PF-metric calculated at both p_{xm}=0 and p_{xm}=1 are positive, then p_{xm} is set to 1 and p_{ym} is set to 0 [40].

In our proposed power allocation method, first, the optimum power allocation factors are calculated as described above. The optimum power level of a user provides higher rate than its rate constraint, the excessive power can be reallocated to the other users in the same NOMA group. Now, let define kth user traffic demand for the current time frame as $\Omega_k[n]$ and the total rate that kth user has been obtained until the current resource block within the current time frame as $R_k^{prev}[n]$. Within the time frame, user throughput can not be greater than this users's traffic demand, so the following constraints for users x and y should hold:

$$R_x^{prev}[n] + R_{xm} \leq \Omega_x[n], \qquad R_y^{prev}[n] + R_{ym} \leq \Omega_y[n] \tag{4.8}$$

Inserting these conditions into the optimization problem using Lagrange multipliers complicates the calculations, so we propose a heuristic such that after optimally calculating the power allocation coefficient for user x, if its total throughput for the current time frame exceeds its demand, then the power level for user x is reduced to the required level and the excessive power is given to user y. In this case, the reduced power level for user x and the corresponding power level for user y can be found using the following formulation:

$$R_{xm} = log(1 + p_{xm}g_{xm}) = \Omega_x[n] - R_x^{prev}[n] \implies p_{xm} = \frac{2^{\Omega_x[n] - R_x^{prev}[n]} - 1}{g_{xm}}, \quad p_{ym} = 1 - p_{xm} \tag{4.9}$$

Similarly, if the total throughput for the time frame exceeds the demand for user y, then its power level is reduced and the excessive power is given to user x. The reduced

Table 4.1 : Summary of UDB-PF algorithm.

For each time frame,
▷ Given: Channel gains for user k at each resource block m and noise power $|h_{km}|^2, W_{0,k}$. For a user, it can vary between time frames as well as between resource blocks.
▷ Given: User demands for the current time frame (Ω_k). For a user, it can vary between time frames.
▷ Update average throughputs as if no one has received a resource in the current time frame: $R_k^c[n] = \frac{t_c-1}{t_c} R_k^c[n-1]$.
▷ Set $R_k^{prev}[n] = 0$ for all users.
▷ For each resource block m within the time frame,
- **User group selection:**
  - For each possible user combination (x, y) such that $x, y \in 1, \ldots, K$,
    - Compute the power allocation factors p_{xm} and p_{ym} using the method described above, and compute the instantaneous rates R_{xm} and R_{ym}.
    - Compute and store the PF index: $PF(x,y) = \frac{R_{xm}}{R_x^c[n]} + \frac{R_{ym}}{R_y^c[n]}$.
  - Choose the user pair (x,y) that has the maximum PF index.
- For the chosen pair, update the $R_k^{prev}[n]$ values as $R_x^{prev}[n] = R_x^{prev}[n] + R_{xm}$ and $R_y^{prev}[n] = R_y^{prev}[n] + R_{ym}$.
- If $R_x^{prev}[n] = \Omega_x[n]$, then remove user x from the search space. Do the same for user y.
- For the chosen pair, update the average throughputs by $R_x^c[n] = R_x^c[n] + R_{xm}/t_c$ and $R_y^c[n] = R_y^c[n] + R_{ym}/t_c$.
- Proceed to the next resource block.

▷ Proceed to the next time frame.

power level for user y and the corresponding level for user x can be found using the following formulation:

$$\begin{aligned} R_{ym} &= log(\frac{(1+g_{ym})}{(1+p_x g_{ym})}) = log(\frac{(1+g_{ym})}{(1+(1-p_{ym})g_{ym})}) = \Omega_y[n] - R_y^{prev}[n] \implies \\ p_{ym} &= \frac{1+g_{ym}}{g_{ym}}(1 - \frac{1}{2^{(\Omega_y[n]-R_y^{prev}[n])}}), \qquad p_{xm} = 1 - p_{ym} \end{aligned} \tag{4.10}$$

Based on the above discussion, the summary of the user-demand based proportional fairness (UDB-PF) algorithm for $N_{max} = 2$ is given in Table 4.1. The extension of the method to more than two users per resource block ($N_{max} > 2$) is straightforward and is not shown here.

4.2.2 Proportional User-Satisfaction Fairness (PUSF)

The PF-based user scheduling method maximizes the product of average user throughputs and it assumes that each user aims to get as much data as possible, which corresponds to the full-buffer case. In reality, each user demands a limited data rate and there is no point allocating more resources whose rate is higher than this user's traffic demand. In the UDB-PF method, which is described above, the PF-based algorithm is modified to reflect the limited nature of user demands. In this sub-section, we propose a new method, namely the Proportional User-Satisfaction Fairness (PUSF) algorithm, that aims to maximize the average user satisfaction.

The user satisfaction can be defined as the ratio of a user's downlink traffic rate to its demand in a time frame [116]. It is a number between zero and one. It can not be larger than 1, because a user can not get more data rate than its demand. Mathematically, it can be stated as follows:

$$\Upsilon_k[n] = \begin{cases} \frac{R_k[n]}{\Omega_k[n]}, & R_k[n] \leq \Omega_k[n] \\ 1, & \text{otherwise} \end{cases} \tag{4.11}$$

$R_k[n]$ is the total data rate that user k receives at time frame n. The ergodic capacity form without time frame n (R_k) is given in equation(4.3) including the power constraint, and $\Omega_k[n]$ represents the demand of the user k for that time frame. Assume that resource blocks are processed within a time frame one-by-one, then for a particular resource block m, $R_k[n]$ can be defined as $R_k[n] = R_k^{prev}[n] + R_{km}$, where $R_k^{prev}[n]$ is the total data rate that the user has been obtained from $m = 1$ to the current resource block within the current time frame, and R_{km} is the calculated throughput for the resource block m.

In our method, we aim to maximize the product of average satisfaction of the users that are calculated over a time window similar to the PF-based methods. Maximizing the product penalizes the cases when a user gets an extremely low satisfaction value, hence it provides fairness among users. Average satisfaction of a user can be defined as:

$$\Upsilon_k^c[n] = \begin{cases} \left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1] + \frac{1}{t_c}\Upsilon_k[n], & k \in I \\ \left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1], & k \notin I \end{cases} \tag{4.12}$$

I is the set of users chosen for the current time frame, and t_c is the constant that specifies the length of the averaging window. Similar to the PF-based methods, for a particular resource block, the underlying optimization problem can be stated as follows:

$$\begin{aligned} I_{opt} &= \arg\max_{I} \left\{ \prod_{k=1}^{K} \Upsilon_k^c[n] \right\} \\ &= \arg\max_{I} \left\{ log\left(\prod_{k=1}^{K} \Upsilon_k^c[n] \right) \right\} \\ &= \arg\max_{I} \left\{ \sum_{k=1}^{K} log\left(\Upsilon_k^c[n]\right) \right\} \end{aligned} \tag{4.13}$$

In equation (4.13), the fact that taking the logarithm of the objective function does not change its optimum point is utilized. If we define the objective function as $J = \sum_{k=1}^{K} log\left(\Upsilon_k^c[n]\right)$ and use equation (4.12):

$$\begin{aligned} J &= \sum_{k=1}^{K} log\left(\Upsilon_k^c[n]\right) = \sum_{k\notin I} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]\right] + \sum_{k\in I} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1] + \frac{1}{t_c}\Upsilon_k[n]\right] \\ &= \sum_{k\notin I} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]\right] + \sum_{k\in I} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]\left(1+\frac{\Upsilon_k[n]}{(t_c-1)\Upsilon_k^c[n-1]}\right)\right] \\ &= \sum_{k\notin I} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]\right] + \sum_{k\in I} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]\right] + \sum_{k\in I} log\left[\left(1+\frac{\Upsilon_k[n]}{(t_c-1)\Upsilon_k^c[n-1]}\right)\right] \\ &= \sum_{k=1}^{K} log\left[\left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]\right] + \sum_{k\in I} log\left[\left(1+\frac{\Upsilon_k[n]}{(t_c-1)\Upsilon_k^c[n-1]}\right)\right] . \end{aligned} \tag{4.14}$$

The first part of the expression does not depend on the scheduling set I, and we can omit it in the optimization process. Also, because t_c is large, the term $\frac{\Upsilon_k[n]}{(t_c-1)\Upsilon_k^c[n-1]}$ is most probably small. Using the property that for small x, $log(1+x) \approx x$, the objective function becomes:

$$J = \sum_{k\in I} \frac{\Upsilon_k[n]}{(t_c-1)\Upsilon_k^c[n-1]} . \tag{4.15}$$

By inserting equation (4.11) into equation (4.15) and by omitting the constant term (t_c-1), we have

$$\begin{aligned} J &= \sum_{k\in I} \frac{R_k^{prev}[n] + R_{km}}{\Omega_k[n]\Upsilon_k^c[n-1]} \\ &= \sum_{k\in I} \frac{R_k^{prev}[n]}{\Omega_k[n]\Upsilon_k^c[n-1]} + \sum_{k\in I} \frac{R_{km}}{\Omega_k[n]\Upsilon_k^c[n-1]} \end{aligned} \tag{4.16}$$

Again, the first term is independent of the selection of I, and the corresponding optimization problem can be stated as follows:

$$I_{opt} = \arg\max_{I} \sum_{k \in I} \frac{R_{km}}{\Omega_k[n]\Upsilon_k^c[n-1]} \qquad s.t. \quad R_k[n] \leq \Omega_k[n] \tag{4.17}$$

Table 4.2 : Summary of PUSF algorithm.

For each time frame,
▷ Given: Channel gains for user k at each resource block m and noise power $|h_{km}|^2, W_{0,k}$. It can vary between time frames as well as between resource blocks.
▷ Given: User demands for the current time frame (Ω_k). For a user, it can vary between time frames.
▷ Update average satisfactions as if no one has received a resource in the current time frame: $\Upsilon_k^c[n] = \left(\frac{t_c-1}{t_c}\right)\Upsilon_k^c[n-1]$.
▷ Set $R_k^{prev}[n] = 0$ for all users.
▷ For each resource block m within the time frame,
- **User group selection:**
  - ◦ For each possible user combination (x,y) such that $x,y \in 1,\ldots,K$,
    - · Compute the power allocation factors p_{xm} and p_{ym} using the method described above, and compute the instantaneous rates R_{xm} and R_{ym}.
    - · Compute and store the PUSF index: $PUSF(x,y) = \frac{R_{xm}}{\Omega_x[n]\Upsilon_x^c[n]} + \frac{R_{ym}}{\Omega_y[n]\Upsilon_y^c[n]}$.
  - ◦ Choose the user pair (x,y) that has the maximum PUSF index.
- For the chosen pair, update the R_k^{prev} values as $R_x^{prev}[n] = R_x^{prev}[n] + R_{xm}$ and $R_y^{prev}[n] = R_y^{prev}[n] + R_{ym}$.
- If user x has reached the maximum satisfaction level of 1, i.e. $R_x^{prev}[n] = \Omega_x$[n], then remove user x from the search space. Do the same for user y.
- For the chosen pair, update the average user satisfaction values by $\Upsilon_x^c[n] = \Upsilon_x^c[n] + (1/t_c)R_x[n]/\Omega_x[n]$ and $\Upsilon_y^c[n] = \Upsilon_y^c[n] + (1/t_c)R_y[n]/\Omega_y[n]$.
- Proceed to the next resource block.

▷ Proceed to the next time frame.

Note that the problem turned out to be the same as PF with $R_k^c[n]$ replaced by $\Omega_k[n]\Upsilon_k^c[n-1]$. Hence, a similar brute-force approach can be used to solve the problem. Also note that the non-linearity in the definition of $\Upsilon_k[n]$, i.e. limiting the total throughput of a user to its demand within a time frame, is ignored when transforming equation (4.15) into equation (4.16). This non-linearity is handled in the power allocation step just like the UDP-PF method. Because the formulation is pretty

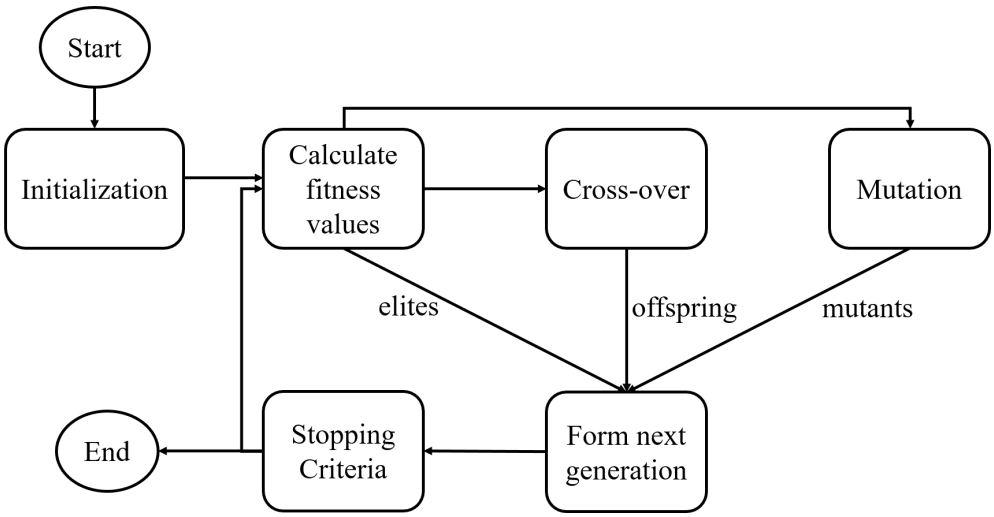


Figure 4.2 : Genetic Algorithm flow chart.

much alike, the optimum power allocation factors can be found by using equation (4.7) and the discussion below equation (4.7) with the only difference in the definitions of w_x and w_y such that $w_x = 1/(\Omega_x[n]\Upsilon_x^c[n-1])$ and $w_y = 1/(\Omega_y[n]\Upsilon_y^c[n-1])$. The case when one of the users gets more rate than its demand is handled just like in the UDB-PF method. As in the UDB-PF approach, the PUSF can also reallocate the excessive power to the other users in the same NOMA group. Based on the above formulation and discussion, the summary of the PUSF algorithm is given in Table 4.2.

4.2.3 Genetic Algorithm approach for user group selection

The optimum user group selection at each resource block can be achieved by evaluating all possible user group combinations with the exhaustive search. This type of brute force method significantly increases the computational complexity when the number of users is high. GA is a powerful heuristic for exploring prohibitively huge search spaces. In this section, we present GA based search heuristics for both UDB-PF and PUSF algorithm to reduce the computational complexity.

In the previous chapter, the GA approach is used for NOMA downlink resource allocation, where, the user group selection for all resource blocks are determined at the same time. Therefore, the size of the search space increases not only with the number of users but also with the number of resource blocks. However, in this chapter, UDB-PF and PUSF algorithms perform user group selection per resource block and exhaustive search is applied within each resource block. Instead of the exhaustive

Table 4.3 : User group selection with GA in UDB-PF algorithm.

- **User group selection with GA:**
  - Employ Genetic Algorithm to find best user (x,y) such that $x,y \in 1,\ldots,K$, In the GA fitness function:
    - Compute the power allocation factors p_{xm} and p_{ym} using the method described above, and compute the instantaneous rates R_{xm} and R_{ym}.
    - Calculate minus PF index as a fitness value:
      $PF(x,y) = \frac{R_{xm}}{R_x^c[n]} + \frac{R_{ym}}{R_y^c[n]}$.
  - GA selects the best user pair (x,y) among the possible solutions.

Table 4.4 : User group selection with GA in PUSF algorithm.

- **User group selection with GA:**
  - Employ Genetic Algorithm to find best user (x,y) such that $x,y \in 1,\ldots,K$, In the GA fitness function:
    - Compute the power allocation factors p_{xm} and p_{ym} using the method described above, and compute the instantaneous rates R_{xm} and R_{ym}.
    - Calculate minus PUSF index as a fitness value:
      $PUSF(x,y) = \frac{R_{xm}}{\Omega_x[n]\Upsilon_x^c[n]} + \frac{R_{ym}}{\Omega_y[n]\Upsilon_y^c[n]}$.
  - GA selects the best user pair (x,y) among the possible solutions.

search for the user group selection in Tables 4.1 and 4.2, the GA is applied within each resource block to select the user group, where the computational complexity is significantly decreased. The pseudo codes of the the user group selection using the GA approach are shown in Tables 4.3 and 4.4 for UDB-PF and PUSF, respectively.

Figure 4.2 presents the flow chart of the proposed GA approach. First, a certain number of individuals corresponding to the candidate solutions in the population is randomly created. Each individual (i.e., chromosome of the GA) corresponds to one of possible user group solutions (e.g., [UE1 UE2] in Figure 4.3). The fitness of each individual is evaluated using a fitness function such that a higher fitness value indicates a preferable user group. Note that two different fitness functions are used for UDB-PF and PUSF algorithms. The individuals with higher fitness values are selected as elites and directly passed to the next generation. To provide the diversity in the population, the remaining individuals are divided into two groups for cross-over and mutation operations such that new offspring and mutants are created for the next generation. The number of individuals for the cross-over operation is determined by the cross-over rate and the rest

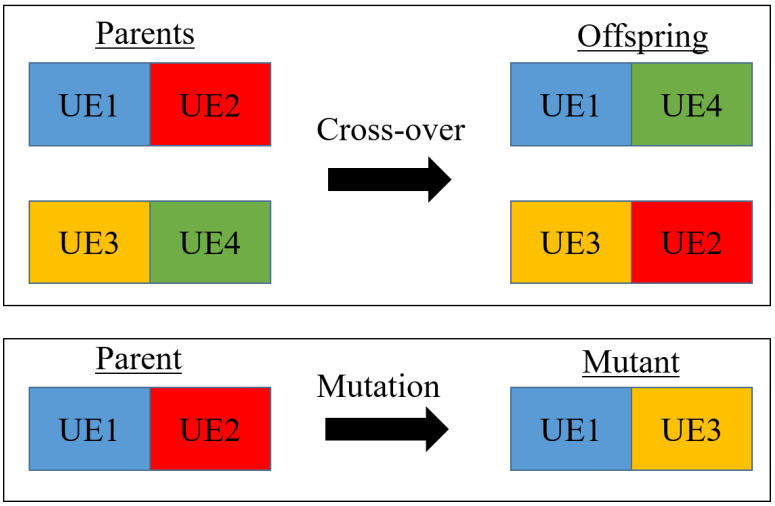


Figure 4.3 : An example of cross-over and mutation operations.

of the individuals processed by the mutation operation. The same procedure continues until the stopping criteria is met.

An example of the cross-over and mutation operations are depicted in Figure 4.3 where two users are allocated for each resource block in NOMA downlink system. The offspring ([UE1 UE4] and [UE3 UE2]) are generated by combining the elements of a pair of parents ([UE1 UE2] and [UE3 UE4]) with the use of cross-over process. In another diversity operation, a random change is applied to the second element of the parent([UE1 UE2]) with the mutation process, and hence a new mutant ([UE1 UE3]) is created.

4.2.4 Complexity analysis of the proposed algorithms

The best user group selection at one resource block can be provided by evaluating all user group combinations with the exhaustive search in UDB-PF and PUSF algorithms. Let the number of users and the maximum number of users per resource block are K and N_{max}, respectively. The number of all possible user sets is $\binom{K}{1}+\binom{K}{2}+\ldots+\binom{K}{N_{max}}$. For $K \gg N_{max}$, $\binom{K}{N_{max}}$ determines the asymptotic behavior of the number of all possible user sets. Since the evaluation function will run for $\binom{K}{N_{max}} = \frac{K\times(K-1)\times\ldots(K-N_{max}+1)}{1\times2\times\ldots N_{max}}$. Therefore, the complexity order can be stated as $O(K^{N_{max}})$.

In the GA approach, the computational complexity mainly depends on the population size (initial solutions), the length of the chromosome and the selection function. The formal complexity analysis of the GA is beyond the scope of this chapter. However,

we will discuss the approximate computational load of the GA approach to indicate the improvement over the exhaustive search. Let the population size, the length of the chromosome, and the selection process be L, N_{max}, and stochastic uniform selection, respectively. At each generation (iteration), the fitness function for candidate solutions is calculated for a population size (L). Therefore, the GA algorithm totally evaluates ($G \times L$) candidate solutions, where G represents the maximum number of generations. In this chapter, G is defined as $G = max(10, K)$, where K represents the number of users. Since the total number of candidate solutions explored is ($G \times L = K \times L$) and L is the constant, the complexity order of the GA approach can be simplified as $O(K)$. Comparing to the complexity order of the exhaustive search ($O(K^{N_{max}})$), the GA heuristics can provide significant reduction on the computational complexity. Note that the GA approach does not guarantee for reaching the global optimal solution since all the solution space may not be explored. However, the results in this chapter show that the throughput and user satisfaction performance of the GA heuristics are comparable to the exhaustive search.

4.3 Simulation Results

The performance of the proposed UDB-PF, PUSF and their GA versions UDB-PF-GA and PUSF-GA methods is evaluated and compared with the power optimized proportional fair (PF) method [40] in various network settings using the MATLAB simulation tool. In the simulation experiments, the users are uniformly distributed over a cell with 1 km radius. A base station located at the center of the cell allocates radio resources to the users by employing NOMA as the multiple access scheme. The resource block is the minimum unit that can be allocated to users and defined as one subcarrier at the frequency and one time slot. The maximum number of users per resource block in the NOMA is set to 2 for all experiments. The parameters used for the simulations are given in Table 4.5. The results are obtained as the average of 1000 experiments for each network setting.

MATLAB optimization toolbox is used to employ the GA heuristics proposed in this chapter. The number of individuals, the number of elites, and the cross-over fraction are set to 10, 2, and 0.8, respectively. Therefore, $(10-2)\times 0.8 \approx 6$ offspring are created by the cross-over function and $10-2-6=2$ mutants are generated by the mutation

function at each iteration. The stopping criteria is selected as the maximum number of generations (G) and it is configured as $G = max(10, K)$ where K is the number of users in the network.

Table 4.5 : Simulation parameters for PF based resource allocation.

Parameter	Value
Subcarrier Bandwidth	180 KHz
Receive/Transmit antenna	SISO
Path Loss Exponent	3
Subcarrier Power	10 dBm
Receiver Noise Density	-169 dBm/Hz
Shadowing standard deviation	Lognormal with 8 dB
Fading Model	Rayleigh flat fading
Cell Radius	1 km
Max # of users per resource block	2
Time window size (t_c)	100 time frames

We report two performance metrics, namely the average sum-rate (network-wide throughput) and the average user satisfaction. We define the user satisfaction metric specifically for non-full buffer traffic scenario as the ratio of carried traffic rate to offered traffic rate (i.e., traffic demand) for a particular user. For time n, when $R_k[n]$ represents the kth user throughput value and there are K connected users at the cell, the sum-rate value is equal to $\sum_{k=1}^{K} R_k[n]$. The average sum-rate is calculated by taking the average of sum-rate values over the simulation time. Similarly, the user satisfaction metric $\Upsilon_k[n]$ can be calculated as $R_k[n]/\Omega_k[n]$, where $\Omega_k[n]$ represents the kth user traffic demand for time n. The average user satisfaction is calculated by taking the average of user satisfaction values over the simulation time.

4.3.1 Effects of traffic loadings and demand variations

In this section, the effects of traffic loadings and user demand variations over the system performance are reported. For the experiments in this section, the number of users and the number of resource blocks are set to constant values of 5 and 2, respectively, while each user traffic demand varies at each time frame according to the uniform distribution with a given mean (μ) and the standard deviation (σ). An example of uniformly distributed user traffic demand is shown in Figure 4.4 when the

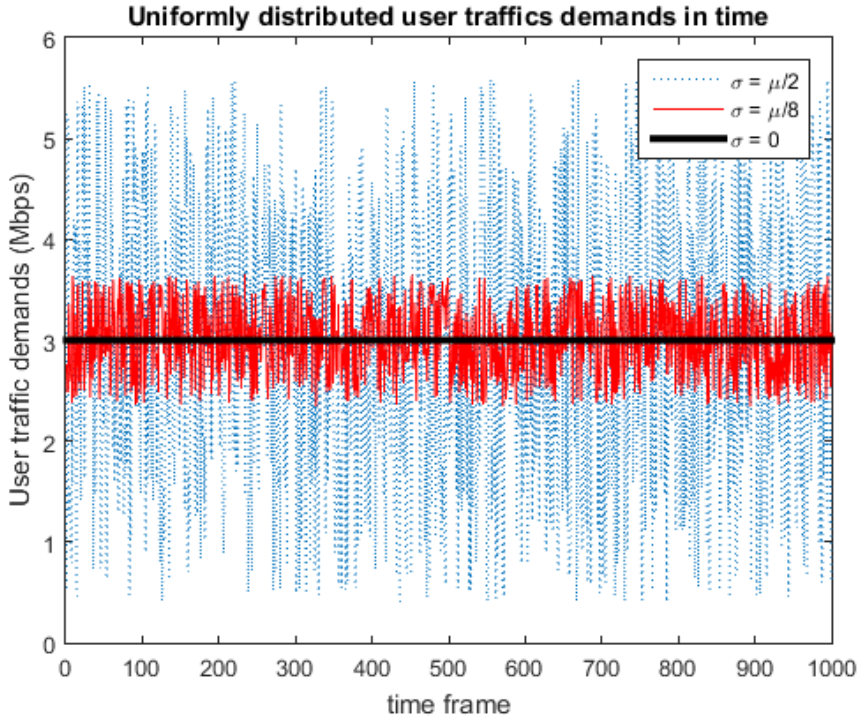


Figure 4.4 : Example uniform user traffic demand distributions.

mean value μ is set to 3 Mbps and standard deviations are set to $\sigma = \mu/2$, $\sigma = \mu/8$, and $\sigma = 0$.

Figure 4.5 illustrates the average sum-rate with respect to the mean value of uniformly distributed user traffic demands using the standard deviation σ of 0, $\mu/8$, and $\mu/2$. The results show that the average sum-rate increases until it reaches to the maximum sum-rate as the user traffic demand increases for all cases. Note that the user traffic demand varies from 0.36 Mbps to 4.32 Mbps and there are only two subcarriers within 180 kHz bandwidth. As observed in the figure, the UDB-PF method provides the highest average sum-rate for all user traffic demands. The sum-rate performance of the PUSF method decreases as the user traffic demand variations increase from 0 to $\mu/2$ since it focuses on maximizing the average user satisfaction. The PF algorithm always yields lower sum-rate compared to UDB-PF and PUSF since it does not consider the user traffic demands in its calculations. We observed higher traffic drops (lower sum-rate) even under low traffic demands in the PF method because it allocates resource blocks to users by considering the full buffer traffic model and may not use the entire resource block effectively if the total user traffic demands are lower than the capacity of the allocated resource blocks.

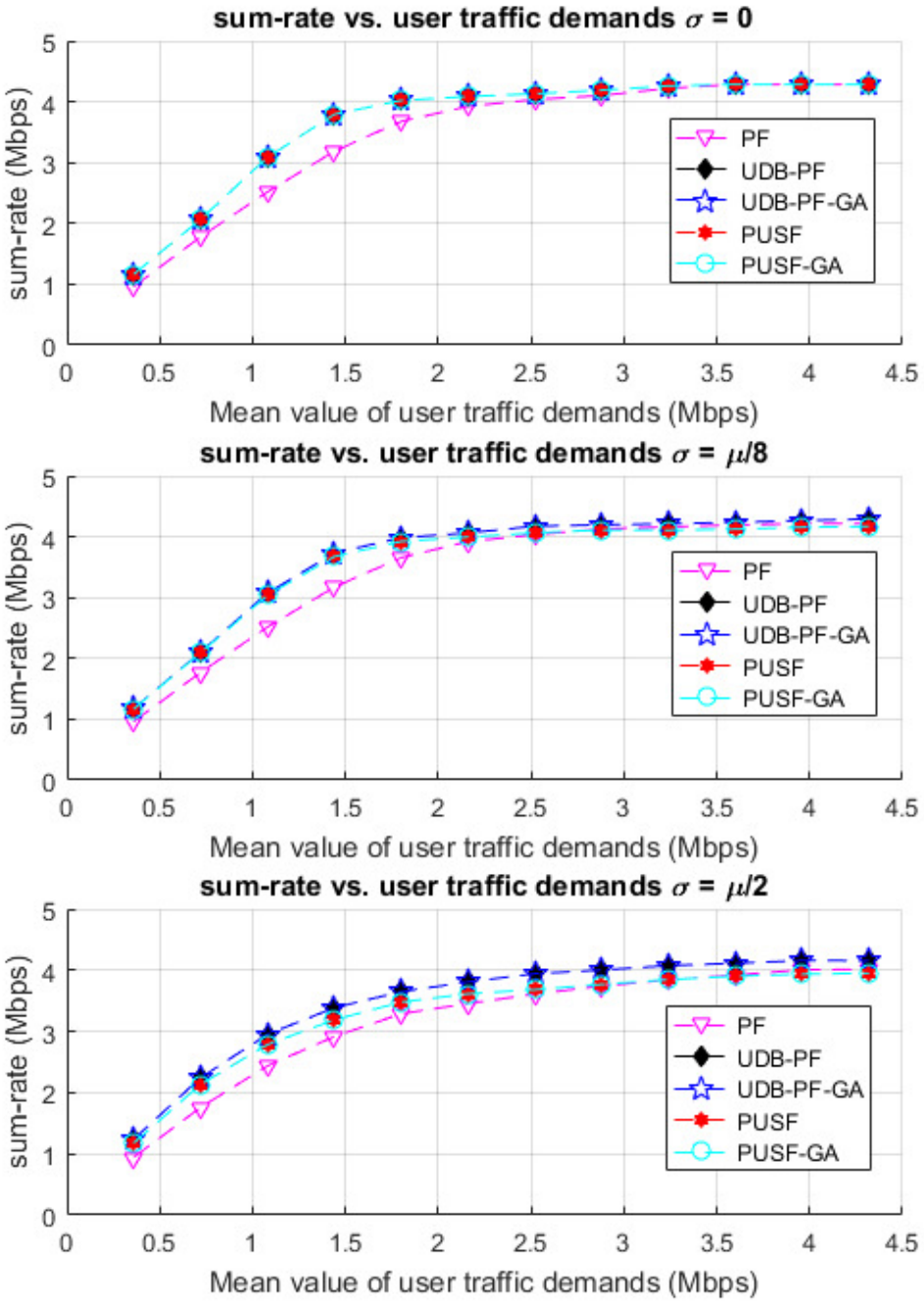


Figure 4.5 : Average sum-rate under various traffic demands.

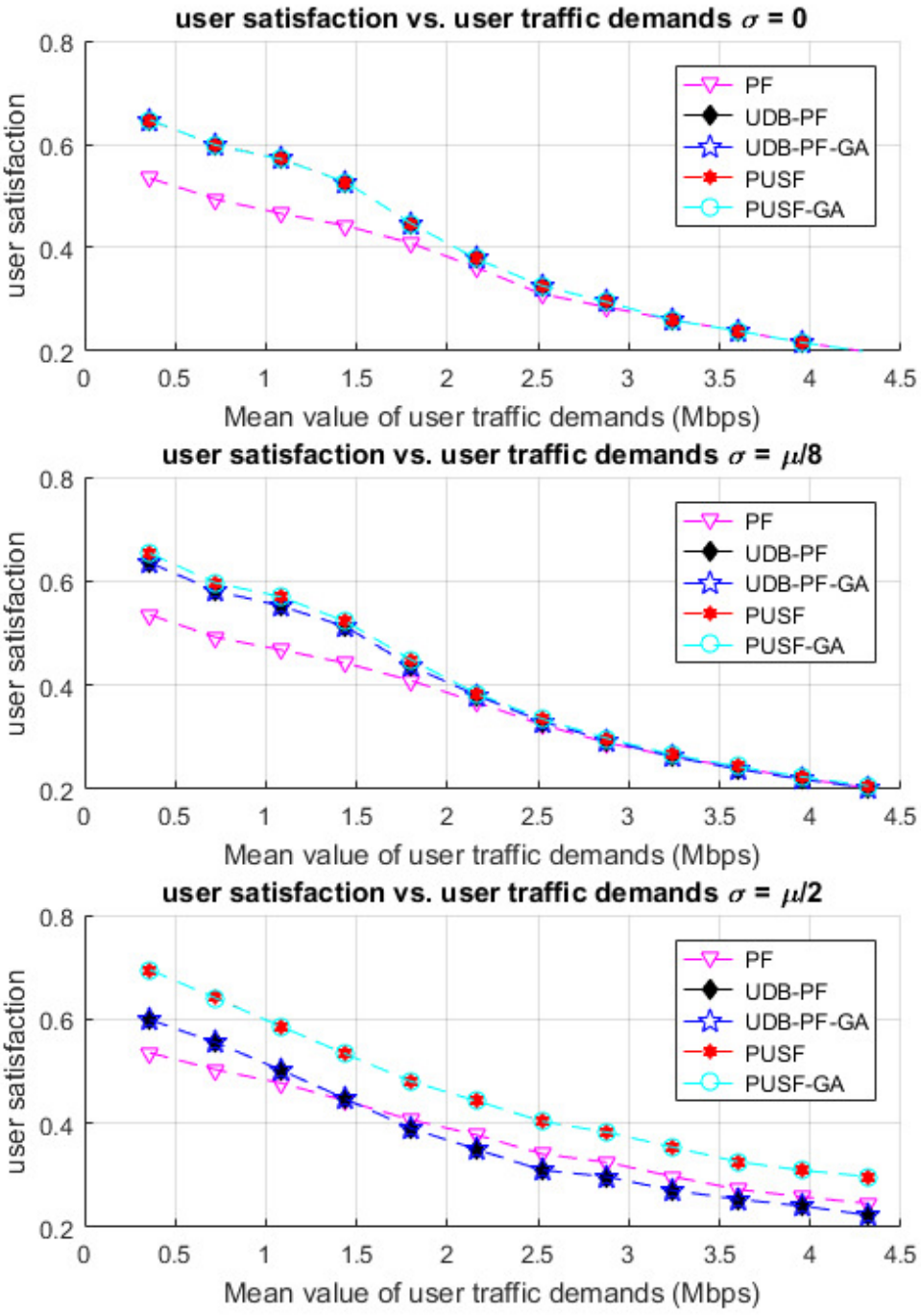


Figure 4.6 : Average satisfaction under various traffic demands.

Figure 4.6 shows the average user satisfaction with respect to the mean value of uniformly distributed user traffic demands using the standard deviation σ of 0, $\mu/8$, and $\mu/2$. The user satisfaction decreases with the increase of user traffic demands since there are limited radio resources. As observed from the figure, the PUSF method ensures the highest user satisfaction for all user traffic demands. In addition, the performance of PUSF increases as the user traffic demand variation increases from 0 to $\mu/2$.

It is an important result that when the user traffic demand remains constant over time (i.e., $\sigma = 0$), the UDB-PF and PUSF methods yield the same performance results in terms of both sum-rate and user satisfaction. The reason for this result is that the objective function will converge to the same function for both UDB-PF and PUSF when the user traffic demand remains constant over time.

Another significant inference is that when the user traffic demands are significantly high and the standard deviation is zero (emulating the full buffer traffic condition), all three methods (PF, UDB-PF, and PUSF) converge to the same performance results in terms of both sum-rate and user satisfaction. The reason for this result is that the UDB-PF algorithm ignores the user traffic requirements under the full buffer traffic condition, and hence it will be the same as the PF method. Similarly, the PUSF method converges to the PF method since the PUSF will be the same as the UDB-PF method when user traffic demands go to infinity (i.e., the full buffer traffic condition) according to equation (4.17). We also provide the results of the GA heuristics, namely UDB-PF-GA and PUSF-GA, in Figures 4.5 and 4.6. Since the number of users K is set to a lower value of 5, the sum-rate and the satisfaction results of the GA heuristics are the same as the UDB-PF and PUSF algorithms.

4.3.2 Effect of the number of users

In the second part of the simulation results, the experiments are performed to evaluate the proposed allocation schemes with respect to the number of connected users in a cell. While the number of users are varied from 5 to 45, the number of resource blocks are set to a constant value of 5. Each user traffic demand varies over time according to the uniform distribution with the mean of 5 Mbps and the standard deviation of 2.5 Mbps.

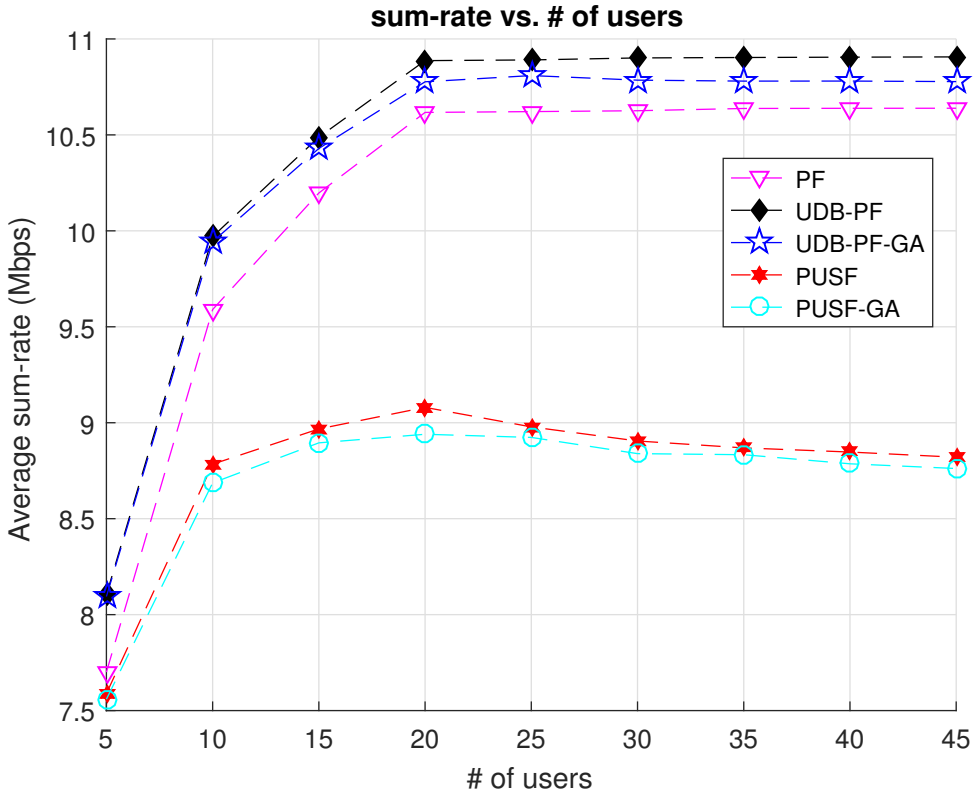


Figure 4.7 : Average throughput with respect to the number of users.

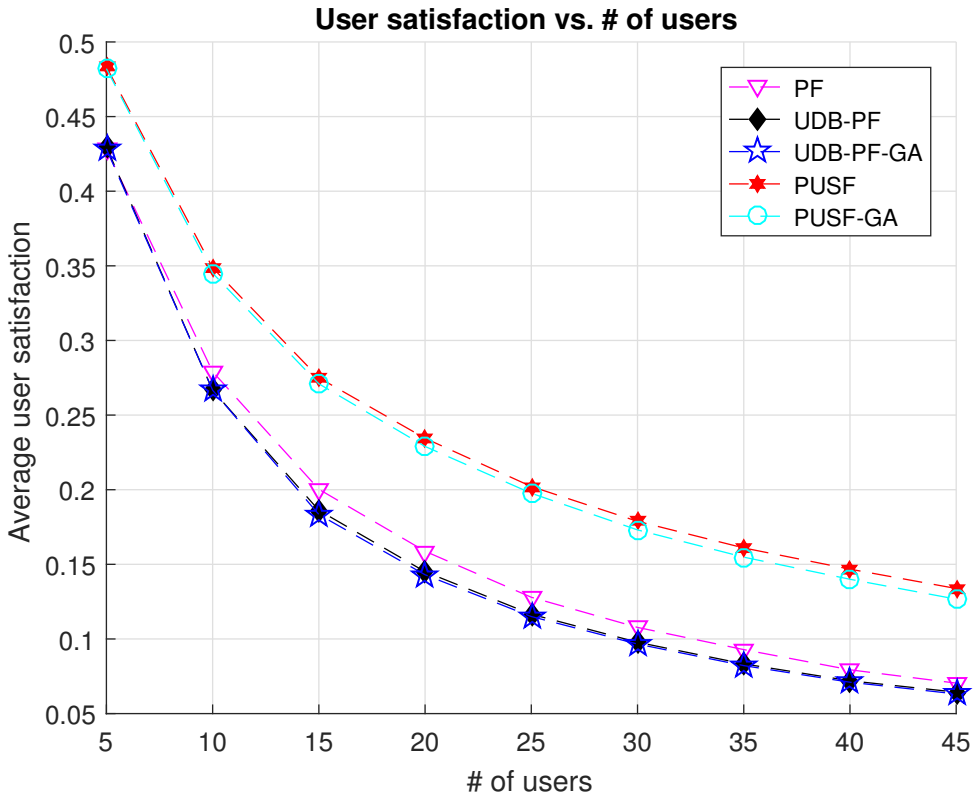


Figure 4.8 : Average user satisfactions with respect to the number of users.

Figure4.7 shows that as the number of users increases the UDB-PF method ensures the highest sum-rate since it maximizes the average user throughputs in a given time window by taking user demand requirements into account. The average sum-rates of all five methods increase as the number of users increases from 5 to 20. When the number of users is beyond 20, the average sum-rates of both PF and UDB-PF remain almost constant. However, the sum-rates of the PUSF and PUSF-GA methods slightly decrease as the number of users increases beyond 20 since it is harder to meet the user satisfaction objective when there are more users demanding the limited resources. A significant result from the figure that, the UDB-PF-GA and PUSF-GA algorithms provide almost the same performance with the the UDB-PF and PUSF algorithms while the number of users is set to 5 and 10. Beyond that, a slight performance degradation can be observed when the GA heuristics are employed. However, the GA approach significantly reduces the computational complexity (see Figure 4.9 for the details). The similar result can be observed from Figure 4.8.

The average user satisfaction with respect to the number of connected users in a cell is illustrated in Figure 4.8. The PUSF method ensures the highest user satisfaction for all number of user settings. When the number of users increases, the average user satisfaction decreases for all five methods as expected because the limited resources are not sufficient to support the demand requirements of higher number of users.

Figure 4.9 shows the total number of candidate solutions explored by all five algorithms with respect to the number of users in a cell. Note that the number of resource blocks in the cell is set to 5 for all experiments and the results in the figure correspond to all resource blocks. The results in this figure confirm the complexity analysis in Section 4.2.4 such that the computational complexity of the PF, UDB-PF, and PUSF methods exponentially increases with the number of users. However, the total number of evaluated candidate solutions in the GA heuristics linearly increase as given in the complexity analysis.

As depicted in Figure 4.9, the GA heuristics explore more candidate solutions when K is less than 20; but the computational complexity is not an issue since the number of candidate solutions explored are low for all algorithms. Beyond K is 20, the computational load is significantly lower for the GA heuristics compared to the exhaustive search based algorithms. When the number of users in a cell is 45, the

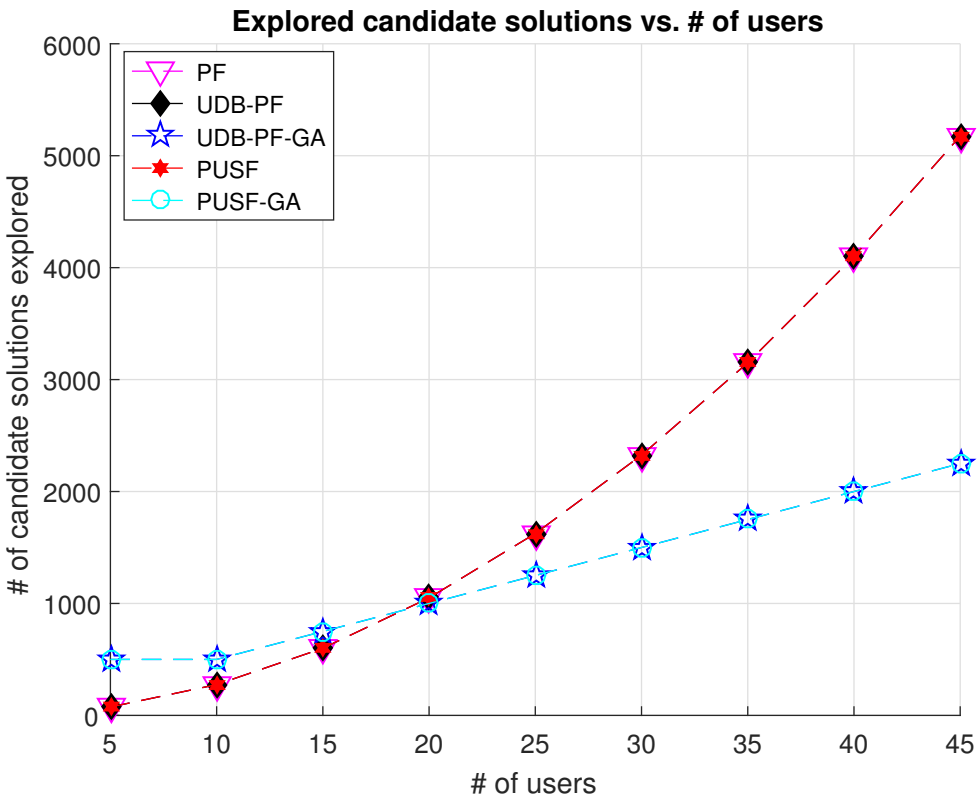


Figure 4.9 : The number of explored candidate solutions.

GA heuristics provide more than 50% improvement on the computational load while the performance degradation is only about 1.2%.

4.4 Summary

In this chapter, two user scheduling and power allocation schemes, namely UDB-PF and PUSF, are proposed for NOMA downlink systems under non-full buffer traffic. UDB-PF extends the PF based scheduling by allocating optimum power levels to satisfy user traffic demand constraints while PUSF maximizes the network-wide user satisfaction. In both schemes, the optimal power level assignment is calculated together with the best user pair selection at each resource block for a given objective function. The GA heuristic is also employed for user group selection at each resource block to reduce the computational complexity. The performance is evaluated by varying the number of users and traffic characteristics of each user. The simulation results show that UDB-PF yields higher sum-rate (throughput) while PUSF provides higher network-wide user satisfaction results compared to the PF based user scheduling. The performance gains of the proposed methods increase as the variation

of user traffic demands increases over time. In addition, when the number of users in the network gets higher, the GA heuristics provide the performance gain on the computational load while the throughput and user satisfaction results are only slightly degraded.

The performance evaluation on this chapter has been carried out using only simulation experiments rate limited traffic demands are considered instead of packet based traffic model with random inter-arrival times and packet sizes. In the next chapters, we will investigate the Poisson traffic arrivals for the downlink resource allocation and investigate the queuing dynamics through the analytical models. Furthermore, we will include the statistical channel state information (i.e., Rayleigh fading) instead of the assumption of perfect CSI at the base station.

5. ANALYTICAL MODEL FOR QUEUING DELAY OF NOMA DOWNLINK SYSTEMS

In the previous chapters, we have presented user grouping and power allocation strategies for downlink NOMA systems under both full and non-full buffer traffic models also published as [117–119]. The performance evaluation of these approaches has been carried out using only simulation experiments such that the CSI is perfectly known at the transmitter and rate limited traffic demands are considered instead of packet based traffic model with random inter-arrival times and packet sizes. In this chapter, we propose an analytical model to characterize the average queuing delay for NOMA downlink systems by utilizing a discrete time M/G/1 queuing model when the probability distribution of channel is known. The packet arrival process is assumed to be Poisson distributed while the departure process depends on network settings and resource allocation.

The complexity of 5G network is expected to be significantly higher due to its inherent support for billions of Internet of Things (IoTs) devices enabling new services with stringent delay and reliability requirements. Three broad categories of 5G services considered by 3GPP are enhanced mobile broadband (eMBB), ultra reliable low latency communication (URLLC), and massive machine-type communications (mMTC). While the eMBB and mMTC services focus on the capacity and scalability aspects of 5G, respectively, URLLC is critical for enabling remote control of time and mission-critical IoT services. Non-orthogonal multiple access (NOMA) is a promising technology for 5G systems due to its higher spectral efficiency potentially yielding lower latency and higher scalability results by allowing simultaneous transmission of multiple users at the same resource block. New analytical models which can characterize the latency dynamics of 5G are of paramount importance to develop high performance resource allocation strategies satisfying the challenging requirements of 5G services.

The latency contribution of the user plane end-to-end (E2E) delay of a packet transmission in 5G can be divided into three main parts: radio access, mobile core, and cloud. The radio access latency between a base station and user equipment includes over-the-air transmission and propagation, queuing, processing, and re-transmission delays [12]. A 5G new radio (5G NR) access technology is introduced with shorter frame duration and wider bandwidth to satisfy the lower latency requirements of URLLC services such as industrial control and automation, augmented and virtual reality, tactile Internet and intelligent transportation [10, 11]. NOMA, which has been studied in 3GPP Releases 13 and 14 and is under consideration at the standardization activities for 5G NR, can be another instrument to potentially decrease the radio access latency for URLLC services [26, 27]. The analytical model proposed in this chapter is directly utilized for quantifying the latency improvements of the 5G NR with the NOMA.

In [105], the authors utilize the stochastic network calculus approach to study the resource allocation problem for uplink NOMA systems by minimizing the delay violation probability. Stable throughput regions for uplink NOMA systems under unsaturated traffic are investigated using the queuing theory approach, where traffic arrival for each user is assumed to be independent Bernoulli process [106]. Similarly, the delay analysis of NOMA is studied using the queuing theory approach in this chapter; however, we focus on downlink channels. In [81], the power control policy for NOMA is studied to meet the delay objectives deriving the effective capacity formulation when the channel state information (CSI) is known. Their delay results are obtained only using the simulations while in our study we analytically expressed the average queuing delay for NOMA downlink systems when the probability distribution of the underlying channel is known.

In [108], a queuing analysis for discrete time discrete state D/G/1 queuing model is presented under a Rayleigh fading channel in the low SNR regime for OMA systems. In another study [109], a general state space Markov chain model is proposed to calculate the throughput regions of OFDMA users under a Rayleigh fading channel by taking the scheduling algorithms into account. The buffer overflow probability providing insights for buffer dimensioning problems is obtained assuming that each user has finite traffic arrival and queue capacity. In this chapter, we adopt a similar

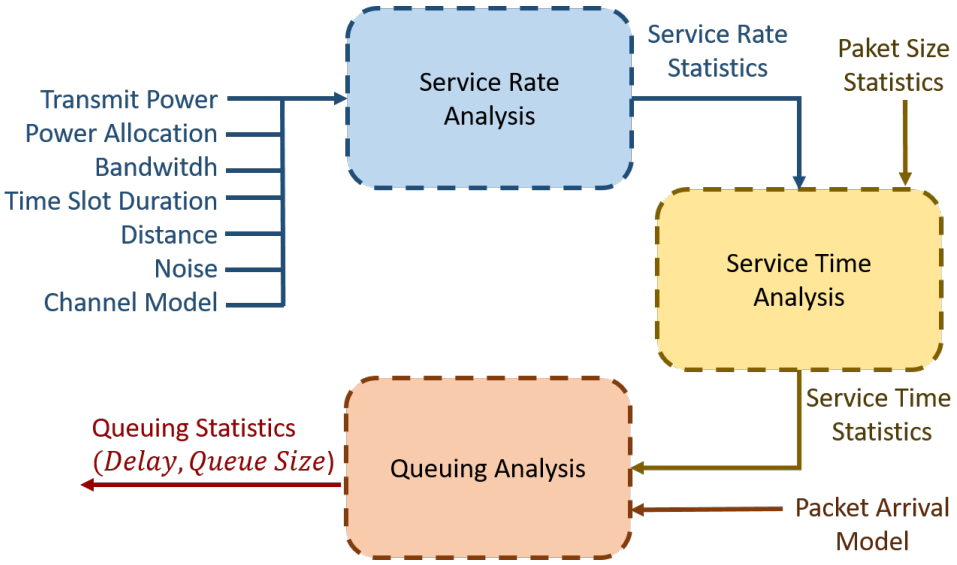


Figure 5.1 : The summary of modelling approach.

queuing formulation for the NOMA downlink system such that each user has its corresponding queue with the packet based random traffic arrival model and the departure process is determined according to the NOMA user service rates under a Rayleigh fading channel. Since we focus on the latency analysis for the URLLC services, we utilize a discrete-time M/G/1 queuing model to obtain the average queuing delay.

In [110], theoretical queuing analysis and system-level simulations are performed to study the system design principles of 5G NR. They emphasize that the queuing delay has an important contribution on the URLLC latency. Although they study both uplink and downlink models for 5G NR, the NOMA technology is not considered in their model. The average queuing delay of 5G NR frame types for both NOMA and OMA downlink systems has been evaluated using our discrete-time M/G/1 queuing model proposed in this chapter.

In this chapter, an analytical model to characterize the average queuing delay for NOMA downlink systems is proposed. The model utilizes a discrete time M/G/1 queuing model where the packet arrival process is assumed to be Poisson distributed. The departure process depends on network settings and resource allocation. The summary of the proposed analytical modelling approach is depicted in Figure 5.1. First, the network settings such as transmit power, bandwidth, and channel model are used to calculate the first and second moment statistics of the user service capacities.

Next, we provide an approximation for the service time statistics under a certain packet size distribution by utilizing the random sums of independent and identically distributed (i.i.d.) random variables. Finally, Pollaczek Khintchine formula and Little's Law are applied to obtain the queuing dynamics such as the average queuing delay. Extensive simulations are carried out to validate the accuracy of the proposed analytical model for both NOMA and OMA under different network settings including bandwidth, traffic arrival rate, and packet size distribution. The results show that the ergodic capacity region of NOMA is a superset of OMA and the NOMA supports higher arrival rates. The numerical results of the analytical model are close to the results of the simulation experiments indicating that the proposed analytical model provides a tight approximation for the average queuing delay. Furthermore, the proposed analytical model is applied to evaluate the performance improvements of the 5G NR concept when the NOMA is utilized with the 5G NR frame types. The results confirm that the 5G NR significantly improves the delay performance as the frame type having wider bandwidth and shorter duration is employed.

5.1 System Model

The system model of downlink transmission including one base station (BS) and K user equipments (UEs) is shown for both non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA) schemes in Figure 5.2. In this model, Power Domain-NOMA (PD-NOMA) is utilized as the NOMA technology while Orthogonal Frequency Division Multiple Acces (OFDMA) is used as the OMA technology. The same radio resources consisting of bandwidth, transmit power, and time slot duration are utilized for both multiple access schemes. In OMA, the bandwidth is equally divided into K subcarriers and each subcarrier is assigned to one UE. The power level of each subcarrier can be determined arbitrarily by the base station by obeying the total transmission power constraint. In NOMA, the whole bandwidth is allocated to all UEs while the total transmission power can be arbitrarily distributed among the UEs. The NOMA concept including the SIC procedure for the full buffer traffic scenario is described in [32].

In the physical layer model considering K-user PD-NOMA downlink system in which all users are allocated to the same frequency subband, the combined signal transmitted

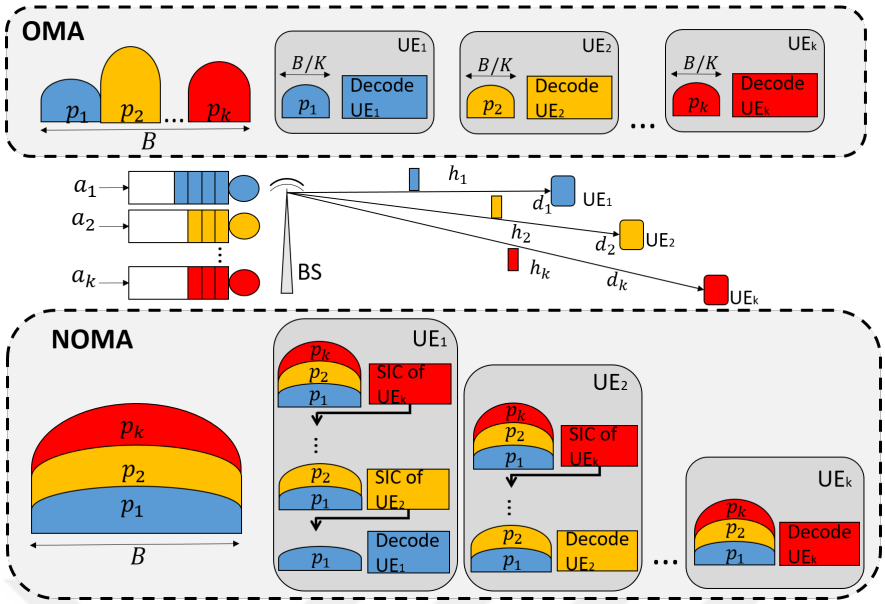


Figure 5.2 : OMA and NOMA downlink system model.

from the base station to users can be represented as $\sqrt{p_1}s_1 + \sqrt{p_2}s_2 + \cdots + \sqrt{p_K}s_K$, where s_k and p_k represent the data symbol per unit energy of k^{th} user and the amount of power allocated for this user, respectively ($1 \leq k \leq K$). The sum of the allocated power levels of all users should be equal to 1 $\left(\sum_{k=1}^{K} p_k = 1\right)$. For a single-input single-output (SISO) system, the received signal of the k^{th} user:

$$y_k = \left(\sqrt{p_k}s_k + \sum_{i=1}^{k-1}\sqrt{p_i}s_i + \sum_{i=k+1}^{K}\sqrt{p_i}s_i\right) \times h_k\sqrt{P_t}\sqrt{PL(d_k)} + w_k \tag{5.1}$$

where w_k represents white Gaussian noise $\sim \mathcal{N}(0, N_0)$ while d_k and h_k represent the distance and the channel gain coefficient between the base station and the k^{th} user, respectively. P_t is the transmit power of the base station. $PL(d_k)$, which is the path loss of the k^{th} user, is calculated using the non-singular path loss model: $PL(d_k) = 1/(1+d_k^{\beta})$, where β is the path loss exponent [120]. The k^{th} user removes the signals of the users having $k+1$ or higher index by employing the SIC method corresponding to the third term in equation (5.1). Assuming that the perfect SIC is performed at the receiver, the signal to interference plus noise ratio (SINR) of the k^{th} user is:

$$\begin{aligned} SINR_k &= \frac{PL(d_k)P_t|h_k|^2 p_k}{PL(d_k)P_t|h_k|^2 \sum_{i=1}^{k-1} p_i + W_{0,k}} \\ s.t. \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k < p_{k+1} \quad k \in [1,2,..K] \,, \end{aligned} \tag{5.2}$$

where, $W_{0,k}$ ($W_{0,k} > 0$) represents the noise power which is calculated according to the double sided white noise such as $W_{0,k} = B \times N_0/2$. Here, B and N_0 represent the bandwidth and noise spectral density, respectively.

If the transmit power of the base station (P_t) is completely assigned to user k, the received power at the user k will be $PR_k = P_t PL(d_k)$. Equation (5.2) can be simplified by defining a new variable $\theta_k = W_{0,k}/PR_k$ as follows:

$$SINR_k = \frac{PR_k X_k p_k}{PR_k X_k \sum_{i=1}^{k-1} p_i + W_{0,k}} = \frac{X_k p_k}{X_k \sum_{i=1}^{k-1} p_i + \theta_k} , \tag{5.3}$$

where $X_k = |h_k|^2$ is the power of the channel gain coefficient.

Let us calculate the SINR equations for OMA. As depicted in Figure 5.2, each UE in OMA is assigned to a separate subcarrier with an equal bandwidth. The power level of each subcarrier can be determined arbitrarily by the base station while the sum of the allocated power coefficients are equal to 1 $\left(\sum_{k=1}^{K} p_k = 1\right)$. As there is no interference effect among subcarriers, UEs can directly decode their corresponding signals without employing the SIC procedure at the receiver side. For a single-input single-output (SISO) system, the received signal of the k^{th} user is:

$$y_k^{oma} = \sqrt{p_k} s_k h_k \sqrt{P_t} \sqrt{PL(d_k)} + w_k . \tag{5.4}$$

Since the entire communication bandwidth is equally divided into K subcarriers, the width of each user subcarrier is equal to $B^{oma} = B/K$ for OMA as depicted in Figure 5.2. Note that the noise power of the k^{th} user is equal to $W_{0,k}^{oma} = BN_0/(2K)$ due to the double sided white noise. Then, the SINR equation of the k^{th} user is expressed as:

$$SINR_k^{oma} = \frac{PL(d_k) P_t |h_k|^2 p_k}{W_{0,k}^{oma}} = \frac{X_k p_k}{\theta_k^{oma}} . \tag{5.5}$$

where $\theta_k^{oma} = W_{0,k}^{oma}/PR_k$. The received power at the user k is $PR_k = P_t PL(d_k)$ assuming that transmit power of the base station (P_t) is completely assigned to user k. We have the following two assumptions for the physical layer models:

Assumption 1: The block (slow) fading Rayleigh channel model is assumed, where the channel gain remains constant at a given time interval nT_s. The channel gain coefficients $h_k[1]$, $h_k[2]$, ..., $h_k[n]$ are i.i.d. sequence of random variables with the Rayleigh distribution which have a finite mean and a finite variance $\forall\, k, n$. Therefore,

the channel power gain ($X_k = |h_k|^2$) is exponentially distributed with the mean value of $E\left[|h_k|^2\right] = 1/\lambda$ and its probability density function (PDF) and cumulative distribution function (CDF) are:

$$\begin{aligned} f_X(x) &= \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & o.w. \end{cases} \\ F_X(x) &= \int_{-\infty}^{x} f_t(t)dt = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} . \end{aligned} \tag{5.6}$$

Assumption 2: The NOMA users are assumed to be ordered for the SIC procedure according to their distances ($d_1 < d_2 .. < d_k$) from the base station instead of their instantaneous channel gains. This assumption is reasonable for practical systems as it is a challenging task to obtain the exact instantaneous users' channel gains at the base station for the SIC ordering. In addition, the perfect SIC procedure is assumed in this chapter.

5.2 Service Capacity Statistics

In this section, the first and second moments of the user service capacity statistics are presented for a single resource block NOMA downlink system under a Rayleigh fading channel. The power allocation coefficients are considered as resource allocation parameters which determine the SINR levels of users, and hence the service capacity statistics.

The capacity in the Rayleigh fading channel can be written as a function of a random variable X, which is exponentially distributed with the density functions in equation (5.6). The instantaneous channel capacity (i.e., instantaneous service capacity) of the k^{th} user is represented as $C_k(X) = B\log_2(1 + SINR_k(X))$. For a K-user NOMA downlink system, UE_k represents the k^{th} user, where $d_k < d_{k+1}$. The service capacity of the k^{th} user is:

$$C_k(X) = B\log_2\left(1 + \frac{p_k X}{X\sum_{i=1}^{k-1} p_i + \theta_k}\right) . \tag{5.7}$$

The first moment of the $C_k(x)$ represents the average service capacity of the user k:

$$\begin{aligned} E[C_k(X)] &= \int_0^\infty C_k(x) f_X(x) dx \\ &= \int_0^\infty B\log_2\left(1+\frac{x p_k}{x\sum_{i=1}^{k-1} p_i+\theta_k}\right)\lambda e^{-\lambda x}dx \\ &= \frac{1}{\log(2)} B\left(\mathrm{e}^{\frac{\lambda\theta_k}{\sum_{i=1}^{k-1} p_i}} Ei\left(-\frac{\lambda\theta_k}{\sum_{i=1}^{k-1} p_i}\right) - \mathrm{e}^{\frac{\lambda\theta_k}{\sum_{i=1}^{k-1} p_i+p_k}} Ei\left(-\frac{\lambda\theta_k}{\sum_{i=1}^{k-1} p_i+p_k}\right)\right) \end{aligned} \tag{5.8}$$

where the function $Ei(z)$ represents the exponential integral function as $Ei(z) = -\int_{-z}^{\infty}\frac{e^{-t}}{t}dt$. Performing the similar approach, the second moment of the $C_k(X)$ is:

$$\begin{aligned} \overline{C_k^2} = E\left[C_k^2(X)\right] &= \int_0^\infty C_k^2(x) f_X(x) dx \\ &= \int_0^\infty \left(B\log_2\left(1+\frac{p_k x}{\sum_{i=1}^{k-1} p_i x+\theta_k}\right)\right)^2 \lambda e^{-\lambda x}dx\,. \end{aligned} \tag{5.9}$$

Upto this point, the service capacity statistics are expressed for K-user NOMA. We now consider the two-user NOMA downlink system, where UE_1 and UE_2 represent the near and far users, respectively ($d_1 < d_2$). The first and second moments of the service capacity can be expressed with special mathematical functions by organizing the equation (5.8) and equation (5.9). First, let us derive the service capacity statistics of the near user UE_1, where interfering signals are removed with the perfect SIC process. The instantaneous service capacity of UE_1:

$$C_1(X) = B\log_2\left(1+\frac{p_1 X}{\theta_1}\right)\,. \tag{5.10}$$

The first moment of the $C_1(X)$ representing the average service capacity of the near user can be expressed as:

$$\begin{aligned} E[C_1] = E[C_1(X)] &= \int_0^\infty B\log_2\left(1+\frac{p_1 x}{\theta_1}\right)\lambda e^{-\lambda x}dx \\ &= -\frac{\mathrm{Be}^{\frac{\lambda\theta_1}{p_1}} Ei\left(-\frac{\lambda\theta_1}{p_1}\right)}{\log(2)}\,. \end{aligned} \tag{5.11}$$

Similarly, the second moment of $C_1(X)$ is:

$$\begin{aligned}\overline{C_1^2} &= E\left[C_1^2(X)\right] = \int_0^\infty C_1^2(x) f_X(x) dx \\ &= \int_0^\infty \left(T_s B \log_2\left(1+\frac{p_1 x}{\theta_1}\right)\right)^2 \lambda e^{-\lambda x} dx \\ &= \frac{2{T_s}^2 \mathrm{B}^2 \mathrm{e}^{\frac{\lambda\theta_1}{p_1}} G_{2,3}^{3,0}\left(\frac{\lambda\theta_1}{p_1}\middle| \begin{matrix}1,1\\0,0,0\end{matrix}\right)}{\log^2(2)}.\end{aligned} \tag{5.12}$$

where, $G_{p,q}^{m,n}\left(z\middle| \begin{matrix}a_1,..,a_n,a_{n+1},..,a_p\\b_1,..,b_m,b_{m+1},..,b_q\end{matrix}\right)$ is the Meijer's G-function [121]. The instantaneous service capacity of the far user UE_2 is:

$$C_2(X) = B\log_2\left(1+\frac{p_2 X}{p_1 X+\theta_2}\right). \tag{5.13}$$

Then, the first moment of the $C_2(X)$ representing the average service capacity of the far user is calculated as:

$$\begin{aligned}E[C_2] = E[C_2(X)] &= \int_0^\infty B\log_2\left(1+\frac{p_2 x}{p_1 x+\theta_2}\right)\lambda e^{-\lambda x} dx \\ &= \frac{1}{\log(2)} B\left(\mathrm{e}^{\frac{\lambda\theta_2}{p_1}} Ei\left(-\frac{\lambda\theta_2}{p_1}\right) - \mathrm{e}^{\frac{\lambda\theta_2}{p_1+p_2}} Ei\left(-\frac{\lambda\theta_2}{p_1+p_2}\right)\right).\end{aligned} \tag{5.14}$$

Similarly, the second moment of the far user service capacity:

$$\begin{aligned}\overline{C_2^2} &= E\left[C_2^2(X)\right] = \int_0^\infty C_2^2(x) f_X(x) dx \\ &= \int_0^\infty \left(B\log_2\left(1+\frac{p_2 x}{p_1 x+\theta_2}\right)\right)^2 \lambda e^{-\lambda x} dx.\end{aligned} \tag{5.15}$$

The derivation of the service capacity statistics for the OMA is not presented since it can be readily obtained by following the similar approach used for the nearest NOMA users with the perfect SIC process. Thus, for the user k in OMA, the first and second moments expressions of the service capacities are:

$$\begin{aligned}E[C_k^{OMA}] &= -\frac{\mathrm{B}_k^{OMA}\mathrm{e}^{\frac{\lambda\theta_k^{OMA}}{p_k}} Ei\left(-\frac{\lambda\theta_k^{OMA}}{p_k}\right)}{\log(2)} \\ \overline{C_{k,OMA}^2} &= 2\left(\frac{\mathrm{B}_k^{OMA}}{\log(2)}\right)^2 \mathrm{e}^{\frac{\lambda\theta_k^{OMA}}{p_k}} G_{2,3}^{3,0}\left(\frac{\lambda\theta_k}{p_k}\middle| \begin{matrix}1,1\\0,0,0\end{matrix}\right).\end{aligned} \tag{5.16}$$

5.3 Queuing Analysis

In our MAC layer model, each user has an infinite First-in-First-out (FIFO) queue at the base station, where incoming packets are stored and forwarded. Similar to the queuing model in [109], for user k, the number of packets in the time slot n represented as $q_k[n]$ can evolve as follows:

$$q_k[n+1] = (q_k[n] + A_k[n] - D_k[n])^+ \tag{5.17}$$

where $(x)^+$ is an operator defined as $max\{0, x_k\}$. The random variable $D_k[n]$ represents the number of packet departures while the random variable $A_k[n]$ represents the number of packet arrivals at the time slot n with the duration of T_s. The service distribution depends on the employed multiple access scheme and power allocation. We have the following assumption for the traffic arrival model:

Assumption 3: The incoming traffic at each queue is assumed to be a Poisson arrival process therefore, $A_k[1]$, $A_k[2]$, ..., $A_k[n]$ are i.i.d. sequence of random variables which have a finite mean and a finite variance $\forall\, k, n$. The random variable $A_k[n]$ is operated for each time interval of nT_s to form K independent Poisson process with the mean value of Λ_k. Furthermore, the packet size is assumed to be an i.i.d. random variable with a finite mean and a finite variance.

The underlying queuing model for downlink multiple access schemes is M/G/1 since the arrival process is Poisson distributed and the departure process is characterized as General distributed, where its statistical information depends on the network settings (e.g., channel model, distance, etc.) and resource allocation decisions (e.g., power allocation, subcarrier assignment, etc.). Let us consider the evolution of the queue size defined in equation (5.17), where the random variable $A_k[n]$ representing the number of arrived packets within one time slot for the user k. The departure process is defined by the random variable $D_k[n]$ representing the number of packets served within one time slot.

Let $S_{k,m}$ be an integer valued random variable representing the k^{th} user service time in terms of the number of time slots required to serve the m^{th} packet with the size of $L_{k,m}$. $R_k[n]$, which represents the amount of served bits within the time slot n, can be calculated by multiplying the channel capacity $C_k[n]$ within the time slot n and the

time slot duration T_s as $R_k[n] = T_s C_k[n]$. Note that the channel capacity $C_k[n]$ remains constant at n^{th} time interval from Assumption 1.

Assuming that $R_k[n]$ is independent and identically distributed, it is a strongly stationary process and its statistical information is independent of time n. Thus the process $(R_k[n]; n \in \mathbb{Z}^+)$ is the joint distribution function of the vector $(R_k[n+1], R_k[n+2], .., R_k[n+j])$ is equal with the one of $(R_k[1], R_k[2], .., R_k[j])$ for any finite set of indices $1, 2, .., j \subset \mathbb{Z}^+$ and any $n \in \mathbb{Z}^+$. The first and the second moment of R_k are:

$$E[R_k] = T_s \times E[C_k] \tag{5.18}$$

$$\overline{R_k^2} = E\left[R_k^2\right] = T_s^2 \times \overline{C_k^2} \tag{5.19}$$

where $E[C_k]$ and $\overline{C_k^2}$ are defined in equation (5.8) and equation (5.9), respectively. Therefore, the variance of the number of bits transmitted within one time slot can be calculated using the equation $Var(R_k) = \overline{R_k^2} - E[R_k]^2$.

Since $R_k[j]$ is finite and T_s is greater than zero, the service time $S_{k,m}$ requires at least one time slot (i.e., $S_{k,m} \geq 1$) to serve a packet having a finite size of $L_{k,m}$. The following equation demonstrates the relation among $L_{k,m}$, $S_{k,m}$, and $R_k[j]$ $(1 \leq j \leq S_{k,m})$:

$$\sum_{j=1}^{S_{k,m}-1} R_k[j] < L_{k,m} \leq \sum_{j=1}^{S_{k,m}} R_k[j] \tag{5.20}$$

When the sum of service capacities within the window of $S_{k,m}$ is greater than the packet size $L_{k,m}$, we assume that it takes $S_{k,m}$ time slots to serve this packet. $Y_{k,m}$ represents the total service capacity within the window of $S_{k,m}$ time slots and can be expressed as:

$$Y_{k,m} = \sum_{j=1}^{S_{k,m}} R_k[j] \ . \tag{5.21}$$

Similar to the approach employed in [108], after the packet m is successfully served, the remaining service capacity at the last time slot $U_{k,m}$ is utilized to serve a portion of the next packet $(m+1)$. Note that the remaining service capability is zero $(U_{k,m} = 0)$ at the beginning. $Y_{k,m}$ can be expressed as follows:

$$Y_{k,m} = L_{k,m} + \Delta U_{k,m} \ , \tag{5.22}$$

where $\Delta U_{k,m} = U_{k,m} - U_{k,m-1}$, $\forall m \in \mathbb{Z}^+$.

Considering when the number of m packet is successfully served, the first moment of Y_k can be calculated using equation (5.22) for m packets:

$$\sum_{i=1}^{m} Y_{k,i} = \sum_{i=1}^{m} \left(L_k[i] + \Delta U_{k,i} \right) \tag{5.23}$$

Since $\sum_{i=1}^{m} \Delta U_{k,i} = U_{k,m}$, the expected value of $E\left[\Delta U_k\right]$ is equal to zero according to the law of large numbers when $m \to \infty$. Then, the first moment of Y_k is:

$$E\left[Y_k\right] = E\left[L_k\right] \ . \tag{5.24}$$

The second moment of Y_k can be calculated using the sum of squares of equation (5.22) for m packets:

$$\sum_{i=1}^{m} \left(Y_{k,i}\right)^2 = \sum_{i=1}^{m} \left(L_k[i] + \Delta U_{k,i} \right)^2 . \tag{5.25}$$

Using the law of large numbers when $m \to \infty$ in (5.25), the second moment of Y_k is:

$$E\left[{Y_k}^2\right] = E\left[{L_k}^2\right] + 2E\left[L_k \Delta U_k\right] + E\left[{\Delta U_k}^2\right] \ . \tag{5.26}$$

Note that $E\left[L_k \Delta U_k\right] = E\left[L_k\right] E[\Delta U_k] + Cov\left(L_k, \Delta U_k\right) = Cov\left(L_k, \Delta U_k\right)$ since $E[\Delta U_k] = 0$, where $Cov\left(L_k, \Delta U_k\right)$ represents the covariance of L_k and ΔU_k. Then, equation (5.26) becomes:

$$E\left[{Y_k}^2\right] = E\left[{L_k}^2\right] + 2Cov\left(L_k, \Delta U_k\right) + E\left[{\Delta U_k}^2\right] \tag{5.27}$$

In this chapter, we assume that the packet size (L_k) is significantly higher than the amount of served bits within one time slot (R_k) so that the remaining service capacity U_k can be neglected compared to L_k. Therefore, we can assume that $E\left[{\Delta U_k}^2\right]$ and $Cov\left(L_k, \Delta U_k\right)$ can be negligible to find the following approximation:

$$E\left[{Y_k}^2\right] \approx E\left[{L_k}^2\right] \ . \tag{5.28}$$

For the sake of simplicity, the amount of served bits for the k^{th} user at the time slot j ($R_k[j]$) will be denoted as R_j. Furthermore, for the m^{th} packet, the service time $S_{k,m}$, the total service capacity within the time window of $S_{k,m}$ ($Y_{k,m}$), and the packet size $L_{k,m}$ will be represented as random variables S, Y, and L, respectively. Then, equation (5.21) can be expressed as random sums of i.i.d. random variables defined in [122]:

$$Y = \sum_{j=1}^{S} R_j \ . \tag{5.29}$$

The first and the second moments of service time are derived in Appendix A and expressed as:

$$\mathrm{E}[S] = \frac{\mathrm{E}[Y]}{\mathrm{E}[R]} . \tag{5.30}$$

$$\mathrm{E}\left[S^2\right] = \overline{S^2} = \frac{\mathrm{E}\left[Y^2\right] - \mathrm{E}[Y]\left(\frac{Var(R)}{\mathrm{E}[R]}\right)}{\mathrm{E}[R]^2} . \tag{5.31}$$

Substituting equation (5.24) and equation (5.28) into equation (5.31) the second moment of the service time can be approximated in terms of the statistics of the random variables R and L as:

$$\overline{S^2} \approx \frac{\mathrm{E}\left[L^2\right] - \mathrm{E}[L]\left(\frac{Var(R)}{\mathrm{E}[R]}\right)}{\mathrm{E}[R]^2} . \tag{5.32}$$

The average service rate of the queue in terms of packets/slot is $\mu = 1/\mathrm{E}[S] = \mathrm{E}[R]/E[L]$. The average arrival rate of the queue Λ packets/slot . The Pollaczek Khintchine formula of the residual service approach together with the Little's Law [123] can be utilized to obtain the average queuing delay of the M/G/1 queuing system in terms of number of time slots:

$$\mathrm{E}[Q] = \frac{\Lambda\overline{S^2}}{2(1-\rho)} + {}^{1}\!/_{\mu} \tag{5.33}$$

where ρ represents the utilization of the queue, which is the ratio of the mean arrival rate over the mean departure rate ($\rho = \Lambda/\mu$). Assuming that the arrival traffic is Poisson with the mean arrival rate of Λ, ρ can be calculated using the mean departure rate in equation (5.8) and equation (5.16) for NOMA and OMA, respectively. Therefore, substituting equation (5.32)) into equation (5.33), the average queuing delay can be obtained for both NOMA and OMA downlink schemes.

5.4 Numerical Results

In this section, the numerical results of the proposed analytical model and Monte Carlo simulation experiments are provided under various network settings such as power allocation rate, bandwidth, traffic arrival rate, and packet size distribution for both NOMA and OMA downlink systems. The performance metrics include individual user service rates and average queuing delays. Unless otherwise is stated, the parameters used for the experiments are given in Table 5.1. We consider two-user scenario, where the transmission bandwidth is 180 KHz for both users in NOMA while it is set to 90

Table 5.1 : Simulation parameters for queuing analysis of NOMA.

Parameter	Value
Transmission Bandwidth	180 KHz
Receive/Transmit Antenna	SISO
Path Loss Exponent (β)	4
Transmit Power (P_t)	0 dBW
Noise Spectral Density (N_0)	-160 dBm/Hz
Rayleigh Fading Parameter (λ)	1
Noise Model	Double-sided White Noise
Path Loss Model	Non-singular Path Loss
Time Slot Duration (TSD)	0.5 ms
Simulation Duration	10^8 TSD
Number of users	2
User distances	$d_1 = 400m$, $d_2 = 1200m$
Packet Size (L)	Constant 4096 bits
User Arrival Rates	2.048×10^5 bps (50 packet/s)

KHz for each user in OMA. Assuming that the total transmit power of the base station is P_t, the transmit power of UE_1 is $p_1 \times P_t$ while the transmit power of the UE_2 is $p_2 \times P_t$ where $p_2 = 1 - p_1$. The distances of UE_1 and UE_2 from the base station are set to $d_1 = 400$ m and $d_2 = 1200$ m, respectively. The numerical results of the average user service capacity and queuing delay are reported for both the analytical model and Monte Carlo simulation experiments.

In the first set of experiments, we focus on the ergodic capacity regions of both NOMA and OMA systems which are calculated by taking all possible power allocations into account, are shown in Figure 5.3. Any vector of arrival rates of UE_1 an d UE_2 lying inside of the ergodic capacity region can yield stable queuing dynamics if the proper power allocation is performed. This demonstrates that NOMA is a superset of OMA in terms of ergodic capacity region. These results indicate that one needs to set the power allocation ratio to an appropriate value to satisfy low latency requirements.

Figure 5.4 shows the average queuing delays of the individual users when the power allocation ratio of UE_1 (p_1) is varied from 0 to 1 for both NOMA and OMA. As p_1 increases, the average queuing delay of UE_1 decreases for both OMA and NOMA while the average queuing delay of UE_2 increases. In addition, when p_1 is higher than 0.42 and 0.94 for NOMA and OMA, respectively, the queuing delay of UE_2

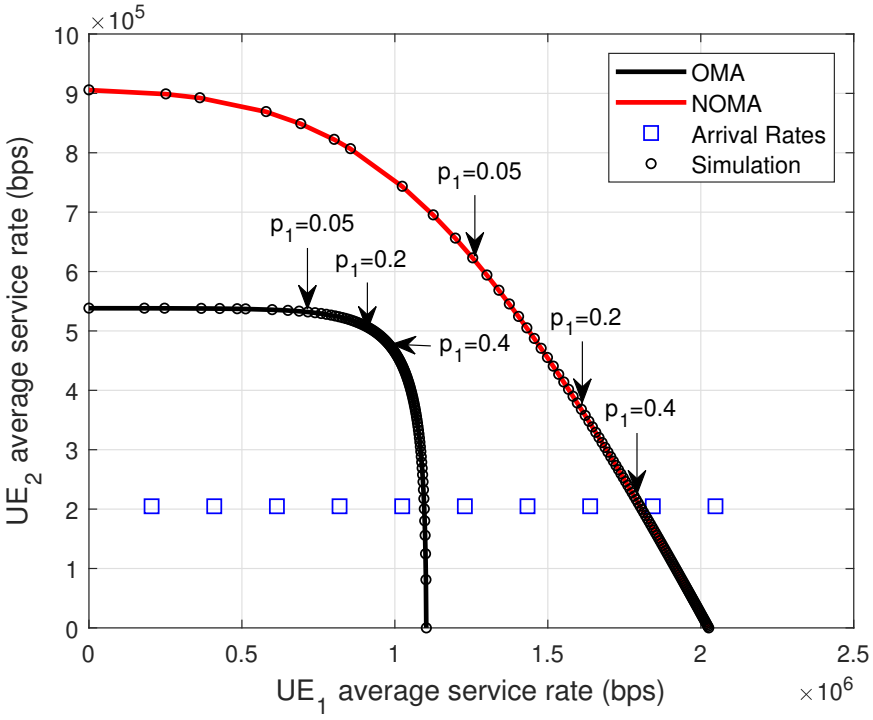


Figure 5.3 : The ergodic capacity regions of OMA and NOMA.

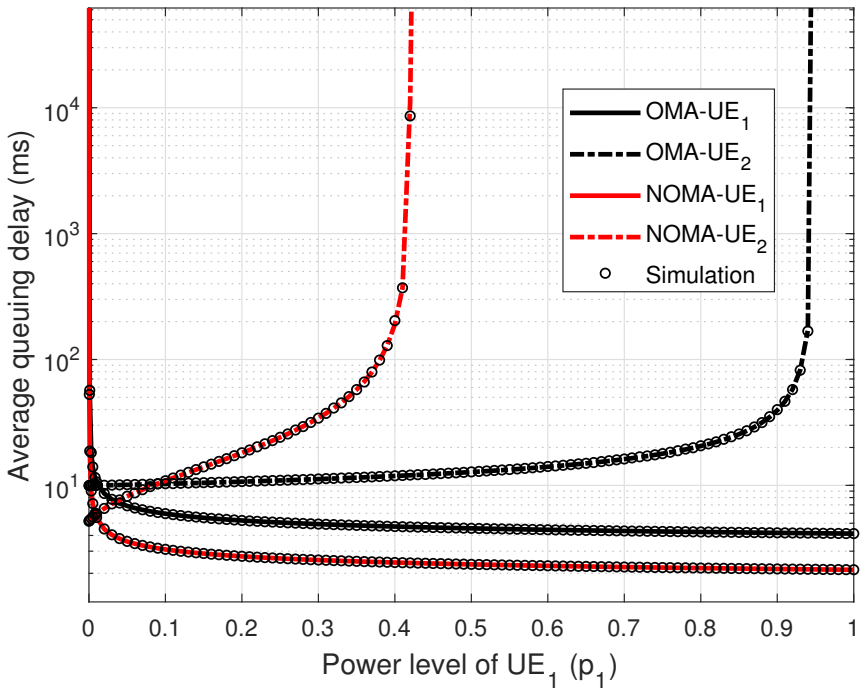


Figure 5.4 : The effects of power allocations on the average queuing delays.

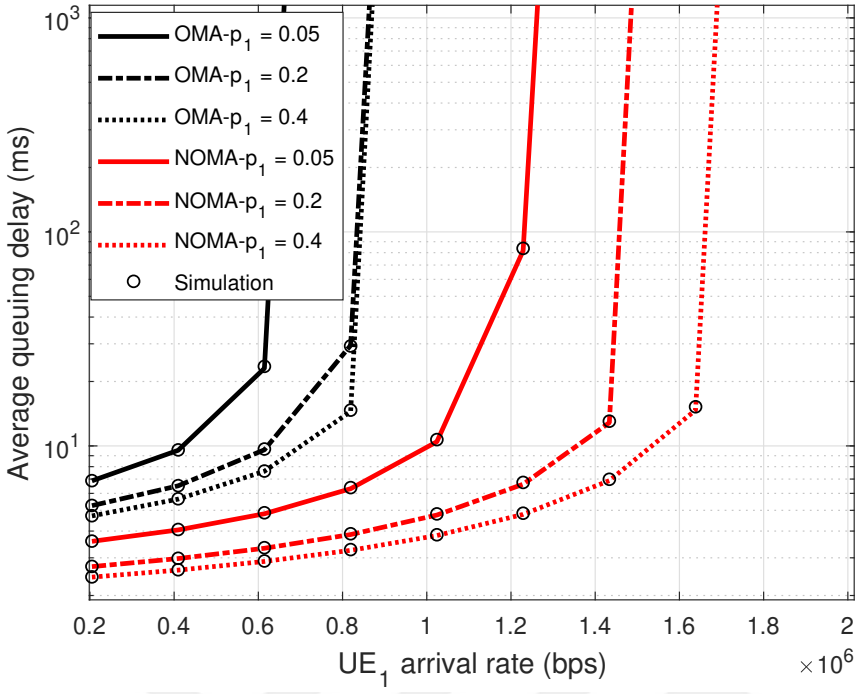


Figure 5.5 : The average queuing delay of the UE_1 versus the UE_1 arrival rates.

exponentially increases since the arrival rate of 2.048×10^5 bps cannot be satisfied using these power allocation ratios.

Figure 5.5 demonstrates the average queuing delays of UE_1 when the arrival rate of UE_2 is remained constant at 2.048×10^5 bps and the arrival rate of UE_1 is varied from 2.048×10^5 bps to 2.048×10^6 bps with the steps of 2.048×10^5 bps as represented by the blue squares in Figure 5.3. The results are obtained for three different power allocation ratios of 0.05, 0.2, and 0.4. The queue of UE_1 is stable when the arrival rate is below the ergodic capacity region shown in Figure 5.3. For example, when the arrival rates of UE_1 and UE_2 are set to 6.144×10^5 and 2.048×10^5 bps, respectively, a stable solution is obtained for both NOMA and OMA. However, when the arrival rates of UE_1 and UE_2 are set to 1.2288×10^6 and 2.048×10^5 bps, respectively, the finite average queuing delay results are obtained only for NOMA. Using the analytical model, we obtain that the arrival rates of UE_1 yielding a stable queue are 1.254×10^6, 1.609×10^6 and 1.788×10^6 bps for NOMA and 7.155×10^5, 8.942×10^5 and 9.839×10^5 bps for OMA when p_1 is set to 0.05, 0.2, and 0.4, respectively.

Table 5.2 : 5G NR frame types.

Frame Type	Subcarrier Spacing (KHz)	RB Bandwidth (KHz)	Time Slot Duration (ms)
0	15	180	0.5
1	30	360	0.25
2	60	720	0.125
3	120	1440	0.0625
4	240	2880	0.03125

5.4.1 Numerical results for 5G NR

The 5G NR, which is based on orthogonal frequency-division multiplexing (OFDM), provides flexibility on the frame structure to support low latency communication. Since a time slot is defined as a fixed number of OFDM symbols, a higher subcarrier spacing leads to a shorter slot duration [124]. The minimum resource allocation unit in LTE is a resource block (RB) and it is composed of 12 consecutive subcarriers and 6 or 7 OFDM symbols corresponding to one time slot with the duration of 0.5 ms. By mapping the similar approach to the 5G NR frame types, when the subcarrier spacing varies as 15 KHz, 30 KHz, 60 KHz, 120 KHz and 240 KHz, the corresponding RB parameters including time slot durations are listed in Table 5.2. Without loss of generality, the same carrier frequency and channel model are used for all experiments in this section. The effects of higher carrier frequencies for wider subcarrier spacing as described in [124] is not within the scope of this work and will be studied as a future work. Excluding the bandwidth and time slot duration values, the simulation parameters are provided in Table 5.1. For the sake of simplicity, the fixed power allocation ratio is employed for both NOMA and OMA. Considering the fairness between UE_1 and UE_2 in terms of the average queuing delays, the p_1 value of 0.05 is selected for further experiments employing the 2.048×10^5 bps of arrival rates for both users as depicted in Figure 5.4.

Figure 5.6 shows the individual average user service rates (bps) for different 5G NR frame structures under the Rayleigh fading channel for NOMA and OMA downlink systems. When the RB bandwidth increases, the SINR values of both users decrease due to the white noise. Therefore, the average user service rates for a single time slot decreases when the frame type varies from 0 to 4. Since the number of time slots per

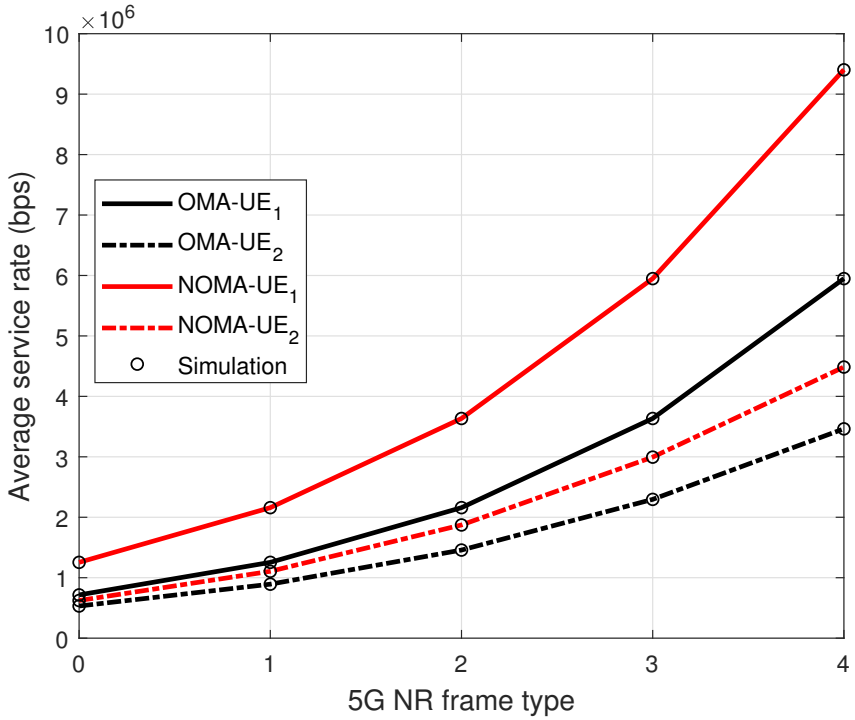


Figure 5.6 : The average user service rates of different 5G NR frame types.

second increases when the time slot duration decreases, the average user service rates increase in terms of bits per second when the frame type varies from 0 to 4.

The service rate of UE_1 are significantly higher when the NOMA is in use for all 5G NR frame structures. However, when the RB bandwidth increases, the rate of average service rate increase for NOMA is lower than OMA for UE_1. For example, the ratio of the average service rate of NOMA over OMA for UE_1 decreases from 1.753 to 1.581 when the frame type varies from 0 to 4. Furthermore, as the noise power increases as a result of the RB bandwidth increase, the rate of average service rate increase for NOMA is higher than OMA for UE_2. For instance, when the frame type varies from 0 to 4, the ratio of the NOMA UE_2 service rate over OMA UE_2 service rate increase from 1.171 to 1.295.

The average queuing delays of individual NOMA and OMA users decrease when 5G NR frame type changes from 0 to 4 as depicted in Figure 5.7. The results indicate that the proposed model and simulation experiments can accurately predict the delay improvements of 5G NR for supporting URLLC services. For example, the delay of less than 1 ms can be achieved for both NOMA users when the 5G NR frame type 4

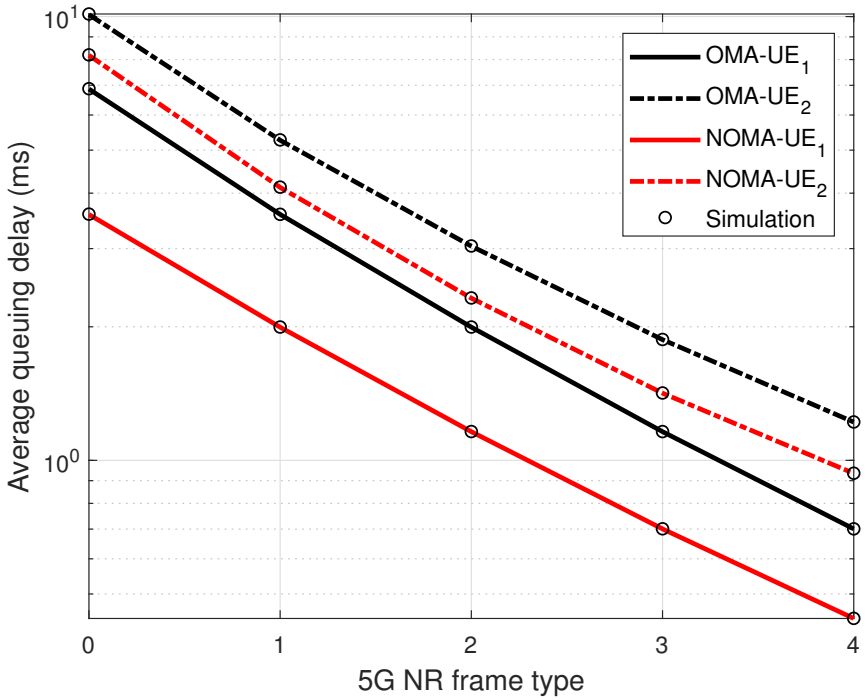


Figure 5.7 : The average user queuing delays of different 5G NR frame types.

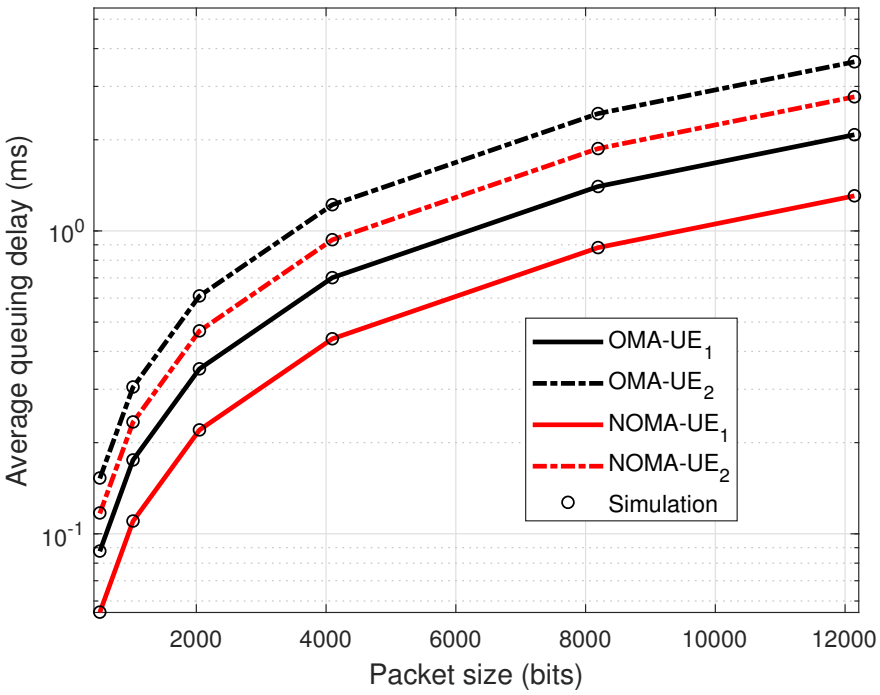


Figure 5.8 : Packet size versus average user queuing delays.

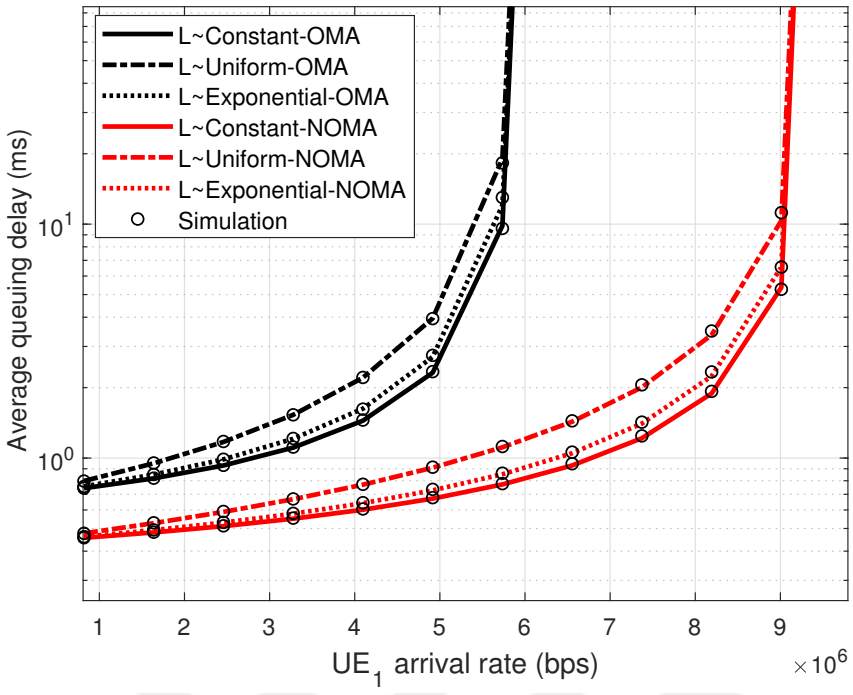


Figure 5.9 : Packet size distribution versus average user queuing delays.

is utilized. Furthermore, the average queuing delays of UE_1 and UE_2 are lower for NOMA compared to OMA users for all 5G NR frame types.

Figure 5.8 shows the effects of packet size on the average queuing delays for OMA and NOMA schemes when the 5G NR frame type 4 is employed. The packet size increases from 64 to 1518 bytes while the user arrival rates are set to the constant value of 2.048×10^5 bps. Without loss of generality, the packet size is an important factor such that the average queuing delay increases with the packet size. NOMA can support higher packet sizes than OMA to provide the same average queuing delay. For example, the results show that the maximum packet size that can result in the average queuing delay of less than 1 ms is 4100 bits for NOMA while it is 3700 bits for OMA.

Figure 5.9 demonstrates the effects of packet size distribution on the average queuing delay of UE_1 for the 5G NR frame type 4 when the arrival rate of UE_1 is varied from 8.192×10^5 bps to 1.024×10^6. The mean packet sizes for three different distributions, namely constant, exponential, and uniform are set to 4096 bits. The standard deviations for exponential and uniform distributions are set to 4096 and 2000, respectively. The numerical results of the analytical model and simulation experiments are close to each other indicating that the proposed analytical model provides a tight

approximation for the average queuing delay for all three packet size distributions. For both NOMA and OMA, when the variance of the packet size is relatively high as in the exponential distribution, the average queuing delay becomes higher compared to the other distributions. When the arrival rate increases, we observe that the effect of packet size distribution on the average queuing delay becomes more visible. Furthermore, the stable queue region is different for OMA and NOMA and independent from the packet size distributions. For example, NOMA can provide a stable queue when the arrival rates of UE_1 is lower than 9.402×10^6 bps while it is 5.948×10^6 bps for OMA.

5.5 Summary

In this chapter, an analytical model utilizing a discrete time M/G/1 queuing model is proposed to characterize the average queuing delay for NOMA downlink systems. The first and second moment statistics of the service time are derived using both packet size and service rate statistics under a Rayleigh fading channel to express the average queuing delay. We perform extensive Monte Carlo simulations and the results verify the accuracy of the proposed analytical model under various network scenarios for both NOMA and OMA schemes. Numerical results show that the ergodic capacity region of NOMA is a superset of OMA indicating that the NOMA can support higher arrival rate and lower latency. Furthermore, the proposed model is applied to demonstrate that the 5G NR frame types having wider bandwidth and shorter duration considerably improves the latency performance. These are promising results such that employing the NOMA technology within the 5G NR concept is a potential enabler to satisfy the challenging latency requirements of time-critical services. For a given set of network parameters including the packet size distribution, the analytical model can be an effective tool for developing the resource allocation techniques that can satisfy the latency requirements of URLLC services.

The outage condition is another important criterion that can needs to be utilized while characterizing the reliability and latency of wireless networks. In the next chapters, we will investigate the outage analysis of NOMA downlink schemes and extend the proposed analytical model by taking the outage constraint into account.

6. OUTAGE PROBABILITY OPTIMIZATION OF NOMA DOWNLINK SYSTEMS

In previous chapters, we have investigated joint power allocation and user grouping for either full buffer or non-full buffer traffic when the CSI is perfectly known at the transmitter. In addition, an analytical model which characterize the queuing delay dynamics for NOMA downlink schemes is presented when the BS has the statistical CSI of the user equipments. In this chapter, we particularly focus the outage analysis of NOMA by providing the closed form expressions of individual user outage probabilities in addition to the system outage probability. Furthermore, the system outage is minimized by optimizing the power allocation among users. The next chapter will study how to control the trade-off between the outage and spectral efficiency in NOMA so that higher throughput and lower latency objectives of 5G can be simultaneously satisfied.

The outage probability is an important metric that can be used to characterize the reliability and latency of wireless networks. For example, the hybrid automatic repeat request (HARQ) is heavily utilized to re-transmit lost data in outage causing additional overhead and latency [74]. The outage event can be defined for cellular systems using various performance metrics such as maximum delay, minimum throughput, minimum BER, and minimum SINR levels. The outage analysis is provided in [38, 66, 71, 75] when each user has a different rate constraint. In [73], the individual and system outage probabilities are used to analyze the secrecy capacity of the NOMA system, where the system outage probability is defined as if both users are in the outage. However, in [77] and [71], a system outage occurs if any or both of the users are in the outage state. We adapt the same definition used in in [77] and [71] for the system outage. The study in [80] investigates the outage probability of OMA downlink transmission, in which the transmitter knows the probability distributions of the fading. By utilizing the similar approach, we analyzed the outage probability of the NOMA downlink transmission under the Rayleigh fading channel model. Further, we present the optimum power

allocation that minimizes the NOMA system outage probability under the assumption that the transmitter knows only the probability distributions of the fading coefficients.

In this chapter, we propose the optimum power allocation that minimizes the system outage probability in PD-NOMA downlink systems. First, the analytical model of the system outage probability is provided as a closed form expression under the Rayleigh fading channel model. Second, we utilize expression to solve the optimum power allocation that minimizes the system outage probability. The proposed power allocation method is compared with fixed NOMA and fractional NOMA and OMA power allocation methods [20]. The accuracy of the theoretical derivations are validated with the Monte Carlo simulations. The results demonstrate that the proposed optimum power allocation method yields the minimum outage probability among all the power allocation schemes of PD-NOMA. The numerical results show that the outage probability depends on the power allocation and the outage probability of OMA with the fractional power allocation is lower than NOMA with the optimum power allocation. However, the spectral efficiency of NOMA is higher since the bandwidth can be utilized by multiple users. These results indicate that the trade-off between the outage and spectral efficiency in NOMA should be carefully controlled to meet higher throughput and lower latency objectives of 5G.

6.1 System Model

The system model of downlink transmission including one base station (BS) and two user equipments (UEs) is shown for both non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA) schemes in Figure 6.1. In this model, PD-NOMA is utilized as the NOMA technology while Orthogonal Frequency Division Multiple Access (OFDMA) is used as the OMA technology. In OMA, the bandwidth is equally divided into two subcarriers and each subcarrier is assigned to one UE. The power level of each subcarrier can be determined arbitrarily by the base station by obeying the total transmission power constraint. However, the whole bandwidth is allocated to both UEs in NOMA while the total transmission power can be arbitrarily distributed among the UEs. The near user (UE_1) employs the SIC procedure to decode its own signal. The far user (UE_2) directly decodes its corresponding signal by considering the signal of UE_1 as interference. The details of the NOMA concept can be found in [32].

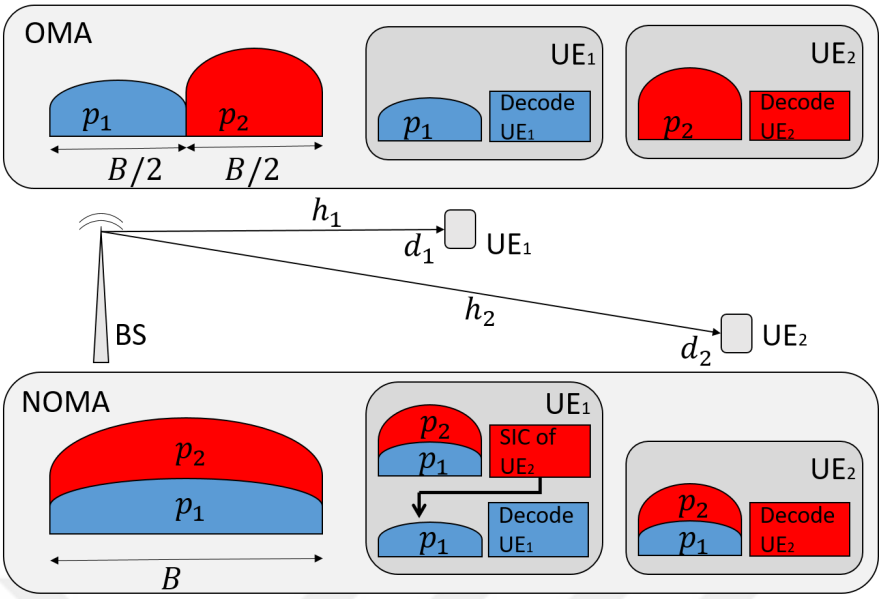


Figure 6.1 : OMA and NOMA downlink system model.

Assuming that PD-NOMA with K users allocated to the same frequency subband, s_k and p_k represent the data symbol per unit energy of kth user and the amount of power allocated for this user, respectively. For a single-input single-output (SISO) system, the received signal of the kth user is:

$$y_k = \left(\sqrt{p_k} s_k + \sum_{i=1}^{k-1} \sqrt{p_i} s_i + \sum_{i=k+1}^{K} \sqrt{p_i} s_i \right) \times h_k \sqrt{P_t} \sqrt{PL(d_k)} + w_k \quad (6.1)$$

$$s.t. \qquad 1 = \sum_{k=1}^{K} p_k$$

where w_k represent the white Gaussian noise $\sim \mathcal{N}(0, N_0)$ while d_k and h_k represents the distance and the channel gain coefficient between the base station and the kth user, respectively. P_t is the transmit power of the base station. $PL(d_k)$ is the path loss of the kth user calculated according to the non-singular path loss model [120]: $PL(d_k) = 1/(1+d_k^{\beta})$, where β is the path loss exponent.

Assume that the channel qualities of the users are arranged in descending order. The kth user removes the signals of the users having k+1 or higher index using the SIC method (equation (6.1)). The signal to interference plus noise (SINR) ratio of the kth

user is:

$$
\begin{aligned}
SINR_k &= \frac{PL(d_k)P_t|h_k|^2 p_k}{PL(d_k)P_t|h_k|^2 \sum_{i=1}^{k-1} p_i + W_{0,k}} \\
s.t. \qquad & 1 = \sum_{k=1}^{K} p_k \ , \\
& p_k < p_{k+1} \quad k \in [1,2,..K] \ .
\end{aligned}
\tag{6.2}
$$

In this equation, $W_{0,k}$ represents the noise power calculated according to the double sided white noise such as $W_{0,k} = B \times N_0/2$, where B and N_0 represent the bandwidth and noise spectral density, respectively.

The users with high channel qualities first decode the signals of the users with lower channel qualities and then the decoded signal is subtracted from the combined received signal to obtain their own signal because the users with lower channel qualities are allocated more power in the NOMA system ($p_k < p_{k+1}$). During the SIC procedure, the SINR level of the lth user's signal at the user k is:

$$
SINR_{k\to l}^{SIC} = \frac{PL(d_k)P_t|h_k|^2 p_l}{PL(d_k)P_t|h_k|^2 \left(\sum_{i=l+1}^{K} p_i\right) + W_{0,k}} \ . \tag{6.3}
$$

We consider slow Rayleigh fading channel with the channel coefficient h_k in our system model. It is assumed that each users' channel coefficient h_k ($k \in [1,2,...K]$) is independent and identically distributed (i.i.d.) random variables with the Rayleigh distribution. The channel power gain ($x_k = |h_k|^2$) is exponentially distributed with the channel mean power of $E\left[|h_k|^2\right] = 1/\lambda$. The following cumulative distribution function (CDF) and probability density function (PDF) will be used in the outage probability analysis in section 6.2.

$$
\begin{aligned}
F_{X_k}(x_k) &= 1 - exp\,(-x_k\lambda\,) \\
f_{X_k}(x_k) &= \lambda\, exp\,(-x_k\lambda\,)
\end{aligned}
\tag{6.4}
$$

6.2 Outage Probability Analysis of NOMA

In this section, the outage probability analysis of the NOMA system is presented. We define a common SINR threshold (τ_{th}) as the minimum required level for successful communication. If the SINR level of a user is greater than or equal to τ_{th} ($\tau_{th} > 0$), it can decode the corresponding signal in both SIC and decoding processes, otherwise the outage event occurs. The predefined SIC order indicates that the near user needs to employ the SIC process to decode its own signal while the far user will receive

near users' signals as interference during the decoding process. For a two-user NOMA scenario, where $UE_1(k=1)$ and $UE_2(k=2)$ representing the near user and far user of a NOMA downlink system ($d_1 < d_2$). The outage SINR condition of user k is represented as γ_k and will be defined below separately for both near and far user according to the SINR threshold τ_{th}.

If the transmit power P_t of the base station were completely assigned to user k, the received power at user k would be $PR_k = P_t PL(d_k)$. To simplify the SINR equations, we define a new variable $\theta_k = W_{0,k}/PR_k$. Then, we obtain the following SINR equations:

$$SINR_{1\to2}^{SIC} = \frac{p_2 PR_1 x_1}{p_1 PR_1 x_1 + W_{0,1}} = \frac{p_2 x_1}{p_1 x_1 + \theta_1} \tag{6.5}$$

$$SINR_1 = \frac{p_1 PR_1 x_1}{W_{0,1}} = \frac{p_1 x_1}{\theta_1} \tag{6.6}$$

$$SINR_2 = \frac{p_2 PR_2 x_2}{p_1 PR_1 x_2 + W_{0,2}} = \frac{p_2 x_2}{p_1 x_2 + \theta_2} \tag{6.7}$$

where $p1$ and $p2$ represent the power allocation ratios at the base station for UE_1 and UE_2, respectively.

First, let us focus on the outage probability of UE_1. The outage event for UE_1 occurs if the required SINR level cannot be reached in neither SIC nor decoding processes. These two conditions are represented as $SINR_{1\to2}^{SIC} < \tau_{th}$ and $SINR_1 < \tau_{th}$. The outage condition γ_1 for UE_1 is defined as either the condition $SINR_{1\to2}^{SIC} < \tau_{th}$ or the condition $SINR_1 < \tau_{th}$ is occurred. The $SINR_{1\to2}^{SIC} < \tau_{th}$ and $SINR_1 < \tau_{th}$ conditions are dependent (overlapping) events as the same random variable of channel power gain (x_1) is used for their SINR calculations. Thus, the outage probability of UE_1 by considering the probabilities of these two conditions can be calculated as follows:

$$\begin{aligned} P(\gamma_1) =& P(SINR_1 < \tau_{th} \cup P(SINR_{1\to2}^{SIC} < \tau_{th}) \\ =& P(SINR_1 < \tau_{th}) + P(SINR_{1\to2}^{SIC} < \tau_{th}) - \\ & P(SINR_1 < \tau_{th} \cap P(SINR_{1\to2}^{SIC} < \tau_{th}) \ . \end{aligned} \tag{6.8}$$

The first term in equation (6.8) represents the probability of UE_1 to decode its own signal. The CDF of this condition is given as:

$$\begin{aligned} F_{SINR_1}(\tau_{th}) &= P(SINR_1 < \tau_{th}) \\ &= 1 - \exp\left(\frac{-\lambda \theta_1 \tau_{th}}{p_1}\right) \ . \end{aligned} \tag{6.9}$$

The probability of UE_1 to decode the signal of UE_2 in the SIC procedure corresponds to the second term in equation (6.8). The CDF of this condition is given as

$$F_{SIC1\to2}(\tau_{th}) = P(SINR^{SIC}_{1\to2} < \tau_{th}) \\ = \begin{cases} 1, & \frac{p_2}{p_1} \leq \tau_{th} \\ 1-\exp\left(\frac{-\lambda\theta_1\tau_{th}}{p_2-p_1\tau_{th}}\right), & o.w\ . \end{cases} \tag{6.10}$$

The third term in equation (6.8) denotes the joint probability that both SIC procedure and decoding process of UE_1 is in the outage and it is given as:

$$= P(SINR_1 < \tau_{th} \cap P(SINR^{SIC}_{1\to2} < \tau_{th}) \\ = \begin{cases} 1-\exp\left(\frac{-\lambda\theta_1\tau_{th}}{p_1}\right), & \frac{p_2}{p_1} < \tau_{th}+1 \\ 1-\exp\left(\frac{-\lambda\theta_1\tau_{th}}{p_2-p_1\tau_{th}}\right), & o.w\ . \end{cases} \tag{6.11}$$

When we place Eqs. 6.9,6.10, and 6.11 into equation 6.8, the outage probability of UE_1 (γ_1) becomes:

$$P(\gamma_1) = \\ \begin{cases} 1, & \frac{p_2}{p_1} \leq \tau_{th} \\ 1-\exp\left(\frac{-\lambda\theta_1\tau_{th}}{p_2-p_1\tau_{th}}\right), & \tau_{th} < \frac{p_2}{p_1} < \tau_{th}+1 \\ 1-\exp\left(\frac{-\lambda\theta_1\tau_{th}}{p_1}\right), & \frac{p_2}{p_1} \geq \tau_{th}+1\ . \end{cases} \tag{6.12}$$

The outage event for UE_2 occurs when the SINR level of UE_2 is lower than the outage SINR threshold ($SINR_2 < \tau_{th}$). This event is represented by the γ_2 condition and hence the outage probability of UE_2 is the same as the probability of meeting the γ_2 condition which is formulated as follows:

$$F_{SINR_2}(\tau_{th}) = P(\gamma_2) = P(SINR_2 < \tau_{th}) \\ = \begin{cases} 1-\exp\left(\frac{-\lambda\theta_2\tau_{th}}{p_2-p_1\tau_{th}}\right), & \frac{p_2}{p_1} \geq \tau_{th} \\ 1, & o.w\ . \end{cases} \tag{6.13}$$

Finally, the system outage can be defined as either one or both of the users is in the outage state. This corresponds to the common outage definition in [77]. Since the outage events γ_1 and γ_2 for UE_1 and UE_2, respectively, are independent, the system outage probability can be calculated as follows:

$$P_{out} = 1-(1-P_{\gamma_1})(1-P_{\gamma_2}) \\ = \begin{cases} 1, & \frac{p_2}{p_1} \leq \tau_{th} \\ 1-\exp\left(-\lambda\tau_{th}\left(\frac{\theta_2+\theta_1}{p_2-p_1\tau_{th}}\right)\right), & \tau_{th} < \frac{p_2}{p_1} < \tau_{th}+1 \\ 1-\exp\left(-\lambda\tau_{th}\left(\frac{\theta_1}{p_1}+\frac{\theta_2}{p_2-p_1\tau_{th}}\right)\right), & \frac{p_2}{p_1} \geq \tau_{th}+1\ . \end{cases} \tag{6.14}$$

The OMA system outage probability analysis can be performed similar to the NOMA. As OMA user signals are transmitted at the base station in different frequency sub-bands, the SIC process and additional interference effects will not be considered. Therefore, the individual outage probability of kth user is calculated according to the $\gamma_k^{OMA} = SINR_k^{OMA} < \tau_{th}$ condition. For the OMA system with two users, the system outage probability can be formulated as:

$$P_{out}^{OMA} = 1 - (1 - P_{\gamma_1}^{OMA})(1 - P_{\gamma_2}^{OMA}) = 1 - \exp\left(-\lambda\,\tau_{th}\left(\frac{p_1\theta_2^{OMA} + p_2\theta_1^{OMA}}{p_1 p_2}\right)\right). \tag{6.15}$$

where $\theta_k^{OMA} = W_{0,k}^{OMA}/PR_k$ corresponds to the ratio of the noise to the received power for OMA, where PR_k is defined in the NOMA outage probability analysis above. Since the length of each user subcarrier is equal to the half of the whole bandwidth ($B^{OMA} = B/2$) in the OMA system as depicted in Figure 5.2, the noise power of the kth user for OMA is $W_{0,k}^{OMA} = B \times N_0/4$.

6.3 Minimizing the Outage Probability

In the previous section, the analysis of the NOMA outage probability is performed and the closed-form expression is obtained in equation 6.14 for two-user NOMA downlink system under the slow Rayleigh Fading channels. In this section, we will utilize the analytical model of the outage probability to calculate the optimum power allocation that minimizes the NOMA system outage. For the NOMA system with two users, it is clear that $p_2 = 1 - p_1$ and the system outage probability in equation (6.14) can be reorganized and simplified to a single variable of p_1 so the power allocation optimization problem can be stated as follows:

$$\begin{aligned} p_1^{opt} &= \arg\min_{p_1} \{P_{out}\} \\ s.t. \quad & 0 < p_1 < 0.5 \end{aligned} \tag{6.16}$$

In this equation, it can be shown that the system outage probability corresponding to the conditions $\frac{1}{p_1} \leq \tau_{th}+1$ and $\tau_{th}+1 < \frac{1}{p_1} < \tau_{th}+2$ are non-convex. We numerically evaluate the system outage probability corresponding to the condition $\frac{1}{p_1} \geq \tau_{th}+2$ using different parameters of λ, τ_{th}, N_0, d_1, d_2, etc. and observe that it is convex when the parameters are selected appropriately. We assume in this chapter that the

appropriate parameters yielding the convex shape are used for the condition $\frac{1}{p_1} \geq \tau_{th}+2$. Then, the optimum power level (p_1^{opt}) can be found where the equation of partial derivation of P_{out} with respect to p_1 is equal to zero:

$$0 = \frac{\partial P_{out}}{p_1 \to p_1^{opt}} = \begin{cases} 0, & \frac{1}{p_1} \leq \tau_{th}+1 \\ \frac{1}{(-1+p_1+p_1\tau_{th})^2} \times \\ \exp\left(\lambda\tau_{th}\left(\frac{\theta_2+\theta_1}{-1+p_1+p_1\tau_{th}}\right)\right) \times \\ \lambda\tau_{th}(1+\tau_{th})(\theta_2+\theta_1), & \tau_{th}+1 < \frac{1}{p_1} < \tau_{th}+2 \\ -\exp\left(\lambda\tau_{th}\left(\frac{\theta_1}{p_1}+\frac{\theta_2}{1-p_1-p_1\tau_{th}}\right)\right) \times \\ \lambda\tau_{th}\left(\frac{\theta_1}{(p_1)^2}+\frac{\theta_2(1+\tau_{th})}{(-1+p_1+p_1\tau_{th})^2}\right), & \frac{1}{p_1} \geq \tau_{th}+2\ . \end{cases} \tag{6.17}$$

The optimum position of power allocation coefficient (p_1^{opt}) is the roots of the third part of equation(6.17):

$$\begin{aligned} 0 = &-\exp\left(\lambda\tau_{th}\left(\frac{\theta_1}{p_1^{opt}}+\frac{\theta_2}{1-p_1^{opt}-p_1^{opt}\tau_{th}}\right)\right) \times \\ &\lambda\tau_{th}\left(\frac{\theta_1}{(p_1^{opt})^2}+\frac{\theta_2(1+\tau_{th})}{(-1+p_1^{opt}+p_1^{opt}\tau_{th})^2}\right)\ . \\ 0 = &\frac{\theta_1}{(p_1^{opt})^2}+\frac{\theta_2(1+\tau_{th})}{(-1+p_1^{opt}+p_1^{opt}\tau_{th})^2}\ . \end{aligned} \tag{6.18}$$

After simplifying equation (6.18), the closed form expression of the optimum power allocation becomes:

$$\begin{aligned} p_1^{opt\,1,2} = &\frac{\theta_1+\theta_1\tau_{th} \mp \sqrt{\theta_1\theta_2+\theta_1\theta_2\tau_{th}}}{\theta_1-\theta_2+2\theta_1\tau_{th}-\theta_2\tau_{th}+\theta_1\tau_{th}^2}\ , \\ s.t.\quad &\frac{1}{p_1^{opt}} \geq \tau_{th}+2\ . \end{aligned} \tag{6.19}$$

6.4 Numerical Results

In this section, the individual and system outage probabilities for three NOMA and one OMA power allocation mechanisms are presented using the proposed analytic model and simulation experiments. Opt-NOMA corresponds to the optimum power

allocation that minimizes the system outage probability, Fix-NOMA corresponds to the fix power allocation such that p_1 is 0.2 and p_2 is 0.8, and Frac-NOMA corresponds to the fractional power allocation based on the received power levels PR_1 and PR_2 for UE_1 and UE_2, respectively (i.e., $p_1 = PR_2/(PR_1 + PR_2)$ and $p_2 = 1 - p_1$). Unless otherwise is stated, the parameters used for the experiments are given in Table 6.1.

Table 6.1 : Simulation parameters for outage analysis of NOMA.

Parameter	**Value**
Transmission Bandwidth	1 Hz
Receive/Transmit Antenna	SISO
Path Loss Exponent (β)	4
Transmit Power (P_t)	1 dB
Noise Spectral Density (N_0)	-100 dBm/Hz
Outage SINR Threshold (τ_{th})	1 dB
Rayleigh Fading Parameter (λ)	1
User Distances (d_1, d_2)	300 m, 800 m
Number of Simulation Trials	10^6
Noise Model	Double-sided White Noise
Path Loss Model	Non-singular Path Loss
Number of users	2

Figure 6.2 shows the outage probabilities of UE_1 and UE_2 when the fractional power allocation is employed for both NOMA and OMA. The power level of UE_1 (p_1) is varied from 0 to 1 for OMA while it is varied from 0 to 0.5 for NOMA since it is not practical to set p_1 beyond 0.5 for NOMA. The results show that the outage probability of each user is always higher for NOMA compared to OMA for all possible power allocations. The outage probability of UE_1 decreases when p_1 increases from 0 to 0.3 while it is exponentially increases when p_1 increases beyond 0.3 and reaches 1 when p_1 is around 0.45. Having p_1 beyond 0.3 increases the SIC failure probability at UE_1 for NOMA and hence the outage probability of UE_1 increases. For OMA, the outage probability of UE_1 decreases while the outage probability of UE_2 increases when p_1 increases since p_2 (1-p_1) decreases. For all experiments, the results of the model match with the simulation results indicating the correctness of the analytical model.

Figure 6.3 shows the system outage probability (P_{out}) when p_1 is varied from 0 to 0.5 for NOMA and from 0 to 1 for OMA. Note that the system is in outage state if any user or both users are in outage state. For all experiments, the analytical and simulation

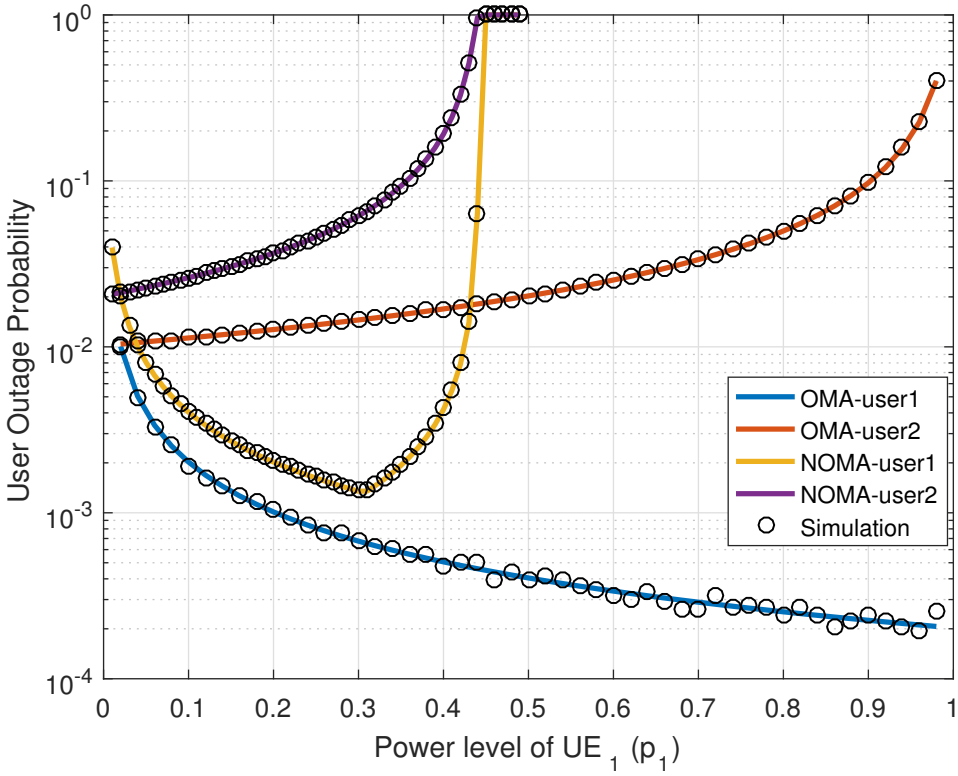


Figure 6.2 : The effects of power allocations on users' outage probabilities.

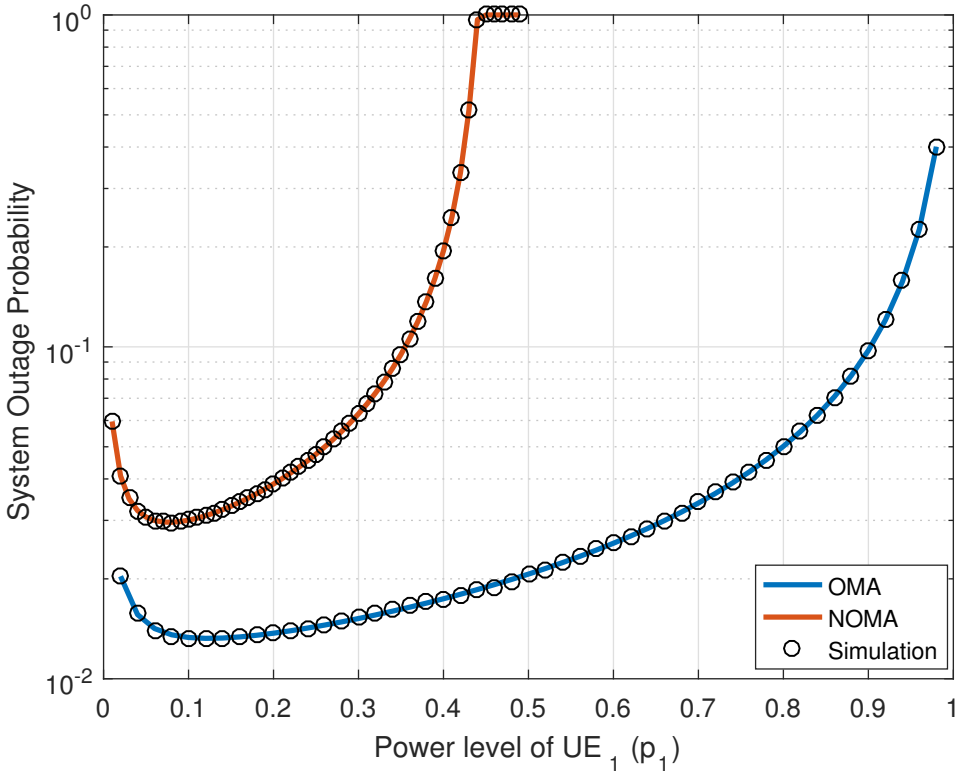


Figure 6.3 : The effects of power allocations on system outage probability.

results match, indicating the correctness of the model. P_{out} is always higher for NOMA compared to OMA for all possible power allocations. P_{out} increases when p_1 increases and reaches 1 when p_1 is around 0.43. These simulation results directly match with the results in equation (6.14), where $p_1 = 0.43, p_2 = 0.57, p_2/p_1 = 1.325$ $\tau_{th} = 1dB = 1.2589W$. The shape of P_{out} for NOMA is convex so that there is an optimum p_1 that minimizes P_{out}. The similar behavior can be observed for OMA but the optimum p_1 value is different compared to NOMA. In this chapter, we propose an analytic model to calculate this optimum p_1 for NOMA. Figure 6.4 and Figure 6.5 compare the outage probabilities of various power allocation methods for NOMA including the optimum one that minimizes P_{out}.

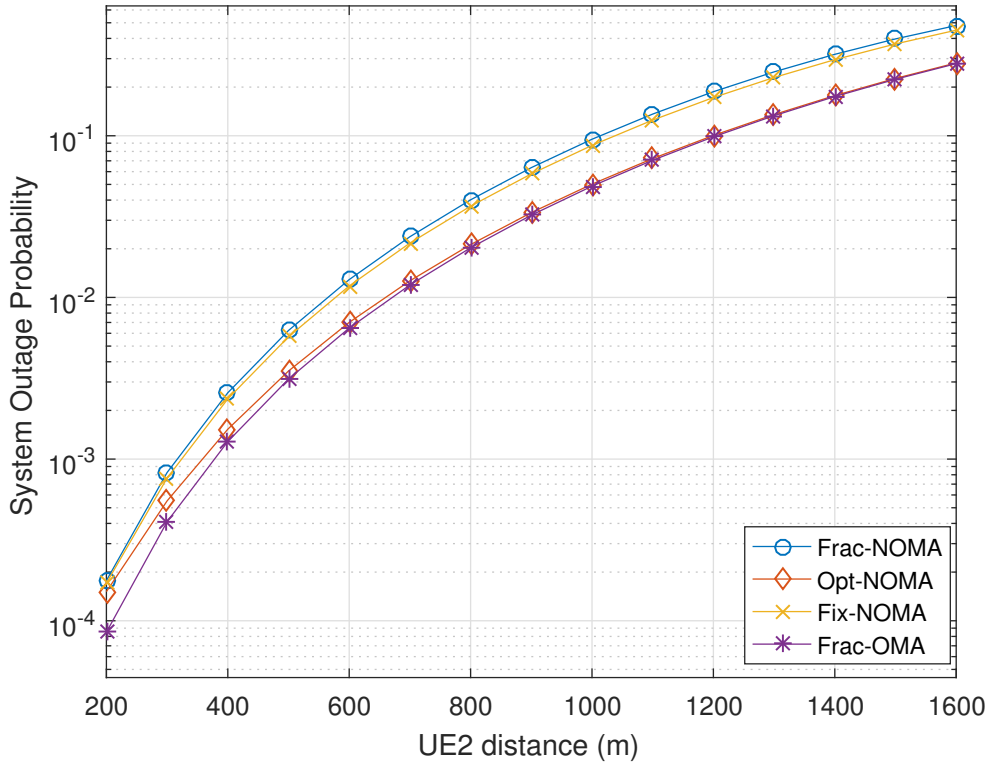


Figure 6.4 : The system outage probabilities versus UE_2 distance.

Figure 6.4 shows the P_{out} results when three NOMA and one OMA power allocation methods are employed when the physical distance from the base station to UE_2 is varied from 200 m to 1600 m while the distance of UE_1 is 100 m. The results show that, for all methods, P_{out} increases when the UE_2 distance increases since the received power and SINR at UE_2 decrease resulting in higher outage probability. Opt-NOMA

yields always the lowest P_{out} among three NOMA power allocation methods while Frac-OMA achieves the lowest P_{out}.

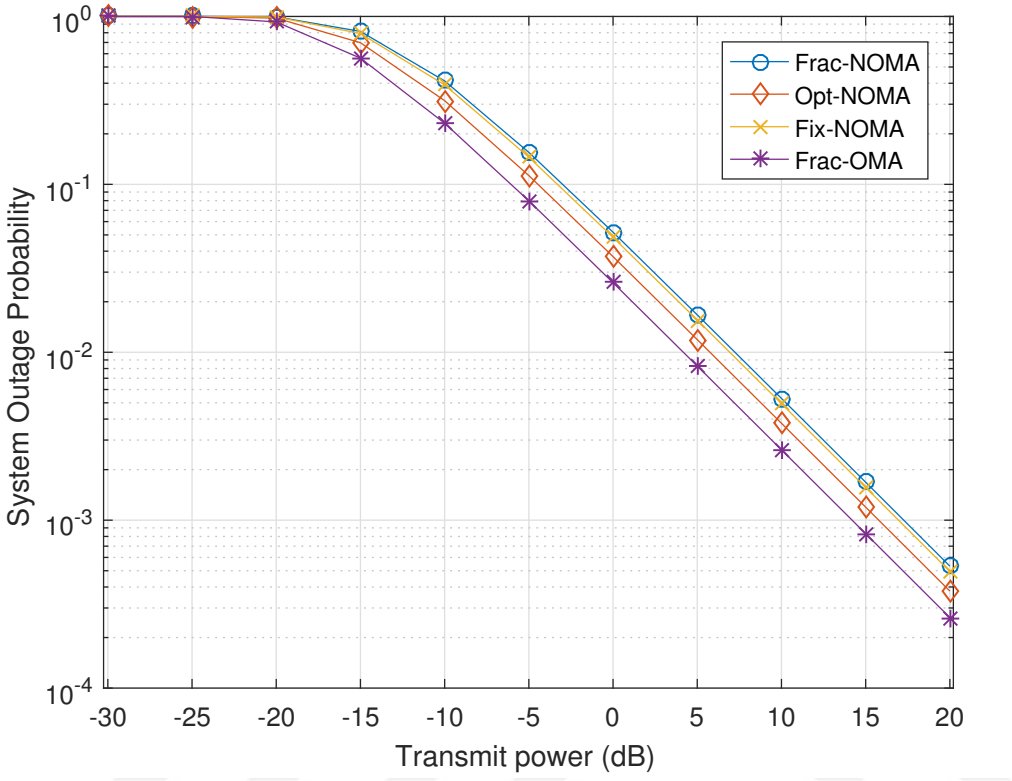


Figure 6.5 : The system outage probability versus the transmit power.

In another set of experiments, the total transmit power (P_t) of the base station is varied from -30 dB to 20 dB while the distances of UE_1 and UE_2 from the base station are set to the fixed values of 300 m and 800 m, respectively. When P_t increases, the system outage probabilities of all power allocation methods including OMA decrease as depicted in Figure 6.5. For Frac-NOMA, the tranmission powers of UE_1 and UE_2 increases with the same rate when P_t increases since the fractional power ratio is constant and does not depend on the absolute value of P_t. In these experiments, p_1 is calculated as 0.0193 and hence p_2 is 0.9807. For all P_t values, Frac-OMA yields the lowest P_{out} among all methods while Opt-NOMA yields the lowest P_{out} among three NOMA methods. When P_t is lower than -25 dB, the P_{out} values of all four methods converge to 1.

6.5 Summary

In this chapter, the analytical model of the system outage probability for the NOMA downlink system under the Rayleigh fading channel is presented. The optimum power allocation yielding the minimum system outage probability is obtained by solving the convex optimization problem. The Monte Carlo simulations demonstrated that the proposed model can be accurately used to characterize the system outage probability. The results show that OMA has lower system outage probability compared to NOMA. However, the spectral efficiency of NOMA is higher since it allows multiple users to share the same radio resource. The next chapter will study how to control the trade-off between the outage and spectral efficiency in NOMA towards meeting higher throughput and lower latency objectives of 5G and provide a comparison between OMA.

7. CROSS-LAYER OPTIMIZATION OF NOMA QUEUING DELAY UNDER SINR OUTAGE CONSTRAINT

In previous chapters, an analytical model of average queuing delay and outage probability analysis are independently studied for NOMA downlink systems. In this chapter, we combine these studies and present a discrete time M/G/1 queuing model by taking the outage event into account such that the user fails either decoding its own signal or performing SIC for the signals of other users at the receiver when the SINR is lower than a predefined outage threshold. The departure process of the queuing model is characterized by obtaining the first and second moment statistics of the service time that depends on the resource allocation strategy and the packet size distribution. The proposed model is utilized to obtain the optimum power allocation that minimizes the maximum of the average queuing delay (MAQD) for a two-user network scenario.

The user plane end-to-end delay of packet transmission can be divided into three main parts: radio access, mobile core, and cloud, where the radio access latency between a base station and user equipment includes over-the-air transmission and propagation, queuing, processing, and re-transmission delays [12]. The outage probability analysis has been taken considerable attention to study the reliability of wireless networks. New analytical models, which can characterize the radio access latency dynamics by taking the outage event into account, are of paramount importance to evaluate the NOMA suitability for URLLC services of 5G NR.

The outage event can be defined for cellular systems using various performance metrics such as maximum delay, minimum throughput, minimum BER, and minimum SINR levels [66, 69, 71, 75]. The study in [78] investigates the outage probability of OMA downlink transmission, in which the transmitter knows the probability distributions of the fading. In [66], the expressions of the average user throughput is provided for both NOMA downlink and uplink systems under the Rayleigh fading channel model by considering target data rates as constraints. The SINR outage constraint is considered in [73] to analyse the individual and system outage probabilities in addition to the

secrecy capacity of the NOMA system under Rayleigh fading channel. These studies assume that the transmitter has the probability distributions of the fading coefficients instead of their realizations. By following a similar approach, our proposed queuing model takes the SINR outage constraint into account for both of the decoding and SIC processes at the receiver. On the other hand, the aforementioned studies focus on only modelling of the throughput, they do not provide higher order statistics of the service rates under the outage constraint. In our study, the first and second moment statistics of the service rate are derived by considering the SINR outage constraint to characterize the latency dynamics of individual users for both NOMA and OMA systems under the Rayleigh fading channel.

The effective capacity approach in [102] is used to accurately predict several link-level QoS metrics such as delay bounds for admission control and resource reservation in wireless communication systems. The effective capacity of NOMA guaranteeing the statistical delay requirements under fading channels has been widely studied [15, 17, 18, 81, 103]. For example, the bisection-based cross-layer power allocation scheme is proposed in [103], where the max-min effective capacity of NOMA is selected as the optimization objective. These studies consider the outage condition as a delay violation constraint while this study presents an analytical model to characterize the average queuing delay under the SINR outage constraint.

[107] investigates the average delay minimization problem for two-user OMA networks and show that the optimal resource allocation policy needs to equalize the queue lengths of both users. We present the optimum cross-layer power allocation framework minimizing the maximum of average queuing delays in two-user NOMA downlink system. Consistent with the proposal in [107], we have analytically shown that the optimal power allocation method yields the minimum average queuing delay by minimizing the difference between the average queuing delays of both users.

The delay violation probability for two-user uplink NOMA systems is presented in [105] by utilizing the stochastic network calculus. They stated that NOMA with the SIC decoding may not be suitable for low latency system under realistic system effects such as imperfect CSI. We have also achieved a similar result for two-user NOMA downlink systems when the SINR outage constraint is set to higher levels.

Stable throughput regions for uplink NOMA systems under unsaturated traffic are investigated using the queuing theory approach, where traffic arrival for each user is assumed to be independent Bernoulli process [106]. In [110], theoretical queuing analysis and system-level simulations are performed to study the system design principles of 5G NR. They emphasize that the queuing effect has an important contribution on the URLLC latency. Although they study both uplink and downlink models for 5G NR, the NOMA technology is not considered in their model. The queuing analysis of block Rayleigh fading channels for conventional OMA system is presented in [108] by utilizing the discrete time discrete state D/G/1 queuing model. They drive the probability distribution of packet service time by taking advantage of the channel distribution of the low SNR regime. In another study [109], a general state space Markov chain model is proposed to calculate the throughput regions of OFDMA users under Rayleigh fading channel by taking the scheduling algorithms into account. The buffer overflow probability providing insights for buffer dimensioning problems is obtained assuming that each user has finite traffic arrival and queue capacity. We adopt a similar system model for the NOMA downlink such that each user has a dedicated queue with the packet based random traffic arrival model and the departure process of each queue is determined by the NOMA resource allocation parameters in addition to the Rayleigh fading channel. By taking both arrival and departure models into account, we utilize a discrete-time M/G/1 queuing model to obtain the average queuing delays of both NOMA and OMA downlink systems with 5G NR frame types.

In Chapter 5), the individual average queuing delay of NOMA users are derived for Poisson distributed packet arrivals with various packet size distributions. However, the proposed approach does not consider the outage condition which is required to accurately model the practical system level enhancements. The closed-form expression of the optimum power allocation that minimizes the system outage probability is provided in Chapter 6 also published as [125].

In this chapter, an analytical model to characterize the queuing delay of NOMA downlink systems under the SINR outage constraint is proposed. The contributions can be summarized as follows:

- The users' service capacities are expressed under a common SINR outage threshold which is the minimum required level to successfully perform both the SIC and

decoding processes. By taking the SINR outage constraint into account, the first and second moment statistics of users' service rates are derived for a NOMA downlink system simultaneously serving K users sharing a single resource block.

- Similar to the queuing model proposed in Chapter 5, for a given probability distribution of the packet size, a fairly close analytical approximation of the first and second moment statistics for the users' service time is obtained by expressing the underlying problem as the random sums of i.i.d random variables. The underlying queuing system with Poisson traffic arrivals becomes M/G/1, where the Pollaczek Khintchine formula of the residual service approach together with the Little's Law are utilized to obtain the average queuing delay.
- We prove that the maximum of average queuing delays for two user NOMA and OMA systems is a unimodel function with a single minimum point for the power allocation yielding stable queues. Using this result, the optimum power allocation framework is proposed by utilizing the M/G/1 queuing model such that the maximum of average queuing delays is minimized for a single resource block simultaneously serving two users.
- The delay optimization framework is applied for the 5G NR concept when the NOMA is utilized.

The proposed analytical model including the approximation of the second moment of the service time is validated by performing the Monte Carlo simulation results. The delay performance of NOMA and OMA is reported using the proposed delay optimization method under various network settings such as SINR outage threshold, user arrival rates and distances. Without considering the SINR outage constraint, the ergodic capacity region of NOMA is a superset of OMA due to its higher spectral efficiency as demonstrated in Chapter 5). As the SINR outage threshold increases, the average queuing delay increases for both NOMA and OMA; however, the rate of increase for NOMA is higher than OMA due to the white noise effect over larger bandwidth. The proposed model in this chapter show that NOMA results in higher delay when the SINR outage threshold is set to higher levels. OMA becomes more preferable than NOMA due to higher noise effect over the 5G NR frame types having wider bandwidth for higher outage thresholds. However, the average queuing

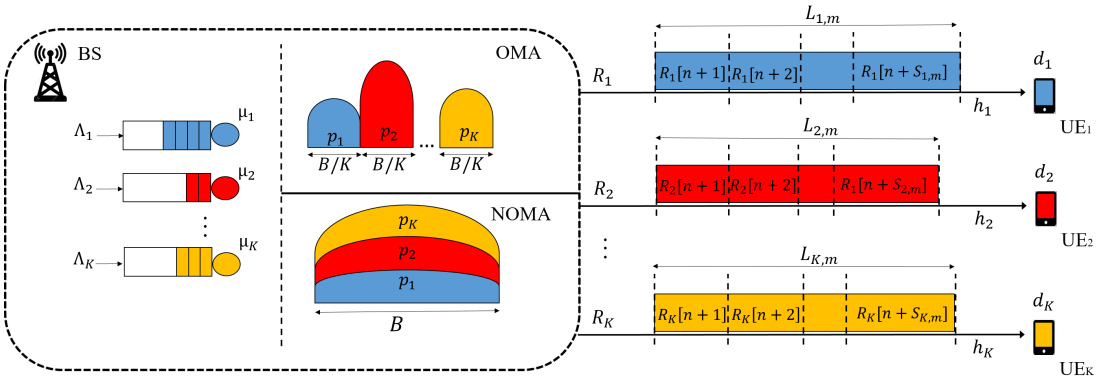


Figure 7.1 : OMA and NOMA downlink system model.

delay performance of NOMA outperforms OMA when the SINR outage threshold is disabled.

7.1 System Model

The downlink transmission system model is shown in Figure 7.1, where one base station utilizes either non-orthogonal multiple access (NOMA) or orthogonal multiple access (OMA) schemes to serve K user equipments (UEs). In this model, Power Domain-NOMA (PD-NOMA) is utilized. The same radio resources consisting of bandwidth, transmit power, and time slot duration are utilized for both multiple access schemes. In OMA, the transmission bandwidth is equally divided into K subcarriers and each subcarrier is assigned to a single UE while the entire bandwidth is allocated to all UEs in NOMA. The total transmission power can be arbitrarily distributed among K UEs for both schemes.

7.1.1 MAC Layer

In the MAC layer, the base station has an infinite First-in-First-out (FIFO) queue for each user to store and forward the corresponding packets. Considering the user k at the time slot n, the number of arrived and served packets within the duration of T_s is represented as independent and identically distributed (i.i.d.) random variables $A_k[n]$ and $D_k[n]$, respectively. The queue size $q_k[n]$ can evolve as:

$$q_k[n+1] = (q_k[n] + A_k[n] - D_k[n])^+ \tag{7.1}$$

where $(x)^+$ is an operator defined as $\max\{0, x_k\}$. The random variable $A_k[n]$ is operated for each time interval of nT_s to form K independent Poisson process with the mean

value of Λ_k packets/slot, where T_s represents the time slot duration. The departure process with the random variable $D_k[n]$ is characterized as General distributed, where its statistical information depends on the network settings (e.g., channel model, distance, etc.) and resource allocation decisions (e.g., power allocation, subcarrier assignment, etc.). Therefore, the underlying queuing model for downlink multiple access schemes forms a discrete time discrete state M/G/1 queue with the mean arrival rate of Λ_k and the mean departure rate of (i.e., service rate) μ_k packets/slot. Furthermore, L_k representing the packet size is assumed to be an i.i.d. random variable with a finite mean ($E[L_k]$) and a finite variance ($var(L_k)$).

Let $S_{k,m}$ be an integer valued random variable representing the service time of the m^{th} packet with the size of $L_{k,m}$ at the user k. The required number of time slots to serve a packet defines the service time. Let $R_k[n]$ represent the amount of served bits within the time slot n. Assuming that $R_k[n]$ is independent and identically distributed, it is a strongly stationary process and its statistical information is independent of time n. Thus, the process $(R_k[n]; n \in \mathbb{Z}^+)$ is the joint distribution function of the vector $(R_k[n+1], R_k[n+2], .., R_k[n+j])$ is equal with the one of $(R_k[1], R_k[2], .., R_k[j])$ for any finite set of indices $1, 2, .., j \subset \mathbb{Z}^+$ and any $n \in \mathbb{Z}^+$.

The service time of the m^{th} packet satisfies the following condition:

$$\sum_{j=1}^{S_{k,m}-1} R_k[j] < L_{k,m} \leq \sum_{j=1}^{S_{k,m}} R_k[j] \tag{7.2}$$

Since $R_k[j]$ is finite, the service time $S_{k,m}$ requires at least one time slot (i.e., $S_{k,m} \geq 1$) to serve a packet having a finite size of $L_{k,m}$. $Y_{k,m}$ representing the consumed service capacity to serve the m^{th} packet is equal to the sum of service capacities within the window of $S_{k,m}$:

$$Y_{k,m} = \sum_{j=1}^{S_{k,m}} R_k[j] \ . \tag{7.3}$$

Equation (7.3) corresponds to the random sums of i.i.d. random variables problem as defined in [122], where the generating functions (g.f.) of random variables can be used to obtain the service time statistics in terms of R_k and Y_k. The g.f of Y_k can be defined as:

$$g_{Y_k}(z) = \mathrm{E}\left[z^i\right] = \sum_{j=1}^{\infty} \mathrm{P}[i=j] z^j \ . \tag{7.4}$$

where $z \in \bar{D}(0:1)$ while $\bar{D}(0:1)$ is the complex closed unit disk centered at 0. Thus, equation (7.3) becomes:

$$g_{Y_k}(z) = g_{S_k}(g_{R_k}(z)) \ . \tag{7.5}$$

where, $g_{S_k}(z)$ and $g_{R_k}(z)$ represents the g.f. of S_k and R_k, respectively. The domain definition of $g_{Y_k}(z)$ contains the open unit disk and the differentiation is possible inside the open disk. Let $z \to 1$, the first and second derivatives of $g_{Y_k}(z)$ are $g'_{Y_k}(1) = \mathrm{E}[Y_k]$ and $g''_{Y_k}(1)) = \mathrm{E}\left[{Y_k}^2 - Y_k\right]$, respectively. Then, $\mathrm{E}[S_k] = \mathrm{E}[Y_k]/\mathrm{E}[R_k]$ using $g'_{Y_k}(1) = g'_{S_k}(1)\, g'_{R_k}(1)$. The second derivative of g_{Y_k} is:

$$g''_{Y_k}(z) = g''_{S_k}(1) \cdot g'_{R_k}(1) \cdot g'_{R_k}(1) + g'_{S_k}(1)\, g''_{R_k}(1) \ . \tag{7.6}$$

When the first and second moments of Y_k, R_k, and S_k are substituted into equation (7.6), the second moment of service time ($\overline{{S_k}^2}$) can be simplified as:

$$\mathrm{E}\left[{S_k}^2\right] = \overline{{S_k}^2} = \frac{\mathrm{E}\left[{Y_k}^2\right] - \mathrm{E}[Y_k]\left(\frac{Var(R_k)}{\mathrm{E}[R_k]}\right)}{\mathrm{E}[R_k]^2} \ . \tag{7.7}$$

We utilize the similar approach in [108] such that after the $(m-1)^{\text{th}}$ packet is successfully served, the remaining service capacity $U_{k,m}$ at the last time slot is used to serve a portion of the m^{th} packet. Note that the remaining service capacity is zero ($U_{k,0} = 0$) at the beginning. Let us define $\Delta U_{k,m} = U_{k,m} - U_{k,m-1}$, $\forall m \in \mathbb{Z}^+$, then $Y_{k,m} = L_{k,m} + \Delta U_{k,m}$. By taking the summation for M packets: $\sum_{m=1}^{M} Y_{k,m} = \sum_{m=1}^{M} \left(L_{k,m} + \Delta U_{k,m}\right)$. Since $\sum_{m=1}^{M} \Delta U_k[m] = U_k[M]$, let $M \to \infty$, utilizing the law of large numbers, $E[Y_k] = E[L_k]$. Similarly, $M \to \infty$ for $\sum_{m=1}^{M} (Y_k[m])^2 = \sum_{m=1}^{M} (L_k[m] + \Delta U_k[m])^2$, we obtain:

$$E\left[{Y_k}^2\right] = E\left[{L_k}^2\right] + 2E[L_k \Delta U_k] + E\left[{\Delta U_k}^2\right] \ . \tag{7.8}$$

Note that $E[L_k \Delta U_k] = E[L_k]E[\Delta U_k] + Cov(L_k, \Delta U_k) = Cov(L_k, \Delta U_k)$ since $E[\Delta U_k] = 0$, where $Cov(L_k, \Delta U_k)$ represents the covariance of L_k and ΔU_k. Then, equation (7.8) becomes:

$$E\left[{Y_k}^2\right] = E\left[{L_k}^2\right] + 2Cov(L_k, \Delta U_k) + E\left[{\Delta U_k}^2\right] \tag{7.9}$$

Assuming that L_k is significantly higher than R_k, U_k can be neglected compared to L_k. Therefore, we can assume that $E\left[{\Delta U_k}^2\right]$ and $Cov(L_k, \Delta U_k)$ can be negligible to find

the following approximation:

$$E\left[{Y_k}^2\right] \approx E\left[{L_k}^2\right] \ . \tag{7.10}$$

One can substitute equation(7.10) into equation (7.7) to approximate $\overline{S_k^2}$ in terms of the statistics of R_k and L_k:

$$\overline{S_k^2} \approx \frac{\mathrm{E}\left[L_k^2\right] - \mathrm{E}\left[L_k\right]\left(\frac{Var(R_k)}{\mathrm{E}[R_k]}\right)}{\mathrm{E}[R_k]^2} \ . \tag{7.11}$$

The Pollaczek Khintchine formula of the residual service approach together with the Little's Law [123] can be utilized to obtain the average queuing delay ($\mathrm{E}[Q_k]$) of the M/G/1 system:

$$\mathrm{E}[Q_k] = \frac{\Lambda_k \overline{S_k^2}}{2\left(1-\rho_k\right)} + \frac{1}{\mu_k} \tag{7.12}$$

where ρ_k represents the utilization of the k^{th} queue, which is the ratio of the mean packet arrival rate over the mean service rate ($\rho_k = \Lambda_k/\mu_k$). The mean service rate of the k^{th} queue in terms of packets/slot is $\mu_k = 1/\mathrm{E}[S_k] = \mathrm{E}[R_k]/E[L_k]$. Therefore, substituting equation (7.11) into equation (7.12):

$$\begin{aligned}\mathrm{E}[Q_k] = &\frac{\mathrm{E}\left[{L_k}^2\right]\Lambda_k \mathrm{E}[R_k] - 2\mathrm{E}[L_k]^2 \Lambda_k \mathrm{E}[R_k]}{2\mathrm{E}[R_k]^2\left(\mathrm{E}[R_k] - \mathrm{E}[L_k]\Lambda_k\right)} + \\ &\frac{\mathrm{E}[R_k]^2 \mathrm{E}[L_k]\left(2+\Lambda_k\right) - \mathrm{E}[L_k]\Lambda_k \mathrm{E}\left[{R_k}^2\right]}{2\mathrm{E}[R_k]^2\left(\mathrm{E}[R_k] - \mathrm{E}[L_k]\Lambda_k\right)} \ .\end{aligned} \tag{7.13}$$

7.1.2 Physical Layer

For K-user PD-NOMA downlink system, the combined transmission signal at the base station is $\sqrt{p_1}s_1 + \sqrt{p_2}s_2 + \cdots + \sqrt{p_K}s_K$, where s_k and p_k represent the data symbol per unit energy and the power allocation level of the k^{th} user, respectively ($1 \leq k \leq K$). The total power allocation levels of all users should be equal to 1 $\left(\sum_{k=1}^{K} p_k = 1\right)$. For a single-input single-output system, the received signal of the k^{th} user:

$$y_k = \left(\sqrt{p_k}s_k + \sum_{i=1}^{k-1}\sqrt{p_i}s_i + \sum_{i=k+1}^{K}\sqrt{p_i}s_i\right) \times h_k\sqrt{P_t}\sqrt{PL(d_k)} + w_k \tag{7.14}$$

where, d_k and h_k represent the distance and the channel gain coefficient between the base station and the k^{th} user, respectively. P_t is the transmit power of the base station

and w_k represents white Gaussian noise $\sim \mathcal{N}(0, N_0)$. We utilize the non-singular path loss model given in [120] as $PL(d_k) = 1/(1+d_k^{\beta})$, where β is the path loss exponent.

At the receiver, the SIC process is employed such that, the user decode and cancel the signals of the other users according to the SIC order. As in [73], the NOMA users are ordered for the SIC procedure according to their distances ($d_1 < d_2 .. < d_k$) from the base station instead of their instantaneous channel gains. Furthermore, the perfect SIC procedure is assumed when the outage constraint is satisfied.

At the user k, the interfering signals of users having $k+1$ or higher index (i.e., the third term in equation (7.14)) is removed by the SIC process. During the SIC procedure, the signal to interference plus noise ratio (SINR) of the l^{th} user's signal (s_l) at the user k is given in equation(7.15). When the SIC procedure is successfully performed, the SINR of the k^{th} user with respect to s_k is given in equation(7.16).

$$SINR_{k\to l}^{SIC} = \frac{PL(d_k)P_t|h_k|^2 p_l}{PL(d_k)P_t|h_k|^2 \left(\sum_{i=l+1}^{K} p_i\right) + W_{0,k}}\,, \tag{7.15}$$

$$\begin{aligned} SINR_k &= \frac{PL(d_k)P_t|h_k|^2 p_k}{PL(d_k)P_t|h_k|^2 \sum_{i=1}^{k-1} p_i + W_{0,k}} \\ s.t. \quad & \sum_{k=1}^{K} p_k = 1\,. \end{aligned} \tag{7.16}$$

Let X_k represent the channel gain power ($X_k = |h_k|^2$) and define a new variable $\theta_k = W_{0,k}/\left(P_t PL(d_k)\right)$ to simplify the equation (7.15) and equation (7.16) as:

$$SINR_{k\to l}^{SIC} = \frac{X_k p_l}{X_k \left(\sum_{i=l+1}^{K} p_i\right) + \theta_k}\,, \tag{7.17}$$

$$SINR_k = \frac{X_k p_k}{X_k \sum_{i=1}^{k-1} p_i + \theta_k}\,. \tag{7.18}$$

When the base station utilize OMA as a downlink multiple access scheme, each UE is assigned to a separate subcarrier with an equal bandwidth of $B^{oma} = B/K$. The power level of each subcarrier can be determined arbitrarily by the base station while the sum of the allocated power coefficients are equal to 1. Hence, the SINR of k^{th} OMA user is:

$$SINR_k^{oma} = \frac{PL(d_k)P_t|h_k|^2 p_k}{W_{0,k}^{oma}} = \frac{X_k p_k}{\theta_k^{oma}} \,. \tag{7.19}$$

where, the noise power is $W_{0,k}^{oma} = BN_0/(2K)$ due to the double sided white noise and $\theta_k^{oma} = W_{0,k}^{oma}/(P_t PL(d_k))$.

The block (slow) fading Rayleigh channel model is assumed, where the channel gain remains constant at a given time interval nT_s. The channel gain coefficients $h_k[1]$, $h_k[2]$, ..., $h_k[n]$ are i.i.d sequence of random variables with the Rayleigh distribution which have a finite mean and a finite variance $\forall\, k,n$. Therefore, the channel power gain ($X_k = |h_k|^2$) is exponentially distributed with the mean value of $E\left[|h_k|^2\right] = 1/\lambda$ and its probability density function is:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & o.w. \,. \end{cases} \tag{7.20}$$

7.2 Service Capacity under Outage Constraint

In this section, the first and second moments of the user service capacity statistics are presented for a single resource block NOMA downlink system under the outage constraint. The service capacity is defined as the amount of served bits within a single time slot. The power allocation coefficients are considered as the resource allocation parameters which determine the SINR levels of users, and hence the service capacity statistics. In Chapter 5 also presented as [125], we have defined an SINR outage constraint for the NOMA downlink scheme and used a predefined SINR threshold τ_{th} for the minimum level required for successful communication. We extend the outage analysis to derive the service capacity statistics of individual NOMA users. For a single time slot, if the SINR level of a user is greater than or equal to τ_{th} ($\tau_{th} \geq 0$), the UEs can decode their corresponding signal in both SIC and decoding processes, otherwise the outage event occurs and the the amount of served bits within a time slot becomes zero. For simplicity, we use X as a channel gain power of the k^{th} user instead of X_k, since $X_1, X_2, .., X_k$ has the same probability distribution.

The outage condition (γ_k) of the user k is expressed in terms of the SINR levels. The outage event for UE_k occurs if the required SINR level cannot be reached in neither

SIC nor decoding processes. Thus, the outage probability of UE_k is:

$$\begin{aligned} P(\gamma_k) =& P\left(SINR_k < \tau_{th}\right) \cup P\left(SINR^{SIC}_{k\to k+1} < \tau_{th}\right) \\ & \cup \ldots \cup P\left(SINR^{SIC}_{k\to K} < \tau_{th}\right) \end{aligned} \tag{7.21}$$

Since the same random variable of channel gain power X is employed for all terms in equation (7.21), the outage probability needs to be calculated by considering the overlapping events. The instantaneous channel capacity of the k^{th} user can be represented as $B\log_2\left(1+SINR_k\left(X\right)\right)$ bits/s.

For the user k, when the SINR condition is satisfied, the amount of served bits within one time slot is represented as $R_k(X)$ and can be calculated by multiplying the instantaneous channel capacity with a time slot duration (T_s), otherwise it will be zero. Note that the channel gain power X remains constant at each time slot interval. For a K-user NOMA downlink system, UE_k represents the k^{th} user, where $d_k < d_{k+1}$. By considering the outage SINR condition, the amount of served bits within one time slot for the k^{th} user can be expressed as:

$$R_k(X) = \begin{cases} T_s B\log_2\left(1+\frac{p_k X}{X\sum\limits_{i=1}^{k-1} p_i+\theta_k}\right), & \gamma_k^c \\ 0, & \gamma_k\ , \end{cases} \tag{7.22}$$

where γ_k^c represents the complement of γ_k such that the user k is not in the outage. The first moment of the $R_k(X)$ representing the average service capacity (bits/slot) for a time slot can be expressed as follows:

$$E\left[R_k\right] = \int\limits_{g(\gamma_k^c)}^{\infty} T_s B\log_2\left(1+\frac{p_k x}{x\sum\limits_{i=1}^{k-1} p_i+\theta_k}\right) f_X\left(x\right)dx\ . \tag{7.23}$$

where $g(\gamma_k^c)$ is the function providing the threshold level of channel gain power according to equation (7.21). Performing the same approach, the second moment of the $R_k(X)$ ($E\left[R_k^2\right]$):

$$\overline{R_k^2} = \int\limits_{g(\gamma_k^c)}^{\infty} \left(T_s B\log_2\left(1+\frac{p_k x}{x\sum\limits_{i=1}^{k-1} p_i+\theta_k}\right)\right)^2 f_X\left(x\right)dx\ . \tag{7.24}$$

Thus, by using equation (7.23) and equation (7.24), the variance of R_k can be calculated as $Var(R_k) = \overline{R_k^2} - E[R_k]^2$.

Up to this point, the first and second moments of the service capacity are expressed for K-user NOMA. We now consider the two-user NOMA downlink system, where UE_1 and UE_2 represent the near and far users, respectively ($d_1 < d_2$). The predefined SIC order indicates that the near user needs to employ the SIC process to decode its own signal while the far user will receive near users' signals as interference during the decoding process. The first and second moments of R_k, for $k = 1,2$ can be expressed with special mathematical functions by organizing the equations equation (7.23) and equation (7.24).

First, let us derive the service capacity statistics of the near user (UE_1) under the SINR outage constraint $\tau_{th} > 0$. The first and second moment statistics of R_1 can be derived by considering the outage events of UE_1. The outage event for UE_1 occurs if the required SINR level cannot be reached in neither SIC nor decoding processes. From equation (7.21) the expression of the outage probability of UE_1 is:

$$P(\gamma_1) = P(SINR_1 < \tau_{th}) \cup P\left(SINR_{1\to 2}^{SIC} < \tau_{th}\right) . \tag{7.25}$$

The $SINR_{1\to 2}^{SIC} < \tau_{th}$ and $SINR_1 < \tau_{th}$ conditions are dependent (overlapping) events as the same random variable of channel gain power (X) is used for their SINR calculations. Using equation (7.17) and equation (7.18) for UE_1, the probability of the complement of γ_1 can be expressed in terms of X:

$$\begin{aligned} P(\gamma_1^c) &= P(X > g(\gamma_1^c)) \\ &= P\left(X > \frac{\theta_1 \tau_{th}}{p_1}\right) \cap P\left(X > \frac{\theta_1 \tau_{th}}{p_2 - p_1 \tau_{th}}\right) . \end{aligned} \tag{7.26}$$

Since θ_1, τ_{th}, and X are always greater than zero, when $p_2/p_1 \leq \tau_{th}$, the probability of satisfying the outage constraint for the SIC procedure is zero (i.e., the second term in equation(7.26)). The intersection point of the first and the second terms satisfies the equation of $\frac{\theta_1 \tau_{th}}{p_1} = \frac{\theta_1 \tau_{th}}{p_2 - p_1 \tau_{th}}$ which yields $p_2/p_1 = \tau_{th} + 1$. Thus, the channel power gain threshold satisfying the outage constraint depends on the values of p_2/p_1 and τ_{th}. The threshold level of the channel gain power is defined with the function $g(\gamma_1^c)$ and expressed as:

$$g(\gamma_1^c)=\begin{cases}\infty, & \frac{p_2}{p_1}\leq\tau_{th}\\ \frac{\theta_1\tau_{th}}{p_2-p_1\tau_{th}}, & \tau_{th}<\frac{p_2}{p_1}<\tau_{th}+1\\ \frac{\theta_1\tau_{th}}{p_1}, & \frac{p_2}{p_1}\geq\tau_{th}+1\ .\end{cases} \tag{7.27}$$

Therefore, the amount of served bits within a single time slot with the duration of T_s for UE_1 is:

$$R_1(X)=\begin{cases}T_sB\log_2\left(1+\frac{p_1X}{\theta_1}\right), & \gamma_1^c\\ 0, & \gamma_1\ .\end{cases} \tag{7.28}$$

Substituting equation (7.27) and equation (7.28) into equation (7.23) and equation (7.24), the first and second moment of R_1 are given by equation (7.29) and equation (7.30), respectively.

$$E[R_1]=\begin{cases}0, & \frac{p_2}{p_1}\leq\tau_{th}\\ \frac{T_sB}{\log(2)}\mathrm{e}^{\frac{\lambda\theta_1\tau_{th}}{-p_2+p_1\tau_{th}}}\left(\log\left(\frac{p_2}{p_2-p_1\tau_{th}}\right)-\mathrm{e}^{\frac{\lambda\theta_1p_2}{p_1(p_2-p_1\tau_{th})}}Ei\left(-\frac{\lambda\theta_1p_2}{p_1(p_2-p_1\tau_{th})}\right)\right), & \tau_{th}<\frac{p_2}{p_1}<\tau_{th}+1\\ \frac{T_sB}{\log(2)}\mathrm{e}^{\frac{-\lambda\theta_1\tau_{th}}{p_1}}\left(\log(1+\tau_{th})-\mathrm{e}^{\frac{\lambda\theta_1(1+\tau_{th})}{p_1}}Ei\left(-\frac{\lambda\theta_1(\tau_{th}+1)}{p_1}\right)\right), & \frac{p_2}{p_1}\geq\tau_{th}+1\ .\end{cases} \tag{7.29}$$

$$\overline{R_1^2}=\begin{cases}0, & \frac{p_2}{p_1}\leq\tau_{th}\\ \left(\frac{T_sB}{\log(2)}\right)^2\mathrm{e}^{\frac{\lambda\theta_1\tau_{th}}{-p_2+p_1\tau_{th}}}\left(\begin{array}{l}\left(\log\left(\frac{p_2}{p_2-p_1\tau_{th}}\right)\right)^2+2\mathrm{e}^{\frac{\lambda\theta_1p_2}{p_1(p_2-p_1\tau_{th})}}\log\left(\frac{p_2}{p_2-p_1\tau_{th}}\right)G_{1,2}^{2,0}\left(\frac{\lambda\theta_1p_2}{p_1(p_2-p_1\tau_{th})}\middle|\begin{array}{c}1\\0,0\end{array}\right)\\ +2\mathrm{e}^{\frac{\lambda\theta_1p_2}{p_1(p_2-p_1\tau_{th})}}G_{2,3}^{3,0}\left(\frac{\lambda\theta_1p_2}{p_1(p_2-p_1\tau_{th})}\middle|\begin{array}{c}1,1\\0,0,0\end{array}\right)\end{array}\right), & \tau_{th}<\frac{p_2}{p_1}<\tau_{th}+1\\ \left(\frac{T_sB}{\log(2)}\right)^2\mathrm{e}^{\frac{-\lambda\theta_1\tau_{th}}{p_1}}\lambda\left(\begin{array}{l}(\log(1+\tau_{th}))^2+2\mathrm{e}^{\frac{\lambda\theta_1(1+\tau_{th})}{p_1}}\log(1+\tau_{th})G_{1,2}^{2,0}\left(\frac{\lambda\theta_1(1+\tau_{th})}{p_1}\middle|\begin{array}{c}1\\0,0\end{array}\right)+\\ 2\mathrm{e}^{\frac{\lambda\theta_1(1+\tau_{th})}{p_1}}G_{2,3}^{3,0}\left(\frac{\lambda\theta_1(1+\tau_{th})}{p_1}\middle|\begin{array}{c}1,1\\0,0,0\end{array}\right)\end{array}\right), & \frac{p_2}{p_1}\geq\tau_{th}+1\ .\end{cases} \tag{7.30}$$

Note that the function $Ei(z)$ represents the exponential integral function as $Ei(z)=-\int_{-z}^{\infty}\frac{e^{-t}}{t}dt$ and $G_{p,q}^{m,n}\left(z\middle|\begin{array}{c}a_1,..,a_n,a_{n+1},..,a_p\\ b_1,..,b_m,b_{m+1},..,b_q\end{array}\right)$ represents the Meijer's G-function [126]. For all $z>0$ the equality of $G_{1,2}^{2,0}\left(z\middle|\begin{array}{c}1\\0,0\end{array}\right)=-Ei(-z)$ is satisfied. Secondly, let us calculate the service capacity statistics of the far user (UE_2) under the SINR outage constraint $\tau_{th}>0$. Since the SIC procedure is not performed in UE_2, the outage condition (γ_2) is satisfied when the SINR level of UE_2 is lower than the outage SINR threshold:

$$P(\gamma_2)=P(SINR_2<\tau_{th})\ . \tag{7.31}$$

Using equation (7.18) for UE_2, the probability of the complement of γ_2 can be expressed in terms of X:

$$P(\gamma_2^c) = P\left(X > g\left(\gamma_2^c\right)\right) = P\left(X > \frac{\theta_2 \tau_{th}}{p_2 - p_1 \tau_{th}}\right). \tag{7.32}$$

The threshold level of the channel gain power for UE_2 is defined with the function $g(\gamma_2^c)$ as:

$$g(\gamma_2^c) = \begin{cases} \infty, & \frac{p_2}{p_1} \leq \tau_{th} \\ \frac{\theta_2 \tau_{th}}{p_2 - p_1 \tau_{th}}, & \frac{p_2}{p_1} > \tau_{th}. \end{cases} \tag{7.33}$$

The amount of served bits within a single time slot for UE_2:

$$R_2(X) = \begin{cases} T_s B \log_2\left(1 + \frac{p_2 X}{p_1 X + \theta_2}\right), & \gamma_2^c \\ 0, & \gamma_2. \end{cases} \tag{7.34}$$

Substituting equation (7.33) and equation (7.34) into equation (7.23) and equation (7.24), the first and second moments of R_2 are:

$$E[R_2] = \begin{cases} 0, & \frac{p_2}{p_1} \leq \tau_{th} \\ \frac{T_s B}{\log(2)} \mathrm{e}^{\frac{\lambda \theta_2}{p_1}} Ei\left(\frac{\lambda \theta_2 p_2}{p_1(-p_2 + p_1 \tau_{th})}\right) - \\ \frac{T_s B}{\log(2)} \mathrm{e}^{\frac{\lambda \theta_2}{p_1 + p_2}} Ei\left(\frac{\lambda \theta_2 p_2 (1+\tau_{th})}{(p_1+p_2)(-p_2+p_1\tau_{th})}\right) & \\ + \frac{T_s B}{\log(2)} \mathrm{e}^{\frac{\lambda \theta_1 \tau_{th}}{-p_2 + p_1 \tau_{th}}} \log(1 + \tau_{th}) & \frac{p_2}{p_1} > \tau_{th}. \end{cases} \tag{7.35}$$

$$\overline{R_2^2} = \begin{cases} 0, & \frac{p_2}{p_1} \leq \tau_{th} \\ \int\limits_{\frac{-\theta_2 \tau_{th}}{-p_2 + p_1 \tau_{th}}}^{\infty} \left(T_s B \log_2\left(1 + \frac{p_2 x}{p_1 x + \theta_2}\right)\right)^2 \lambda e^{-\lambda x} dx, & \frac{p_2}{p_1} > \tau_{th}. \end{cases} \tag{7.36}$$

The derivation of the service capacity statistics for the OMA is not presented since it can be readily obtained by following the similar approach used for the nearest NOMA users with the perfect SIC process. Thus, for the user k in OMA, the first and second moments of the service capacity are given by equation (7.37) and equation (7.38), respectively.

$$E[R_{k,OMA}] = \frac{T_s \mathrm{B}_k^{OMA}}{\log(2)} \mathrm{e}^{\frac{-\lambda \theta_k^{OMA} \tau_{th}}{p_k}} \left(\log(1 + \tau_{th}) - \mathrm{e}^{\frac{\lambda \theta_k^{OMA}(1+\tau_{th})}{p_k}} Ei\left(-\frac{\lambda \theta_k^{OMA}(\tau_{th}+1)}{p_k}\right)\right). \tag{7.37}$$

$$\overline{R^2_{k,OMA}} = \left(\frac{T_s \mathrm{B}_k^{OMA}}{\log(2)}\right)^2 \mathrm{e}^{\frac{-\lambda\theta_k^{OMA}\tau_{th}}{p_k}} \lambda \left(\begin{array}{l} \left(\log^2(1+\tau_{th})\right)^2 + 2\mathrm{e}^{\frac{\lambda\theta_k^{OMA}(1+\tau_{th})}{p_k}} \log(1+\tau_{th}) G^{2,0}_{1,2}\left(\frac{\lambda\theta_k^{OMA}(1+\tau_{th})}{p_k} \middle| \begin{array}{c} 1 \\ 0,0 \end{array} \right) + \\ 2\mathrm{e}^{\frac{\lambda\theta_k^{OMA}(1+\tau_{th})}{p_1}} G^{3,0}_{2,3}\left(\frac{\lambda\theta_k^{OMA}(1+\tau_{th})}{p_k} \middle| \begin{array}{c} 1,1 \\ 0,0,0 \end{array} \right) \end{array} \right). \tag{7.38}$$

7.3 Queuing Delay Optimization

The optimum power allocation coefficients ($\overline{\mathrm{P}_{\mathrm{opt}}}$) that minimize the maximum of average queuing delays (MAQD) for K-user downlink scheme can be expressed as:

$$\begin{aligned} \overline{P_{opt}} &= \underset{\overline{\mathrm{P}}}{argmin} \left\{ \max\left(\overline{Q}\right) \right\} \\ s.t. \quad & \mu_k > \Lambda_k \quad \forall k \in [1,2,..K] \\ & \sum_{k=1}^{K} p_k = 1 \ , \end{aligned} \tag{7.39}$$

where $\overline{\mathrm{P}}$ and $\overline{Q}$ represent the power allocation coefficients $(p_1, p_2, ...p_k)$ and the average queuing delays $(Q_1, Q_2, ...Q_k)$, respectively. Note that for a selected power level allocation P, if any user k does not satisfy the constraint of $\mu_k > \Lambda_k$, the queuing system is not stable and hence, Q_k becomes infinity at the steady state. $\overline{\mathrm{P}_{\mathrm{opt}}}$ yields the minimum MAQD by minimizing the differences among users' average queuing delays. The similar observation is also stated in [107].

Lemma 1 *For two-user NOMA and OMA downlink systems without the SINR outage constraint ($\tau_{th} = 0$), the average service rate of UE$_1$ (μ_1) increases and the average service rate of UE$_2$ (μ_2) decreases when p_1 increases from* 0 *to* 1.

Proof: When the SINR outage constraint is disabled ($\tau_{th} = 0$), the average service rate μ_k (packets/slot) of the k^{th} user:

$$\mu_k = \int_0^{\infty} \frac{T_s B}{E\left[L_k\right]} \log_2\left(1 + SINR_k\left(x\right)\right) f_X(x) dx \ . \tag{7.40}$$

Firstly, let us focus on μ_1. The Leibniz integral rule is used to calculate the derivative of μ_1 with respect to p_1 as:

$$\begin{aligned}\frac{\partial \mu_1}{\partial p_1} &= \frac{\partial \left(\int_0^{\infty} \frac{T_s B}{E[L_1]} \log_2 \left(1+\frac{p_1 x}{\theta_1}\right) \lambda e^{-\lambda x} dx \right)}{\partial p_1} \\ &= \int_0^{\infty} \frac{\frac{T_s B}{E[L_1]}\, x}{\theta_1 \left(1+\frac{p_1 x}{\theta_1}\right) \log(2)} \lambda e^{-\lambda x} dx \\ &= -\frac{T_s B}{E[L_1] \log(2)} \mathrm{e}^{\frac{\lambda \theta_1}{p_1}} Ei\left(-\frac{\lambda \theta_1}{p_1}\right).\end{aligned} \tag{7.41}$$

Let $\phi = \frac{\lambda \theta_1}{p_1}$, $-\mathrm{e}^{\phi} Ei(-\phi) > 0$ for all $\phi > 0$. Since θ_1, λ, $E[L_1]$, and B are always greater than zero, equation (7.41) proves that $\frac{\partial \mu_1}{\partial p_1} > 0$ for all $p_1 \in (0,1)$. Therefore, μ_1 increases when p_1 increases from 0 to 1.

Secondly, substituting $p_2 = 1 - p_1$ into equation (7.18) and the Leibniz integral rule is used to calculate the derivative of μ_2 with respect to p_1:

$$\begin{aligned}\frac{\partial \mu_2}{\partial p_1} &= \frac{\partial \left(\int_0^{\infty} \frac{T_s B}{E[L_2]} \log_2 \left(1+\frac{(1-p_1)x}{p_1 x+\theta_2}\right) \lambda e^{-\lambda x} dx \right)}{\partial p_1} \\ &= \int_0^{\infty} \frac{\partial \left(\frac{T_s B}{E[L_2]} \log_2 \left(1+\frac{(1-p_1)x}{p_1 x+\theta_2}\right) \right)}{\partial p_1} \lambda e^{-\lambda x} dx \\ &= \int_0^{\infty} \frac{T_s B}{E[L_2]} \log_2 \left(\frac{-\frac{(1-p_1)x^2}{(\theta_2+p_1 x)^2} - \frac{x}{\theta_2+p_1 x}}{\left(1+\frac{(1-p1)x}{\theta_2+p1x}\right) \log(2)} \right) \lambda e^{-\lambda x} dx .\end{aligned} \tag{7.42}$$

Since θ_2, λ, $E[L_2]$, and B are greater than zero, equation (7.41) proves that $\frac{\partial \mu_2}{\partial p_1} < 0$ for all $p_1 \in (0,1)$. Therefore, μ_2 decreases when p_1 increases from 0 to 1. Performing the similar approach for two-user OMA system, it can be shown that μ_1^{OMA} increases and μ_2^{OMA} decreases when p_1 increases from 0 to 1, respectively.

Lemma 2 *For a two-user NOMA downlink system, when the SINR outage constraint is enabled ($\tau_{th} > 0$), the average service rate of UE_1 (μ_1) increases as p_1 increases from 0 to $p_{1,opt}^{\mu_1}$, where $p_{1,opt}^{\mu_1}$ represents the optimum power level providing the maximum average service rate of UE_1.*

Proof: When we apply the Leibniz integral rule after substituting $p_2 = 1 - p_1$ and $E[R_1]$ in equation (7.23) into $\mu_1 = E[R_1]/E[L_1]$, the derivative of μ_1 with respect to p_1:

$$\frac{\partial \mu_1}{\partial p_1} = \begin{cases} \frac{\partial}{\partial p_1} \int\limits_{\frac{\theta_1 \tau_{th}}{p_1}}^{\infty} \frac{T_s B}{E[L_1]} \log_2\left(1 + \frac{p_1 x}{\theta_1}\right) \lambda e^{-\lambda x} dx, & p_1 \leq \frac{1}{\tau_{th}+2} \\ \frac{\partial}{\partial p_1} \int\limits_{\frac{\theta_1 \tau_{th}}{(1-p_1)-p_1\tau_{th}}}^{\infty} \frac{T_s B}{E[L_1]} \log_2\left(1 + \frac{p_1 x}{\theta_1}\right) \lambda e^{-\lambda x} dx, & \frac{1}{\tau_{th}+1} > p_1 > \frac{1}{\tau_{th}+2} \\ 0, & p_1 \geq \frac{1}{\tau_{th}+1}\ . \end{cases} \tag{7.43}$$

Case 1: Let us consider the first case for $p_1 \leq \frac{1}{\tau_{th}+2}$:

$$\frac{\partial \mu_1}{\partial p_1} = \frac{T_s B}{E[L_1] {p_1}^2 \log(2)} e^{\frac{-\lambda \theta_1 \tau_{th}}{p_1}} \left(\begin{array}{l} p_1 + e^{\frac{\lambda \theta_1 (1+\tau_{th})}{p_1}} \lambda \theta_1 (1+\tau_{th}) Ei\left(-\frac{\lambda \theta_1 (1+\tau_{th})}{p_1}\right) + \\ \lambda \theta_1 \tau_{th} \left(-e^{\frac{\lambda \theta_1 (1+\tau_{th})}{p_1}} Ei\left(-\frac{\lambda \theta_1 (1+\tau_{th})}{p_1}\right) + \log(1+\tau_{th})\right) \end{array} \right) \tag{7.44}$$

Let $\psi = \lambda \theta_1 (1+\tau_{th})$, ψ is always greater than zero since λ, θ_1, and τ_{th} are always greater than zero. Substituting ψ into the first row inside the parenthesis of equation (7.44), we obtain $p_1 + e^{\frac{\psi}{p_1}} \psi Ei\left(-\frac{\psi}{p_1}\right) > 0$ for $p_1 \in [0,1]$. Then, let $\phi = \frac{\lambda \theta_1}{p_1}$, $-\mathrm{e}^{\phi} Ei(-\phi) > 0$ for all $\phi > 0$ and hence the second row inside the parenthesis of equation (7.44) is always positive.

Since B, $E[L_1]$, and T_s are always greater than zero, $\frac{\partial \mu_1}{\partial p_1} > 0$ is satisfied for $0 < p_1 \leq \frac{1}{\tau_{th}+2}$. Thus, the average service rate μ_1 increases when p_1 increases from 0 to $\frac{1}{\tau_{th}+2}$.

Case 2: For $p_1 \in (\frac{1}{\tau_{th}+2}, \frac{1}{\tau_{th}+1})$ in equation (7.43), the outage event due to the SIC failure becomes a dominating factor compared to the outage event due to the failure of decoding its own signal. The derivative of μ_1 with respect to p_1 is:

$$\frac{\partial \mu_1}{\partial p_1} = \frac{T_s B}{E[L_1] p_1^2 \log(2)} e^{\frac{\lambda \theta_1 (-1+p_1\tau_{th})}{p_1(-1+p_1+p_1\tau_{th})}} \left(\begin{array}{l} e^{\frac{\lambda \theta_1}{-1+p_1+p_1\tau_{th}}} \lambda \theta_1 Ei\left(\frac{-\lambda \theta_1 (-1+p_1)}{p_1(-1+p_1+p_1\tau_{th})}\right) + \\ e^{\frac{\lambda \theta_1}{p_1(-1+p_1+p_1\tau_{th})}} \left(p_1 + \frac{\lambda \theta_1 {p_1}^2 \tau_{th}(1+\tau_{th}) \log\left(1+\frac{p_1\tau_{th}}{-1+p_1}\right)}{(-1+p_1+p_1\tau_{th})^2}\right) \end{array} \right) \tag{7.45}$$

Note that all the parameters in (7.45) always greater than zero. T_s, $E[L_1]$, and B affect the magnitude of the derivative function in equation (7.45) while the values of λ, θ_1,

τ_{th}, and p_1 determine the sign of the function to be either positive or negative. When p_1 increases from $\frac{1}{\tau_{th}+2}$, μ_1 increases to the maximum level for the interval satisfying $\frac{\partial \mu_1}{\partial p_1} > 0$. The maximum value of μ_1 is obtained for $p_{1,opt}^{\mu_1}$ providing $\frac{\partial \mu_1}{\partial p_1} = 0$. For the p_1 values satisfying $\frac{\partial \mu_1}{\partial p_1} < 0$, μ_1 decreases to zero. Note that if there is no p_1 values satisfying $\frac{\partial \mu_1}{\partial p_1} \geq 0$, $p_{1,opt}^{\mu_1} = \frac{1}{\tau_{th}+2}$. Higher p_1 values result in more frequent outage event due to the SIC procedure and μ_1 becomes zero for $p_1 \geq \frac{1}{\tau_{th}+1}$. When p_1 increases from $\frac{1}{\tau_{th}+2}$ to $\frac{1}{\tau_{th}+1}$, the μ_1 either increases first, then decreases or decreases directly depending on the values of λ, θ_1, and τ_{th}. Therefore, the optimum power level ($p_{1,opt}^{\mu_1}$) that maximize μ_1 is located inside the interval of $p_1 \in [\frac{1}{\tau_{th}+2}, \frac{1}{\tau_{th}+1})$ and provides the equality of $\frac{\partial \mu_1}{\partial p_1} = 0$. As a result, μ_1 increases when p_1 increases from 0 to $p_{1,opt}^{\mu_1}$.

Lemma 3 *For a two-user NOMA downlink system, when the SINR outage constraint is enabled ($\tau_{th} > 0$), the average service rate of UE$_2$ (μ_2) decreases as p_1 increases from 0 to $\frac{1}{\tau_{th}+1}$.*

Proof: Substituting $p_2 = 1 - p_1$ into equation (7.35), the average service rate of UE$_2$ (μ_2) is available only for $p_1 < \frac{1}{\tau_{th}+1}$. Thus, the derivative of $\mu_2 = E[R_2]/E[L_2]$ with respect to p_1:

$$\begin{aligned}\frac{\partial \mu_2}{\partial p_1} =& \frac{\partial}{\partial p_1} \int_{\frac{-\theta_2\tau_{th}}{-(1-p_1)+p_1\tau_{th}}}^{\infty} \frac{T_s B}{E[L_2]} \log_2\left(1+\frac{(1-p_1)x}{p_1 x+\theta_2}\right)\lambda e^{-\lambda x} dx \\ =& -\frac{T_s B}{E[L_2]{p_1}^2 \log(2)} e^{\frac{\lambda\theta_2}{p_1}} \lambda\theta_2 Ei\left(\frac{\lambda\theta_2(1-p_1)}{p_1(-1+p_1+p_1\tau_{th})}\right) - \\ & \frac{T_s B}{E[L_2]{p_1}^2 \log(2)} \frac{1}{(-1+p_1+p_1\tau_{th})^2} e^{\frac{\lambda\theta_2\tau_{th}}{-1+p_1+p_1\tau_{th}}} \\ & p_1\left((-1+p_1+p_1\tau_{th})^2 + \lambda\theta_2 p_1\tau_{th}(1+\tau_{th})\log(1+\tau_{th})\right) .\end{aligned} \tag{7.46}$$

The T_s, $E[L_2]$, and B parameters are greater than zero and only affect the magnitude of the derivative function in equation (7.46). The numerical values of $\frac{\partial \mu_2}{\partial p_1}$ are extensively calculated for various θ_2, λ, and τ_{th} values. Since θ_2, λ, and τ_{th} are always greater than zero, it is observed that $\frac{\partial \mu_2}{\partial p_1} < 0$ is satisfied when $p_1 \in \left(0, \frac{1}{\tau_{th}+1}\right)$. Hence, the average service rate of UE$_2$ (μ_2) decreases as p_1 increases from 0 to $\frac{1}{\tau_{th}+1}$.

Lemma 4 *The average queuing delay (Q_k) decreases as the average service capacity $E[R_k]$ increases under a certain condition between packet size and service capacity statistics.*

Proof: The utilization value $\rho_k = \Lambda_k/\mu_k$ decreases when $E[R_k]$ increases for a given distribution L_k with a finite mean of $E[L_k]$, where $\mu_k = E[R_k]/E[L_k]$. However, this is not a sufficient condition to show that the queuing delay decreases as $E[R_k]$ increases. From equation (7.12), Q_k increases as the second moment of the service time ($\overline{S_k^2}$) increases. The $\overline{S_k^2}$ in equation (7.11) can be expressed as:

$$\overline{S_k^2} = \frac{\mathrm{E}\left[L_k^2\right] E[R_k] - \mathrm{E}[L_k]\,\mathrm{E}[\mathrm{R_k}]^2 - \mathrm{E}[L_k]\,\mathrm{E}\left[R^2\right]}{E[R_k]^3}. \tag{7.47}$$

The derivative of $\overline{S_k^2}$ with respect to $E[R_k]$ is:

$$\frac{\partial \overline{S_k^2}}{\partial E[R_k]} = -\frac{2\mathrm{E}\left[L_k^2\right] E[R_k] + \mathrm{E}[L_k]\,\mathrm{E}[\mathrm{R_k}]^2 - 3\mathrm{E}[L_k]\,\mathrm{E}\left[R^2\right]}{E[R_k]^4}. \tag{7.48}$$

The inequality of $\frac{\partial \overline{S_k^2}}{\partial E[R_k]} < 0$ is required to prove that $\overline{S_k^2}$ decreases as $E[R_k]$ increases. Since $E[R_k]$, $\mathrm{E}\left[R^2\right]$, $\mathrm{E}[L_k]$, and $\mathrm{E}\left[L_k^2\right]$ are always greater then zero, the condition that provide $\overline{S_k^2}$ to be a decreasing function is:

$$\mathrm{E}\left[L_k^2\right] > \frac{-\mathrm{E}[L_k] E[R_k]^2 + 3\mathrm{E}[L_k]\,\mathrm{E}\left[R^2\right]}{2E[R_k]}. \tag{7.49}$$

As a result, when the condition defined in equation (7.49) is satisfied, Q_k decreases as $E[R_k]$ increases.

For the finite values of Q_k, let $p_{k,max}^{\Lambda_k}$ and $p_{k,min}^{\Lambda_k}$ represent the power level assignments providing the maximum and minimum average queuing delays of UE_k for the arrival rate of Λ_k, respectively. We combine Lemma 1 through 4 to obtain the following theorem for minimizing the maximum of average queuing delays (MAQD)

Theorem 1 *For two-user NOMA downlink systems with and without the SINR outage threshold, if any p_1 satisfying the constraint of $\mu_k > \Lambda_k \forall k \in [1,2]$ exists, $\max\{Q_1, Q_2\}$ is a unimodel function for all $p_1 \in \left(\max\left\{p_{1,max}^{\Lambda_1}, p_{1,min}^{\Lambda_2}\right\}, \min\left\{p_{1,min}^{\Lambda_1}, p_{1,max}^{\Lambda_2}\right\}\right)$, which has a single minimum point.*

Proof: Firstly, let us consider the p_1 values satisfying the $\mu_1 > \Lambda_1$ condition. The interval of $\left(p_{1,max}^{\Lambda_1}, p_{1,min}^{\Lambda_1}\right)$ represents the p_1 values to meet the $\mu_1 > \Lambda_1$ constraint yielding a stable queue for UE_1. For the case that the SINR outage constraint is enabled ($\tau_{th} > 0$), $p_{1,min}^{\Lambda_2}$ needs to be less then $\frac{1}{\tau_{th}+1}$ from equation (7.27) and it equals to the p_1 value providing the maximum service rate of UE_1 ($p_{1,opt}^{\mu_1}$) from Lemma 4. When p_1 increases from 0 to $p_{1,min}^{\Lambda_2}$, μ_1 increases and hence Q_1 decreases according to the Lemma 4.

Secondly, the interval of $\left(p_{1,min}^{\Lambda_2}, p_{1,max}^{\Lambda_2}\right)$ is defined as the p_1 values satisfying the $\mu_2 > \Lambda_2$ condition. Since μ_2 decreases when p_1 increases from $p_{1,min}^{\Lambda_2}$ to $p_{1,max}^{\Lambda_2}$, Q_2 decreases according to the Lemma 3 and 4.

The interval of p_1 that meets $\mu_1 > \Lambda_1$ and $\mu_2 > \Lambda_2$ is the intersection interval of both conditions and can be represented as $\left(\max\left\{p_{1,max}^{\Lambda_1}, p_{1,min}^{\Lambda_2}\right\}, \min\left\{p_{1,min}^{\Lambda_1}, p_{1,max}^{\Lambda_2}\right\}\right)$. When p_1 increases inside this interval, Q_1 decreases and Q_2 increases. Hence, the function $\max\{Q_1, Q_2\}$ is a unimodel function with a single minimum point. Furthermore, the optimum p_1 value that provides the minimum of $\max\{Q_1, Q_2\}$ satisfies the equation of $Q_1 = Q_2$. Using the similar approach and Lemma 1, it can be shown that this result is also valid for two-user NOMA and OMA downlink systems when the SINR outage threshold is disabled.

An example scenario is performed to numerically calculate the average user service rates and queuing delays for both two-user NOMA and OMA downlink schemes when the power level of UE_1 (p_1) varies from 0 to 1 ($p_2 = 1 - p_1$). For the sake of visualizing Theorem 1 clearly, the average packet size $E[L]$ is set to 2000 bits/packet and the user traffic arrivals rates are set to 500 packets/s and 200 packets/s for UE_1 and UE_2, respectively. Note that the SINR outage threshold τ_{th} is set to 8 dB and the rest of the parameters are given in Table 7.1. Fig 7.2 shows the individual average user service capacities in addition to individual user arrival rates to visualize the power level assignment of $p_{1,min}^{\Lambda_1}$, $p_{1,max}^{\Lambda_1}$, $p_{1,max}^{\Lambda_2}$, and $p_{1,min}^{\Lambda_2}$ for a two-user NOMA downlink system. The average service capacity of UE_1 increases when p_1 increases from 0 to $p_{1,opt}^{\mu_1}$, which is located inside the interval of $\frac{1}{\tau_{th}+1} \geq p_1 > \frac{1}{\tau_{th}+2}$ as stated in Lemma 2. The average service capacity of UE_2 decreases when p_1 increases from 0 to $p_{1,max}^{\Lambda_2}$ as stated in Lemma 3.

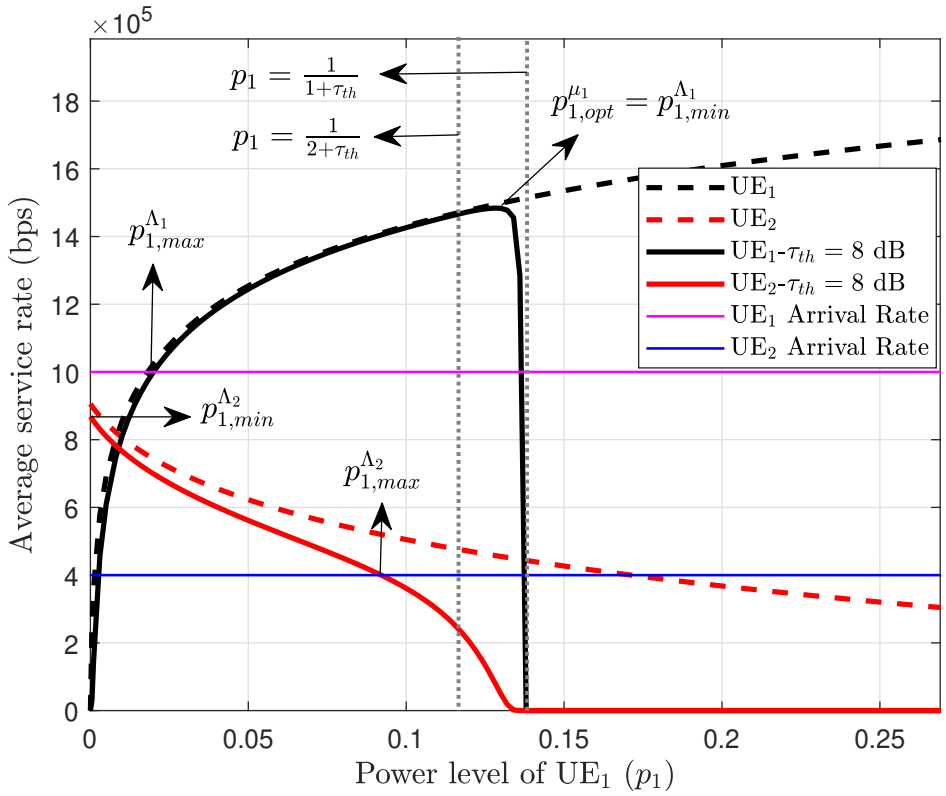


Figure 7.2 : The average user service capacities versus power level assignments.

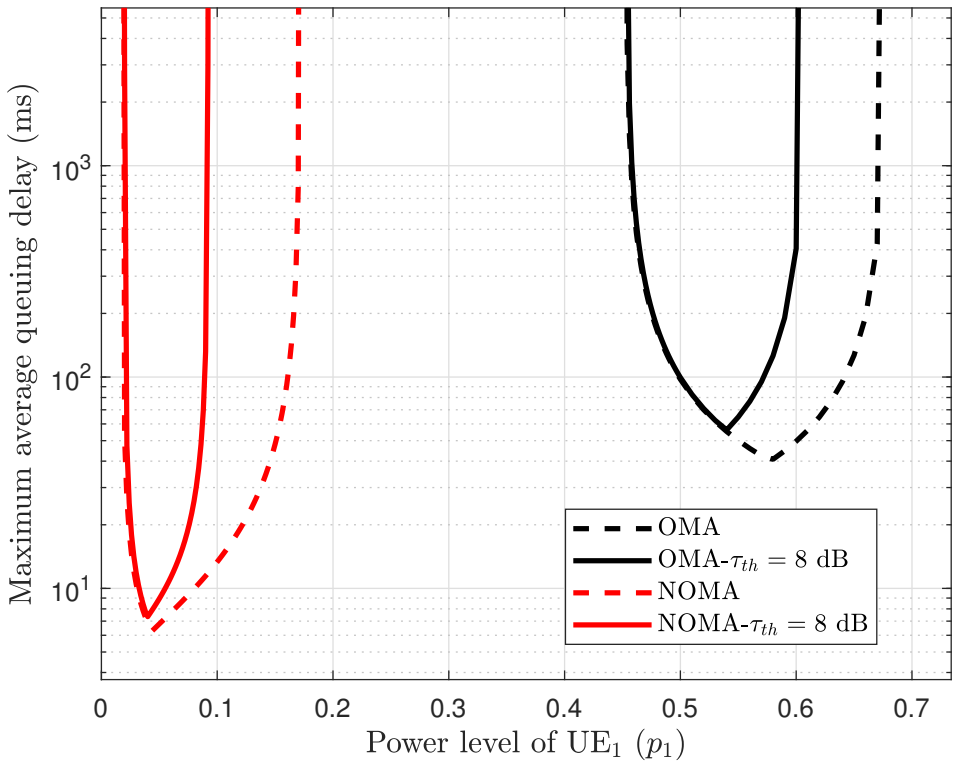


Figure 7.3 : The maximum average queuing delay versus power level assignments.

The maximum value of the users' average queuing delays is shown in Fig 7.3 as the power level of UE_1 (p_1) varies from 0 to 1. The maximum average queuing delay (MAQD) is determined by Q_1 when p_1 is between 0 and a certain power level (i.e., 0.04 and 0.54 for NOMA-$\tau_{th} = 8$ dB, OMA-$\tau_{th} = 8$ dB, respectively) while it is determined by Q_2 beyond that power level. Outside the interval defined in Theorem 1, the MAQD is infinity since one of the users' queue is not stable and hence its average queuing delay becomes infinity at the steady state. This power level corresponds to the optimum operating point resulting in the lowest MAQD value, where the difference between Q_1 and Q_2 is minimum. As depicted in the figure, the optimum power allocation depends not only on the multiple access scheme but also the SINR outage threshold.

The result of the Theorem 1 indicates the existence of the optimum power level of p_1 within a fixed interval for the optimization problem in equation (7.39). For two-user scenario, equation (7.39) can be expressed as:

$$\begin{aligned} &\min_{p_1} \max\{Q_1, Q_2\} \\ &s.t.\ \max\left\{p_{1,max}^{\Lambda_1}, p_{1,min}^{\Lambda_2}\right\} < p_1 < \min\left\{p_{1,min}^{\Lambda_1}, p_{1,max}^{\Lambda_2}\right\} \\ &\quad \mu_k > \Lambda_k \quad \forall k \in [1,2] \\ &\quad p_2 = 1 - p_1. \end{aligned} \tag{7.50}$$

When any p_1 satisfying the constraints of $\mu_1 > \Lambda_1$ and $\mu_2 > \Lambda_2$ exists, the function $\max\{Q_1, Q_2\}$ is a unimodel function and the optimum p_1 is inside the interval of $\left(\max\left\{p_{1,max}^{\Lambda_1}, p_{1,min}^{\Lambda_2}\right\}, \min\left\{p_{1,min}^{\Lambda_1}, p_{1,max}^{\Lambda_2}\right\}\right)$. The solution can be found by several optimization approaches. A golden section search and parabolic interpolation based search algorithm [127] is used in this study to find the optimum p_1 value yielding the minimum of MAQD within a given interval. The used approach effectively explores the minimum point of a cost function within a fixed interval.

7.4 Numerical Results

In this section, the numerical results of the proposed analytical model and Monte Carlo simulation experiments are provided for two-user downlink schemes under various network settings such as SINR outage threshold, traffic arrival rates, and user distances from the base station. The theoretical derivations, including the approximation of the second moment of the service time in Section 7.1, are validated with the simulation results. The optimum power level assignment satisfying the minimum of

Table 7.1 : Simulation parameters for queuing analysis of NOMA with SINR outage.

Parameter	Value
Transmission Bandwidth (B)	180 KHz
Receive/Transmit Antenna	SISO
Path Loss Exponent (β)	4
Transmit Power (P_t)	0 dBW
Noise Spectral Density (N_0)	-160 dBm/Hz
Rayleigh Fading Parameter (λ)	1
User Distances (d_1, d_2)	400 m, 1200 m
Noise Model	Double-sided White Noise
Path Loss Model	Non-singular Path Loss
Time Slot Duration (T_s)	0.5 ms
Number of Packets	10^8
Number of Users (K)	2
Packet Size (L)	Uniform Distributed $E[L] = 2000$ bits/packet $\sigma_L = 1000$ bits/packet
UE_1 Arrival Rate	50 packets/s (100 Kbps)
UE_2 Arrival Rate	50 packets/s (100 Kbps)

maximum average queuing delays (MAQD) is employed for various scenarios and the queuing delay performances of NOMA and OMA considering the outage threshold is compared. Unless otherwise is stated, the parameters used for the experiments are given in Table 7.1. We consider two-user scenario, where the transmission bandwidth is 180 KHz for both users in NOMA while it is set to 90 KHz for each user in OMA. Assuming that the total transmit power of the base station is P_t, the transmit power of UE_1 is $p_1 \times P_t$ while the transmit power of the UE_2 is $p_2 \times P_t$ where $p_2 = 1 - p_1$. The distances of UE_1 and UE_2 from the base station are set to $d_1 = 400$ m and $d_2 = 1200$ m, respectively. The packet size L is uniform distributed with the mean value of 2000 bits/packet and the standard deviation of 1000 bits/packet while the user traffic arrivals are set to 50 packets/s for both UE_1 and UE_2. The provided numerical results for both service capacity and average queuing delay are given in the units of seconds (i.e., bps for service capacity, ms for the average queuing delay) to clarify the effects of 5G NR frame types having different time slot duration.

Figure 7.4 shows the ergodic capacity regions of both NOMA and OMA systems, which are calculated by taking all possible power allocations into account. Any vector of arrival rates of UE_1 and UE_2 lying inside of the ergodic capacity region

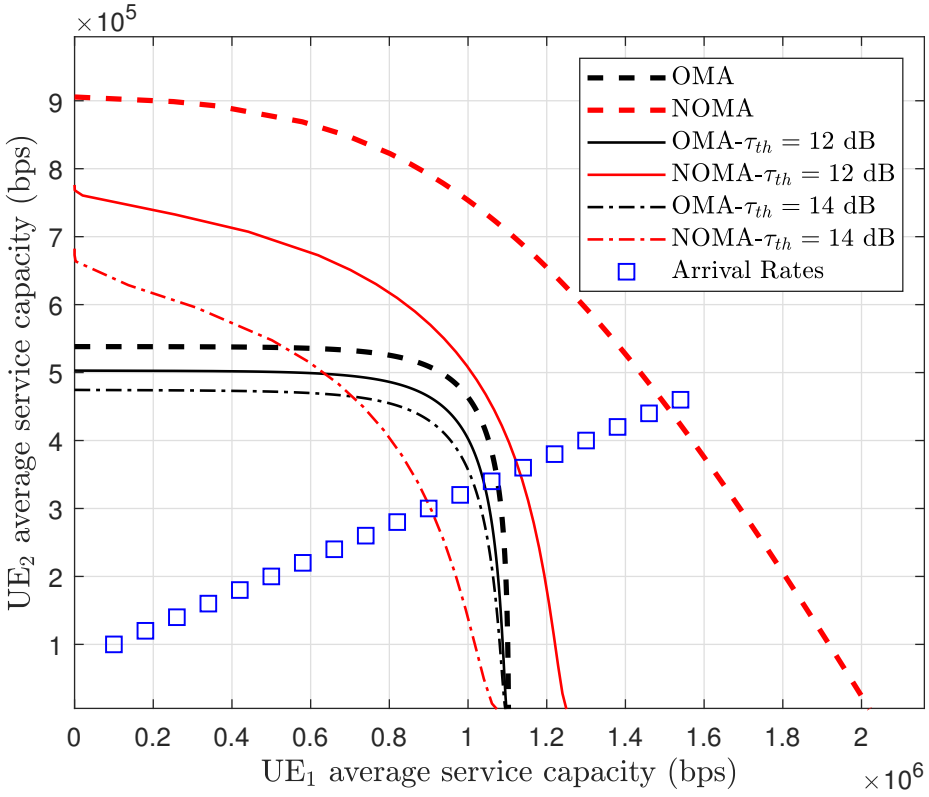


Figure 7.4 : The ergodic capacity regions of OMA and NOMA.

can yield stable queuing dynamics if the proper power allocation is performed. When the SINR outage threshold (τ_{th}) is disabled, NOMA is a superset of OMA in terms of the ergodic capacity region. As the the SINR outage threshold increases the ergodic capacity region of NOMA decreases more than OMA. When τ_{th} is set to 14 dB, the precise superiority of NOMA over OMA can not be observed such that OMA can support more capacity when both UE_1 and UE_2 are jointly considered within a certain part of the ergodic capacity region. Since the bandwidth of the OMA is half of the bandwidth of NOMA, the power of double sided white Gaussian noise is half of NOMA. Therefore, NOMA users have lower SINR levels than OMA users when the same power allocation coefficients are used. As a summary, NOMA users tend to have lower service capacities as τ_{th} gets higher. Another drawback of NOMA for higher SINR threshold scenarios is that it becomes challenging to satisfy the SINR constraint of UE_1 in the SIC process in addition to the decoding process.

Figure 7.5 demonstrates that the MAQD of OMA and NOMA downlink schemes increase when τ_{th} increases from 4 dB to 24 dB. Since the optimum power allocation scheme, which minimizes the maximum average queuing delay for each network

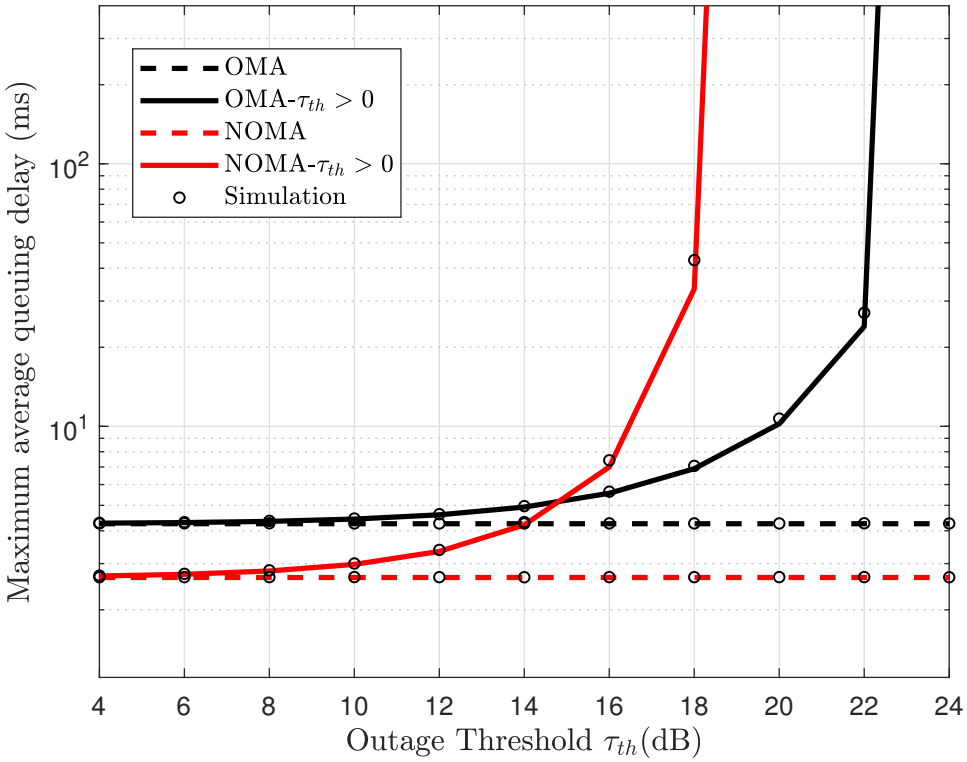


Figure 7.5 : The maximum average queuing delay versus outage threshold.

settings, is employed, the best performances of both multiple access schemes are presented. The higher SINR condition causes higher service times due to the amount of served bits within a time slot is zero when the SINR constraint is not met. Therefore, the MAQD significantly increases for higher SINR threshold conditions. When τ_{th} is disabled, the queuing delay of NOMA is lower due to its higher spectral efficiency. When τ_{th} is enabled, as τ_{th} increases, the MAQD increases for both NOMA and OMA. However, the rate of increase for NOMA is higher than OMA due to the white noise effect. The results show that NOMA provides lower MAQD for the τ_{th} values below 15 dB, while OMA provides lower MAQD for τ_{th} values above 16 dB. Furthermore, the queues are stable for both users when τ_{th} is lower than 18 dB for NOMA, while it is 22 dB for OMA.

Figure 7.6 demonstrates the maximum average queuing delays (MAQD) when the arrival rate of UE_1 and UE_2 varies from 100 to 1540 Kbps with the steps of 80 Kbps and from 100 to 460 Kbps with the steps of 20 Kbps, respectively (i.e., the arrival rates represented by the blue squares in Figure 7.4). The results are provided for the SINR outage threshold (τ_{th}) values of 12 dB and 14 dB in addition to the case

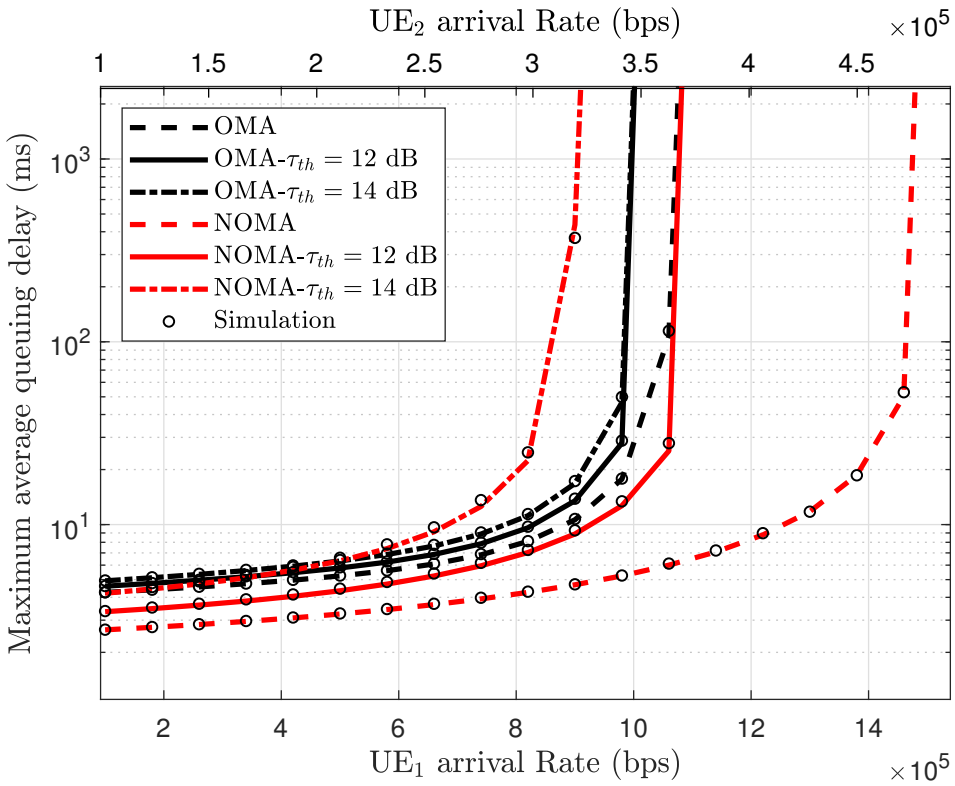


Figure 7.6 : The maximum average queuing delay versus different arrival rates.

that when the outage is disabled. The optimum power allocation coefficients yielding the minimum MAQD is utilized for all scenarios. When the SINR outage threshold is disabled ($\tau_{th} = 0$) or set to 12 dB, NOMA provides lower MAQD than OMA. The queues of UE_1 and UE_2 are stable when the arrival rates are inside the ergodic capacity region. For example, for $\tau_{th} = 0$, both users' queues are stable for NOMA when the UE_1 and UE_2 arrival rates are equal to or lower than 1460 and 440 Kbps while they are stable for OMA for 1060 and 440 Kbps, respectively. On the other hand, for $\tau_{th} = 14$, OMA can support higher arrival rates compared to NOMA for the same MAQD results when the UE_1 and UE_2 arrival rates are beyond 500 and 200 Kbps, respectively.

Figure 7.7 demonstrates that the MAQD of OMA and NOMA downlink schemes increases when the distance of UE_2 from the base station (d_2) increases from 400 m to 2200 m. The reason for this increase is that the average service rate of UE2 decreases and the average queuing delay Q_2 increases when d_2 increases. When the outage constraint is disabled ($\tau_{th} = 0$) NOMA provides lower MAQD than OMA since its higher spectral efficiency. The increase on MAQD for NOMA is higher than OMA when d_2 increases when τ_{th} is enabled. NOMA provides lower MAQD when the d_2

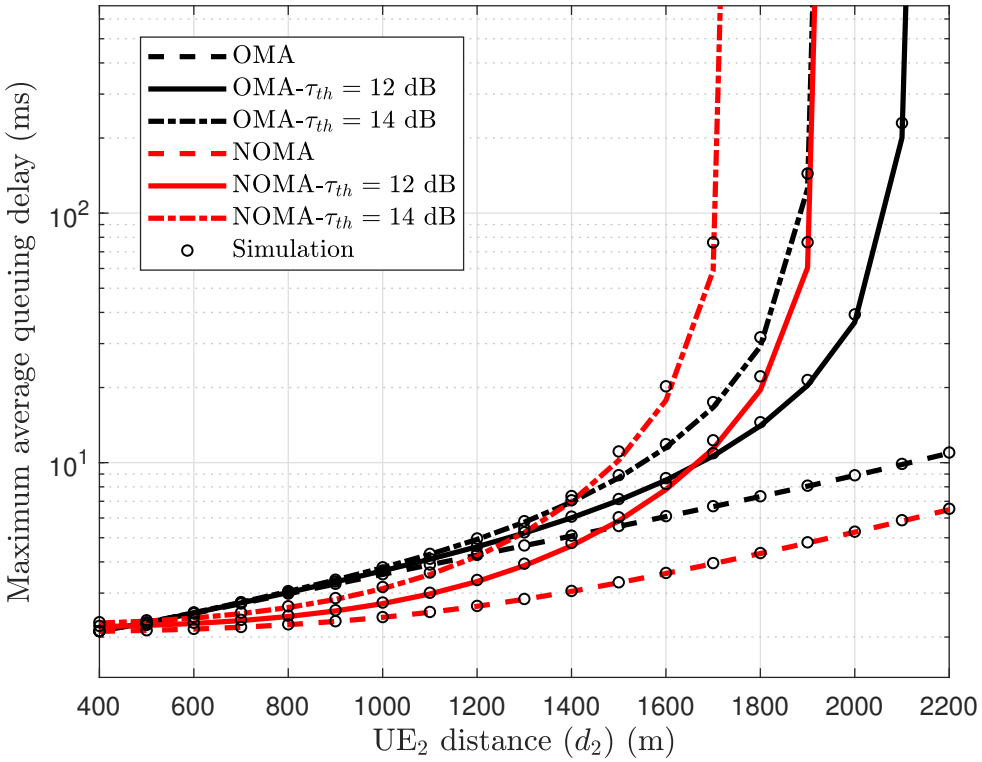


Figure 7.7 : The maximum average queuing delay versus UE$_2$ distance.

value is lower than 1400 m and 1700 m for the SINR outage threshold values (τ_{th}) of 12 dB and 14 dB, respectively. However, when d_2 is greater than 1500 m for $\tau_{th} = 14$ dB and 1800 m for $\tau_{th} = 12$ dB OMA yields lower MAQD. The proposed model is capable of determining a multiple access scheme that achieves the lowest delay for a given network scenario.

7.4.1 Numerical results for 5G NR

The 5G NR, which is based on orthogonal frequency-division multiplexing (OFDM), provides flexibility on the frame structure to support low latency communication. Since a time slot is defined as a fixed number of OFDM symbols, a higher subcarrier spacing leads to a shorter slot duration [124]. Table 5.2 shows subcarrier spacing, RB bandwidth, and time slot duration for different 5G NR frame types, where the frame type 0 corresponds to the LTE setting. The rest of the simulation parameters are used from Table 7.1. Without loss of generality, it is assumed that the same carrier frequency and channel model are used for all 5G NR experiments in this section. The effects of

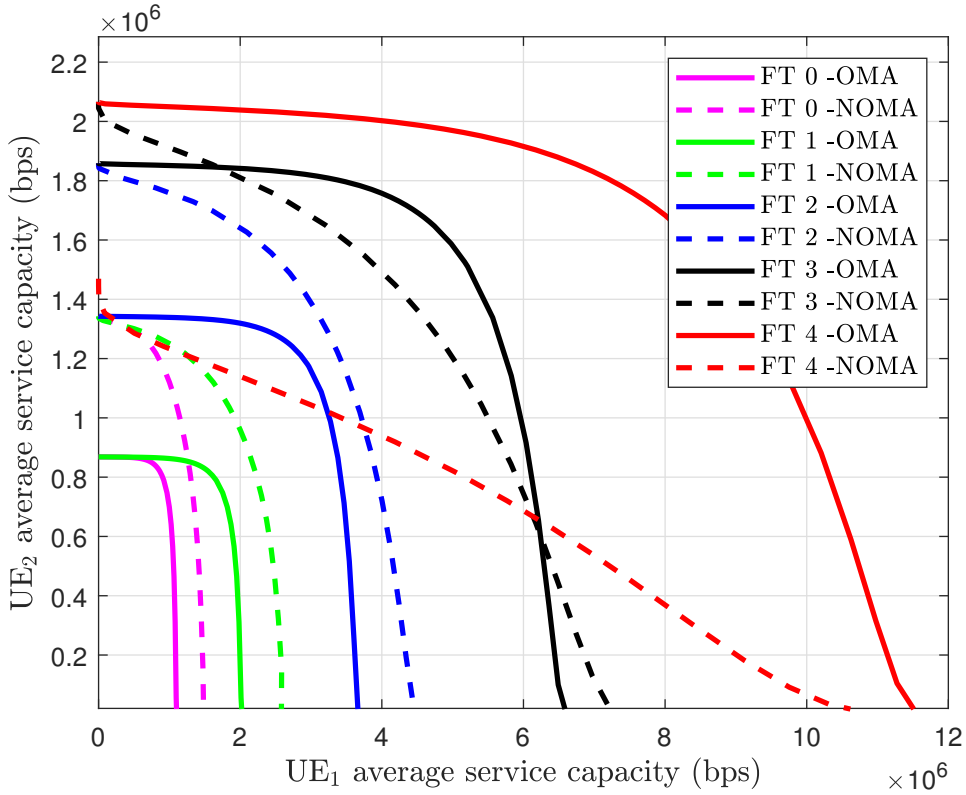


Figure 7.8 : The ergodic capacity regions versus 5G NR frame types.

higher carrier frequencies for wider subcarrier spacing as described in [124] is not within the scope of this work and will be studied as a future work.

The ergodic capacity regions of both NOMA and OMA downlink systems for the 5G NR frame types are shown in Figure 7.8 when the SINR outage threshold τ_{th} is set to 8 dB. Since the noise power increases with the RB bandwidth, the average service capacities for a single time slot decrease when the frame type increases from 0 to 4. Since the number of time slots per second increases when the time slot duration decreases, the average service capacities in terms of bits per second are expected to increase when the frame type varies from 0 to 4. However, the lower SINR levels make the average service capacities significantly decrease when the SINR outage constraint is taken into account. For example, NOMA is a superset of OMA in terms of ergodic capacity region for the 5G NR frame types 0, 1, and 2. When the 5G NR frame type 3 is utilized, there is no precise superiority of ergodic capacity of NOMA over OMA. Furthermore, OMA is preferable when the 5G NR frame type 4 is employed since it provides higher ergodic capacity.

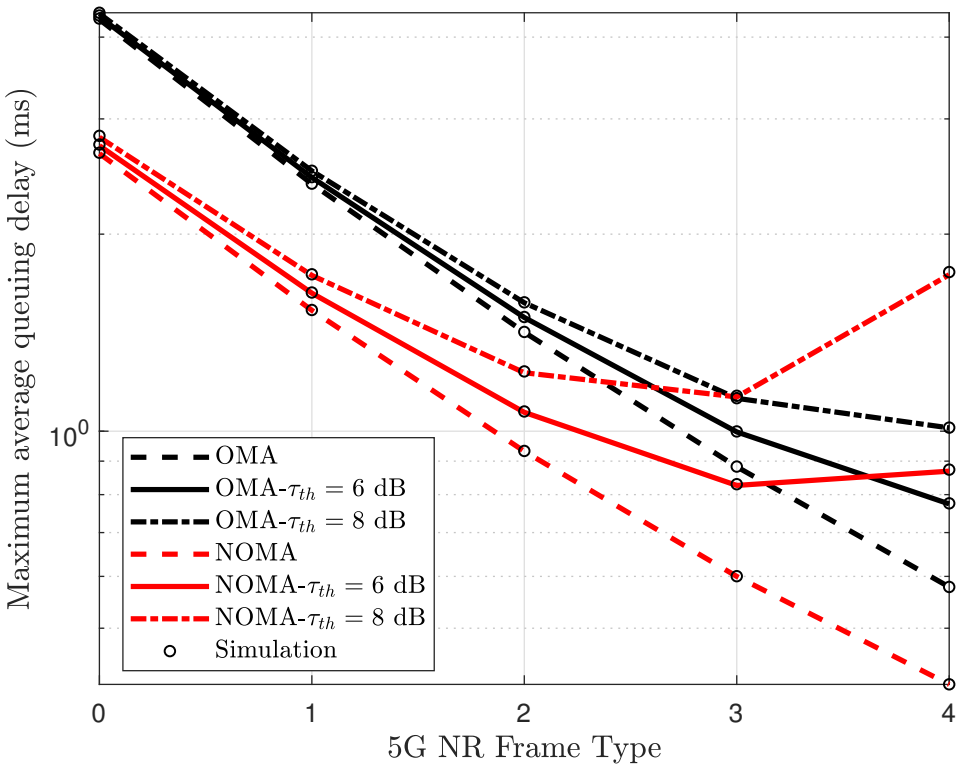


Figure 7.9 : The maximum average queuing delay for 5G NR frame types.

Figure 7.9 shows the MAQD results of 5G NR frame types for different SINR outage thresholds (τ_{th}). When the outage constraint is disabled, the MAQD decreases for both NOMA and OMA when the frame type varies from 0 to 4 while NOMA provides lowest MAQD values. For example, the delay of less than 1 ms can be achieved for NOMA with the 5G NR frame types 2, 3, and 4 while it is observed for OMA with frame types 3 and 4. When τ_{th} is enabled and set to 6 and 8 dB, NOMA provides lower maximum average queuing delays for 5G NR frame types 0, 1, 2, and 3. However, for the 5G NR frame types having wider bandwidth (5G NR frame types 4), the MAQD of NOMA increases and becomes higher than OMA for both τ_{th} values. The results indicate that the delay improvements of 5G NR without the outage constraint and the drawbacks of NOMA for higher outage threshold values can be accurately predicted.

7.5 Summary

In this chapter, the queuing delay performance of a NOMA downlink system is studied by utilizing a discrete time M/G/1 queuing model for the radio access network. The analytical model includes the SINR outage threshold to successfully perform SIC

and decoding procedures at the receiver. The proposed queuing model is utilized within an optimization framework to obtain the lowest queuing delay performances of both NOMA and OMA. The Monte Carlo simulation experiments are performed to numerically validate the model by providing lowest delay results for both NOMA and orthogonal multiple access (OMA) schemes. The numerical results show that NOMA provides higher ergodic capacity region than OMA indicating that NOMA can achieve lower latency when the SINR outage threshold is disabled or set to low values. However, when the SINR outage threshold is set to high levels, the superiority of OMA over NOMA is observed in terms of the ergodic capacity and delay performances. Furthermore, the optimization framework using the analytical model is applied for the performance evaluations of the 5G NR concept when the NOMA is utilized. For higher SINR outage thresholds, the delay performance is highly affected by the frame type due to higher noise effect over wider bandwidth. For a given network scenario including the SINR outage threshold that satisfy reliability requirement of 5G URLLC services, our proposed model is capable of determining the frame type that achieves the lowest delay performance for both NOMA and OMA.

8. CONCLUSION AND FUTURE WORK

In this thesis, we have studied the resource allocation for NOMA downlink schemes in wireless cellular networks. We have first focused on radio resource management of multi user multi resource block NOMA systems by proposing a novel genetic algorithm based approaches under full buffer and the rate limited traffic models. In addition, we have proposed analytical models for NOMA downlink systems which can be utilized to develop resource allocation strategies satisfying the challenging requirements of 5G services.

In Chapter 3, a downlink radio resource management for NOMA system is studied from the multi-user scheduling perspective towards maximizing the geometric mean of user throughputs. Genetic algorithm (GA) approach is proposed to reach the best resource allocation solution, where the power level optimization is employed for each candidate user group. Simulation experiments show that the proposed method quickly converges to the target solution that balances the tradeoff between total system throughput and fairness among users.

NOMA downlink resource allocation including both user scheduling and power allocation is studied under rate limited traffic arrivals in Chapter 4. We first propose two proportional fairness (PF) based user scheduling and power allocation schemes, namely UDB-PF and PUSF, by takein the rate limited user traffic demand requirements into account. UDB-PF extends the PF based scheduling by allocating optimum power levels to satisfy user traffic demand constraints while PUSF maximizes the network-wide user satisfaction. In both schemes, the optimal power level assignment is calculated together with the best user pair selection at each resource block for a given objective function. The GA heuristic is also employed for user group selection at each resource block to reduce the computational complexity. The performance is evaluated by varying the number of users and traffic characteristics of each user. The simulation results show that UDB-PF yields higher sum-rate (throughput) while PUSF provides higher network-wide user satisfaction results compared to the PF based user

scheduling. The performance gains of the proposed methods increase as the variation of user traffic demands increases over time. In addition, when the number of users in the network gets higher, the GA heuristics provide the performance gain on the computational load while the throughput and user satisfaction results are only slightly degraded.

While the proportional fairness based resource allocation approaches assuming the rate-limited traffic demand as a QoS requirement, the delay dynamics can not be studied since the packet-based traffic model with random inter-arrival times and packet sizes are not considered. In Chapter 5, an analytical model utilizing a discrete time M/G/1 queuing model is proposed to characterize the average queuing delay for NOMA downlink systems. The first and second moment statistics of the service time are derived using both packet size and service rate statistics under a Rayleigh fading channel to express the average queuing delay. We perform extensive Monte Carlo simulations and the results verify the accuracy of the proposed analytical model under various network scenarios for both NOMA and OMA schemes. Numerical results show that the ergodic capacity region of NOMA is a superset of OMA indicating that the NOMA can support higher arrival rate and lower latency. Furthermore, the proposed model is applied to demonstrate that the 5G NR frame types having wider bandwidth and shorter duration considerably improves the latency performance. These are promising results such that employing the NOMA technology within the 5G NR concept is a potential enabler to satisfy the challenging latency requirements of time-critical services.

Due to the importance of the outage condition in real-life scenarios of wireless communication systems, in Chapter 6, the outage probability for the NOMA downlink system under the Rayleigh fading channel is analyzed. The optimum power allocation yielding the minimum system outage probability is obtained by solving the convex optimization problem. The Monte Carlo simulations demonstrated that the proposed model can be accurately used to characterize the system outage probability. The results show that OMA has lower system outage probability compared to NOMA. However, the spectral efficiency of NOMA is higher since it allows mutiple users to share the same radio resource.

Then, utilizing the outage probability analysis, the proposed analytical model is extended to support outage constraint in Chapter 7. The extended analytical model includes the SINR outage threshold to successfully perform SIC and decoding procedures at the receiver. The proposed extended queuing model is utilized within an optimization framework to obtain the lowest queuing delay performances of both NOMA and OMA. The numerical results show that NOMA provides higher ergodic capacity region than OMA indicating that NOMA can achieve lower latency when the SINR outage threshold is disabled or set to low values. However, when the SINR outage threshold is set to high levels, the superiority of OMA over NOMA is observed in terms of the ergodic capacity and delay performances. Furthermore, the optimization framework using the extended analytical model is applied for the performance evaluations of the 5G NR concept when the NOMA is utilized. For higher SINR outage thresholds, the delay performance is highly affected by the frame type due to higher noise effect over wider bandwidth. For a given network scenario including the SINR outage threshold that satisfy reliability requirement of 5G URLLC services, our proposed model is capable of determining the frame type that achieves the lowest delay performance for both NOMA and OMA.

8.1 Future Work

A significant reduction in end-to-end latency is one of the major concern for 5G cross-layer radio resource management schemes. The proposed analytical models in this thesis can potentially have a wide range of application scenarios for the delay aware cross-layer radio resource management.

Future work will focus on minimizing the maximum queuing delay of multi-user multi-RB NOMA downlink systems, where optimum user grouping and power level assignment will be performed by extending the M/G/1 queuing model. In addition, a hybrid NOMA and OMA system will be investigated as the proposed model can predict the higher performance regions of these multiple access schemes according to the network settings. The proposed models also allow us to design new radio resource management methods while queuing delay requirements are considered as a QoS constraint to reach different objectives such as sum-rate, energy efficiency, fairness. In the literature, the effective capacity approach has been heavily utilized

to study the performance of resource allocation strategies. As another extension to our work, the resource allocation strategies using both the effective capacity approach and the proposed model can be compared in terms of the latency performances.

Modelling the end-to-end latency of the wireless networks will be a significant contribution and can be provided by combining core-network, propagation, processing, and re-transmission delays with the proposed queuing delay. Utilizing the corresponding channel models and capacity equations, we are planning to investigate the cross-layer designs of NOMA with different technologies such as beamforming, massive MIMO, cooperative communication, mmWave, or visible light communication.

REFERENCES

[1] **Benjebbour, A.**, **Saito, Y.**, **Kishiyama, Y.**, **Li, A.**, **Harada, A. and Nakamura, T.** (2013). Concept and Practical Considerations of Non-orthogonal Multiple Access (NOMA) for Future Radio Access, *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, IEEE, pp.770–774.

[2] **Saito, Y.**, **Kishiyama, Y.**, **Benjebbour, A.**, **Nakamura, T.**, **Li, A. and Higuchi, K.** (2013). Non-orthogonal Multiple Access (NOMA) for Cellular Future Radio Access, *2013 IEEE 77th vehicular technology conference (VTC Spring)*, IEEE, pp.1–5.

[3] **Tao, Y.**, **Liu, L.**, **Liu, S. and Zhang, Z.** (2015). A Survey: Several Technologies of Non-orthogonal Transmission for 5G, *China communications*, *12*(10), 1–15.

[4] **Jänis, P.**, **Koivunen, V. and Ribeiro, C.B.** (2011). Interference-aware Radio Resource Management for Local Area Wireless Networks, *EURASIP Journal on Wireless Communications and Networking*, *2011*, 1–15.

[5] **Otao, N.**, **Kishiyama, Y. and Higuchi, K.** (2012). Performance of Non-orthogonal Access with SIC in Cellular Downlink Using Proportional Fair-based Resource Allocation, *2012 international symposium on wireless communication systems (ISWCS)*, IEEE, pp.476–480.

[6] **Parida, P. and Das, S.S.** (2014). Power allocation in OFDM based NOMA systems: A DC programming approach, *2014 IEEE Globecom Workshops (GC Wkshps)*, IEEE, pp.1026–1031.

[7] **He, J.**, **Tang, Z. and Che, Z.** (2016). Fast and Efficient User Pairing and Power Allocation Algorithm for Non-orthogonal Multiple Access in Cellular Networks, *Electronics letters*, *52*(25), 2065–2067.

[8] **Di, B.**, **Song, L. and Li, Y.** (2016). Sub-channel Assignment, Power Allocation, and User Scheduling for Non-orthogonal Multiple Access Networks, *IEEE Transactions on Wireless Communications*, *15*(11), 7686–7698.

[9] **Lei, L.**, **Yuan, D.**, **Ho, C.K. and Sun, S.** (2016). Power and Channel Allocation for Non-orthogonal Multiple Access in 5G systems: Tractability and Computation, *IEEE Transactions on Wireless Communications*, *15*(12), 8580–8594.

[10] **Parvez, I.**, **Rahmati, A.**, **Guvenc, I.**, **Sarwat, A.I. and Dai, H.** (2018). A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions, *IEEE Communications Surveys Tutorials*, *20*(4), 3098–3130.

[11] **Siddiqi, M.A.**, **Yu, H. and Joung, J.** (2019). 5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices, *Electronics*, *8*(9), 981.

[12] **Bennis, M.**, **Debbah, M. and Poor, H.V.** (2018). Ultra-reliable and Low-Latency Wireless Communication: Tail, Risk, and Scale, *Proceedings of the IEEE*, *106*(10), 1834–1853.

[13] **Feng, D.**, **She, C.**, **Ying, K.**, **Lai, L.**, **Hou, Z.**, **Quek, T.Q.**, **Li, Y. and Vucetic, B.** (2019). Toward Ultra-reliable Low-Latency Communications: Typical Scenarios, Possible Solutions, and Open Issues, *IEEE Vehicular Technology Magazine*, *14*(2), 94–102.

[14] **Zhou, Y. and Wong, V.W.** (2017). Stable Throughput Region of Downlink NOMA Transmissions with Limited CSI, *2017 IEEE International Conference on Communications (ICC)*, IEEE, pp.1–7.

[15] **Amjad, M. and Musavian, L.** (2018). Performance Analysis of NOMA for Ultra-reliable and Low-latency Communications, *2018 IEEE Globecom Workshops (GC Wkshps)*, IEEE, pp.1–5.

[16] **Salehi, F.**, **Neda, N. and Majidi, M.H.** (2020). Max-min fairness in downlink non-orthogonal multiple access with short packet communications, *AEU-International Journal of Electronics and Communications*, *114*, 153028.

[17] **Yu, W.**, **Musavian, L. and Ni, Q.** (2018). Link-layer Capacity of NOMA under Statistical Delay QoS Guarantees, *IEEE Transactions on Communications*, *66*(10), 4907–4922.

[18] **Xiao, C.**, **Zeng, J.**, **Ni, W.**, **Liu, R.P.**, **Su, X. and Wang, J.** (2019). Delay Guarantee and Effective Capacity of Downlink NOMA Fading Channels, *IEEE Journal of Selected Topics in Signal Processing*, *13*(3), 508–523.

[19] **Zhao, X. and Chen, W.** (2019). Non-Orthogonal Multiple Access for Delay-Sensitive Communications: A Cross-Layer Approach, *IEEE Transactions on Communications*.

[20] **Benjebbovu, A.**, **Li, A.**, **Saito, Y.**, **Kishiyama, Y.**, **Harada, A. and Nakamura, T.** (2013). System-level Performance of Downlink NOMA for Future LTE Enhancements, *2013 IEEE Globecom Workshops (GC Wkshps)*, IEEE, pp.66–70.

[21] **Viswanathan, H. and Weldon, M.** (2014). The Past, Present, and Future of Mobile Communications, *Bell Labs Technical Journal*, *19*, 8–21.

[22] **Huff, D.** (1979). Advanced Mobile Phone Service: The Developmental System, *Bell System Technical Journal*, *58*(1), 249–269.

[23] **Mehrotra, A.** (1997). *GSM System Engineering*, Artech House, Inc.

[24] **Holma, H. and Toskala, A.** (2005). *WCDMA for UMTS: Radio access for third generation mobile communications*, john wiley & sons.

[25] **Ghosh, A.**, **Zhang, J.**, **Andrews, J.G. and Muhamed, R.** (2010). *Fundamentals of LTE*, Pearson Education.

[26] **Yuan, Y. and Yan, C.** (2018). NOMA Study in 3GPP for 5G, *2018 IEEE 10th International Symposium on Turbo Codes & Iterative Information Processing (ISTC)*, IEEE, pp.1–5.

[27] **Liu, Y.**, **Qin, Z.**, **Elkashlan, M.**, **Ding, Z.**, **Nallanathan, A. and Hanzo, L.** (2017). Nonorthogonal Multiple Access for 5G and Beyond, *Proceedings of the IEEE*, *105*(12), 2347–2381.

[28] 3GPP TD RP-150496: "Study on Downlink Multiuser Superposition Transmission".

[29] **Yuan, Y.**, **Yuan, Z.**, **Yu, G.**, **Hwang, C.h.**, **Liao, P.k.**, **Li, A. and Takeda, K.** (2016). Non-orthogonal transmission technology in LTE evolution, *IEEE Communications Magazine*, *54*(7), 68–74.

[30] **Ge, X.** (2019). Ultra-reliable Low-latency Communications in Autonomous Vehicular Networks, *IEEE Transactions on Vehicular Technology*, *68*(5), 5005–5016.

[31] **Chang, B.**, **Zhang, L.**, **Li, L.**, **Zhao, G. and Chen, Z.** (2019). Optimizing Resource Allocation in URLLC for Real-time Wireless Control Systems, *IEEE Transactions on Vehicular Technology*, *68*(9), 8916–8927.

[32] **Saito, Y.**, **Benjebbour, A.**, **Kishiyama, Y. and Nakamura, T.** (2013). System-level Performance Evaluation of Downlink Non-orthogonal Multiple Access (NOMA), *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, IEEE, pp.611–615.

[33] **Hoshyar, R.**, **Wathan, F.P. and Tafazolli, R.** (2008). Novel low-density signature for synchronous CDMA systems over AWGN channel, *IEEE Transactions on Signal Processing*, *56*(4), 1616–1626.

[34] **Nikopour, H. and Baligh, H.** (2013). Sparse code multiple access, *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, IEEE, pp.332–336.

[35] **Zeng, J.**, **Li, B.**, **Su, X.**, **Rong, L. and Xing, R.** (2015). Pattern division multiple access (PDMA) for cellular future radio access, *2015 international conference on wireless communications & signal processing (WCSP)*, IEEE, pp.1–5.

[36] **Yuan, Z.**, **Yu, G.**, **Li, W.**, **Yuan, Y.**, **Wang, X. and Xu, J.** (2016). Multi-user shared access for internet of things, *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, IEEE, pp.1–5.

[37] **Dai, L.**, **Wang, B.**, **Yuan, Y.**, **Han, S.**, **Chih-Lin, I. and Wang, Z.** (2015). Non-orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends, *IEEE Communications Magazine*, *53*(9), 74–81.

[38] **Ding, Z.**, **Yang, Z.**, **Fan, P. and Poor, H.V.** (2014). On the Performance of Non-orthogonal Multiple Access in 5G Systems with Randomly Deployed Users, *IEEE signal processing letters*, *21*(12), 1501–1505.

[39] **Di, B.**, **Bayat, S.**, **Song, L. and Li, Y.** (2015). Radio Resource Allocation for Downlink Non-orthogonal Multiple Access (NOMA) Networks Using Matching Theory, *2015 IEEE global communications conference (GLOBECOM)*, IEEE, pp.1–6.

[40] **Datta, S.N. and Kalyanasundaram, S.** (2016). Optimal Power Allocation and User Selection in Non-orthogonal Multiple Access Systems, *2016 IEEE Wireless Communications and Networking Conference*, IEEE, pp.1–6.

[41] **Wang, C.L.**, **Chen, J.Y. and Chen, Y.J.** (2016). Power Allocation for a Downlink Non-orthogonal Multiple Access System, *IEEE wireless communications letters*, *5*(5), 532–535.

[42] **Al-Abbasi, Z.Q. and So, D.K.** (2015). Power Allocation for Sum Rate Maximization in Non-orthogonal Multiple Access System, *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, IEEE, pp.1649–1653.

[43] **Choi, J.** (2016). Power Allocation for Max-sum Rate and Max-min Rate Proportional Fairness in NOMA, *IEEE Communications Letters*, *20*(10), 2055–2058.

[44] **Ding, Z.**, **Fan, P. and Poor, H.V.** (2015). User Pairing in Non-orthogonal Multiple Access Downlink Transmissions, *2015 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pp.1–5.

[45] **Zhang, H.**, **Zhang, D.K.**, **Meng, W.X. and Li, C.** (2016). User Pairing Algorithm with SIC in Non-orthogonal Multiple Access System, *2016 IEEE International Conference on Communications (ICC)*, IEEE, pp.1–6.

[46] **Shahab, M.B.**, **Kader, M.F. and Shin, S.Y.** (2016). A Virtual User Pairing Scheme to Optimally Utilize the Spectrum of Unpaired Users in Non-orthogonal Multiple Access, *IEEE Signal Processing Letters*, *23*(12), 1766–1770.

[47] **Liang, W.**, **Ding, Z.**, **Li, Y. and Song, L.** (2017). User Pairing for Downlink Non-orthogonal Multiple Access Networks Using Matching Algorithm, *IEEE Transactions on Communications*, *65*(12), 5319–5332.

[48] **Kang, J.M. and Kim, I.M.** (2018). Optimal User Grouping for Downlink NOMA, *IEEE Wireless Communications Letters*, *7*(5), 724–727.

[49] **Liu, F.**, **Mähönen, P. and Petrova, M.** (2015). Proportional Fairness-based User Pairing and Power Allocation for Non-orthogonal Multiple Access, *2015 IEEE 26th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, IEEE, pp.1127–1131.

[50] **Higuchi, K. and Benjebbour, A.** (2015). Non-orthogonal Multiple Access (NOMA) with Successive Interference Cancellation for Future Radio Access, *IEICE Transactions on Communications*, *98*(3), 403–414.

[51] **Umehara, J.**, **Kishiyama, Y. and Higuchi, K.** (2012). Enhancing User Fairness in Non-orthogonal Access with Successive Interference Cancellation for Cellular Downlink, *2012 IEEE International Conference on Communication Systems (ICCS)*, IEEE, pp.324–328.

[52] **Wang, K.**, **Cui, J.**, **Ding, Z. and Fan, P.** (2019). Stackelberg game for user clustering and power allocation in millimeter wave-NOMA systems, *IEEE Transactions on Wireless Communications*, *18*(5), 2842–2857.

[53] **Zhang, H.**, **Wang, B.**, **Jiang, C.**, **Long, K.**, **Nallanathan, A.**, **Leung, V.C. and Poor, H.V.** (2018). Energy Efficient Dynamic Resource Optimization in NOMA System, *IEEE Transactions on Wireless Communications*, *17*(9), 5671–5683.

[54] **Wei, Z.**, **Ng, D.W.K.**, **Yuan, J. and Wang, H.M.** (2017). Optimal Resource Allocation for Power-efficient MC-NOMA with Imperfect Channel State Information, *IEEE Transactions on Communications*, *65*(9), 3944–3961.

[55] **Wang, W.**, **Liu, Y.**, **Luo, Z.**, **Jiang, T.**, **Zhang, Q. and Nallanathan, A.** (2018). Toward Cross-layer Design for Non-orthogonal Multiple Access: A Quality-of-experience Perspective, *IEEE Wireless Communications*, *25*(2), 118–124.

[56] **Saito, Y.**, **Benjebbour, A.**, **Li, A.**, **Takeda, K.**, **Kishiyama, Y. and Nakamura, T.** (2015). System-level Evaluation of Downlink Non-orthogonal Nultiple Access (NOMA) for Non-full Buffer Traffic Model, *2015 IEEE Conference on Standards for Communications and Networking (CSCN)*, IEEE, pp.94–99.

[57] **Liu, Y.**, **Ding, Z.**, **Eïkashlan, M. and Poor, H.V.** (2015). Cooperative Non-orthogonal Multiple Access in 5G Systems with SWIPT, *2015 23rd European signal processing conference (EUSIPCO)*, IEEE, pp.1999–2003.

[58] **Sun, Q.**, **Han, S.**, **Chin-Lin, I. and Pan, Z.** (2015). On the Ergodic Capacity of MIMO NOMA systems, *IEEE Wireless Communications Letters*, *4*(4), 405–408.

[59] **Kimy, B.**, **Lim, S.**, **Kim, H.**, **Suh, S.**, **Kwun, J.**, **Choi, S.**, **Lee, C.**, **Lee, S. and Hong, D.** (2013). Non-orthogonal Multiple Access in a Downlink Multiuser Beamforming System, *MILCOM 2013-2013 IEEE Military Communications Conference*, IEEE, pp.1278–1283.

[60] **He, G.**, **Li, L.**, **Li, X.**, **Chen, W.**, **Yang, L.L. and Han, Z.** (2017). Secrecy Sum Rate Maximization in NOMA Systems with Wireless Information and Power Transfer, *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, IEEE, pp.1–6.

[61] **Diamantoulakis, P.D.**, **Pappi, K.N.**, **Ding, Z. and Karagiannidis, G.K.** (2016). Optimal Design of Non-orthogonal Multiple Access with Wireless Power Transfer, *2016 IEEE international conference on communications (ICC)*, IEEE, pp.1–6.

[62] **Liu, Y.**, **Qin, Z.**, **Elkashlan, M.**, **Gao, Y. and Hanzo, L.** (2017). Enhancing the Physical Layer Security of Non-orthogonal Multiple Access in Large-scale Networks, *IEEE Transactions on Wireless Communications*, *16*(3), 1656–1672.

[63] **Su, X.**, **Nkurunziza, P.**, **Gu, J.**, **Castiglione, A. and Choi, C.** (2017). Inter-beam Interference Cancellation and Physical Layer Security Constraints by 3D Polarized Beamforming in Power Domain NOMA Systems, *IEEE Transactions on Sustainable Computing*.

[64] **Zhang, H.**, **Yang, N.**, **Long, K.**, **Pan, M.**, **Karagiannidis, G.K. and Leung, V.C.** (2018). Secure Communications in NOMA System: Subcarrier Assignment and Power Allocation, *IEEE Journal on Selected Areas in Communications*, *36*(7), 1441–1452.

[65] **Ding, Z.**, **Fan, P. and Poor, H.V.** (2015). Impact of User Pairing on 5G Non-orthogonal Multiple-access Downlink Transmissions, *IEEE Transactions on Vehicular Technology*, *65*(8), 6010–6023.

[66] **Yang, Z.**, **Ding, Z.**, **Fan, P. and Al-Dhahir, N.** (2016). A General Power Allocation Scheme to Guarantee Quality of Service in Downlink and Uplink NOMA Systems, *IEEE transactions on wireless communications*, *15*(11), 7244–7257.

[67] **Fang, F.**, **Zhang, H.**, **Cheng, J. and Leung, V.C.** (2016). Energy-efficient Resource Allocation for Downlink Non-orthogonal Multiple Access Network, *IEEE Transactions on Communications*, *64*(9), 3722–3732.

[68] **Fang, F.**, **Zhang, H.**, **Cheng, J.**, **Roy, S. and Leung, V.C.** (2017). Joint User Scheduling and Power Allocation Optimization for Energy-efficient NOMA Systems with Imperfect CSI, *IEEE Journal on Selected Areas in Communications*, *35*(12), 2874–2885.

[69] **Cui, J.**, **Ding, Z. and Fan, P.** (2016). A Novel Power Allocation Scheme Under Outage Constraints in NOMA Systems, *IEEE Signal Processing Letters*, *23*(9), 1226–1230.

[70] **Cui, J.**, **Ding, Z. and Fan, P.** (2016). A Novel Power Allocation Scheme under Outage Constraints in NOMA Systems, *IEEE Signal Processing Letters*, *23*(9), 1226–1230.

[71] **Tang, Z.**, **Wang, J.**, **Wang, J. and Song, J.** (2018). On the Achievable Rate Region of NOMA Under Outage Probability Constraints, *IEEE Communications Letters*, *23*(2), 370–373.

[72] **Do, D.T. and Van Nguyen, M.S.** (2019). Outage Probability and Ergodic Capacity Analysis of Uplink NOMA Cellular Network with and without Interference from D2D Pair, *Physical Communication*, 100898.

[73] **Chu, T.M.C. and Zepernick, H.J.** (2017). Outage Probability and Secrecy Capacity of a Non-orthogonal Multiple Access System, *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, IEEE, pp.1–6.

[74] **Trillingsgaard, K.F. and Popovski, P.** (2017). Generalized HARQ Protocols with Delayed Channel State Information and Average Latency Constraints, *IEEE Transactions on Information Theory*, *64*(2), 1262–1280.

[75] **Ding, Z.**, **Yang, Z.**, **Fan, P. and Poor, H.V.** (2014). On the Performance of Non-orthogonal Multiple Access in 5G Systems with Randomly Deployed Users, *IEEE Signal Processing Letters*, *21*(12), 1501–1505.

[76] **Wang, X.**, **Wang, J.**, **He, L. and Song, J.** (2017). Outage Analysis for Downlink NOMA with Statistical Channel State Information, *IEEE Wireless Communications Letters*, *7*(2), 142–145.

[77] **Xu, P.**, **Yuan, Y.**, **Ding, Z.**, **Dai, X. and Schober, R.** (2016). On the Outage Performance of Non-orthogonal Multiple Access with 1-bit Feedback, *IEEE Transactions on Wireless Communications*, *15*(10), 6716–6730.

[78] **Zhang, W.**, **Kotagiri, S.P. and Laneman, J.N.** (2009). On Downlink Transmission Without Transmit Channel State Information and With Outage Constraints, *IEEE Transactions on Information Theory*, *55*(9), 4240–4248.

[79] **Wang, X.**, **Wang, J.**, **He, L. and Song, J.** (2017). Outage Analysis for Downlink NOMA with Statistical Channel State Information, *IEEE Wireless Communications Letters*, *7*(2), 142–145.

[80] **Zhang, W.**, **Kotagiri, S.P. and Laneman, J.N.** (2009). On Downlink Transmission without Transmit Channel State Information and with Outage Constraints, *IEEE transactions on information theory*, *55*(9), 4240–4248.

[81] **Choi, J.** (2017). Effective Capacity of NOMA and a Suboptimal Power Control Policy with Delay QoS, *IEEE Transactions on Communications*, *65*(4), 1849–1858.

[82] **Chen, Y.**, **Wu, K. and Zhang, Q.** (2014). From QoS to QoE: A tutorial on video quality assessment, *IEEE Communications Surveys & Tutorials*, *17*(2), 1126–1165.

[83] **Cui, J.**, **Liu, Y.**, **Ding, Z.**, **Fan, P. and Nallanathan, A.** (2018). QoE-based resource allocation for multi-cell NOMA networks, *IEEE Transactions on Wireless Communications*, *17*(9), 6160–6176.

[84] **Islam, S.R.**, **Avazov, N.**, **Dobre, O.A. and Kwak, K.S.** (2016). Power-domain Non-orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges, *IEEE Communications Surveys & Tutorials*, *19*(2), 721–742.

[85] **Fang, Z.**, **Li, J. and Lu, Y.** (2020). Cooperative non-orthogonal multiple access for two-way relay networks, *AEU-International Journal of Electronics and Communications*, *115*, 153021.

[86] **Ding, Z.**, **Dai, H. and Poor, H.V.** (2016). Relay Selection for Cooperative NOMA, *IEEE Wireless Communications Letters*, *5*(4), 416–419.

[87] **Yuan, C.**, **Tao, X.**, **Li, N.**, **Ni, W.**, **Liu, R.P. and Zhang, P.** (2019). Analysis on Secrecy Capacity of Cooperative Non-orthogonal Multiple Access with Proactive Jamming, *IEEE Transactions on Vehicular Technology*, *68*(3), 2682–2696.

[88] **Deng, P.**, **Wu, W.**, **Shen, X.**, **Li, P. and Wang, B.** (2019). Precoding Design of NOMA-enabled D2D Communication System with Low Latency, *EURASIP Journal on Wireless Communications and Networking*, *2019*(1), 185.

[89] **Ali, S.**, **Hossain, E. and Kim, D.I.** (2016). Non-orthogonal Multiple Access (NOMA) for Downlink Multiuser MIMO Systems: User Clustering, Beamforming, and Power Allocation, *IEEE access*, *5*, 565–577.

[90] **Chinnadurai, S.**, **Selvaprabhu, P. and Lee, M.H.** (2017). A Novel Joint User Pairing and Dynamic Power Allocation Scheme in MIMO-noma System, *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, pp.951–953.

[91] **Ding, Z.**, **Fan, P. and Poor, H.V.** (2017). Random Beamforming in Millimeter-wave NOMA Networks, *IEEE access*, *5*, 7667–7681.

[92] **Liu, Y.**, **Ding, Z.**, **Elkashlan, M. and Poor, H.V.** (2016). Cooperative Non-orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer, *IEEE Journal on Selected Areas in Communications*, *34*(4), 938–953.

[93] **Ye, Y.**, **Li, Y.**, **Wang, D. and Lu, G.** (2017). Power Splitting Protocol Design for the Cooperative NOMA with SWIPT, *2017 IEEE International Conference on Communications (ICC)*, IEEE, pp.1–5.

[94] **Alsaba, Y.**, **Leow, C.Y. and Rahim, S.K.A.** (2018). Full-duplex Cooperative Non-orthogonal Multiple Access with Beamforming and Energy Harvesting, *IEEE Access*, *6*, 19726–19738.

[95] **Kizilirmak, R.C.**, **Rowell, C.R. and Uysal, M.** (2015). Non-orthogonal Multiple Access (NOMA) for Indoor Visible Light Communications, *2015 4th international workshop on optical wireless communications (IWOW)*, IEEE, pp.98–101.

[96] **Fu, Y.**, **Hong, Y.**, **Chen, L.K. and Sung, C.W.** (2018). Enhanced Power Allocation for Sum Rate Maximization in OFDM-NOMA VLC systems, *IEEE Photonics Technology Letters*, *30*(13), 1218–1221.

[97] **Zhang, X.**, **Gao, Q.**, **Gong, C. and Xu, Z.** (2016). User Grouping and Power Allocation for NOMA Visible Light Communication Multi-cell Networks, *IEEE communications letters*, *21*(4), 777–780.

[98] **Wei, Z.**, **Guo, J.**, **Ng, D.W.K. and Yuan, J.** (2017). Fairness Comparison of Uplink NOMA and OMA, *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, IEEE, pp.1–6.

[99] **Zeng, M.**, **Yadav, A.**, **Dobre, O.A. and Poor, H.V.** (2019). Energy-efficient Joint User-RB Association and Power Allocation for Uplink Hybrid NOMA-OMA, *IEEE Internet of Things Journal*, *6*(3), 5119–5131.

[100] **Nomikos, N.**, **Charalambous, T.**, **Vouyioukas, D.**, **Karagiannidis, G.K. and Wichman, R.** (2019). Hybrid NOMA/OMA with Buffer-aided Relay Selection in Cooperative networks, *IEEE Journal of Selected Topics in Signal Processing*, *13*(3), 524–537.

[101] **Yang, Z.**, **Ding, Z.**, **Fan, P. and Ma, Z.** (2016). Outage Performance for Dynamic Power Allocation in Hybrid Non-orthogonal Multiple Access Systems, *IEEE Communications Letters*, *20*(8), 1695–1698.

[102] **Wu, D. and Negi, R.** (2003). Effective Capacity: A Wireless Link Model for Support of Quality of Service, *IEEE Transactions on wireless communications*, *2*(4), 630–643.

[103] **Liu, G.**, **Ma, Z.**, **Chen, X.**, **Ding, Z.**, **Yu, F.R. and Fan, P.** (2017). Cross-layer Power Allocation in Non-orthogonal Multiple Access Systems for Statistical QoS Provisioning, *IEEE Transactions on Vehicular Technology*, *66*(12), 11388–11393.

[104] **Gan, K.**, **Chen, Q.**, **Shen, X. and Nie, Y.** (2019). Energy Efficient Short-packet Downlink Transmission with Non-orthogonal Multiple Access, *Physical Communication*, *37*, 100839.

[105] **Schiessl, S.**, **Skoglund, M. and Gross, J.** (2019). NOMA in the Uplink: Delay Analysis with Imperfect CSI and Finite-Length Coding, *arXiv preprint arXiv:1903.09586*.

[106] **Liu, L.**, **Sheng, M.**, **Liu, J.**, **Dai, Y. and Li, J.** (2019). Stable Throughput Region and Average Delay Analysis of Uplink NOMA Systems with Unsaturated Traffic, *IEEE Transactions on Communications*.

[107] **Yang, J. and Ulukus, S.** (2009). Delay Minimization in Multiple Access Channels, *2009 IEEE International Symposium on Information Theory*, IEEE, pp.2366–2370.

[108] **Dong, Y. and Fan, P.** (2013). Queueing analysis for block fading Rayleigh channels in the low SNR regime, *2013 International Conference on Wireless Communications and Signal Processing*, IEEE, pp.1–6.

[109] **Wunder, G. and Zhou, C.** (2009). Queueing Analysis for the OFDMA Downlink: Throughput Regions, Delay and Exponential Backlog Bounds, *IEEE Transactions on Wireless Communications*, *8*(2), 871–881.

[110] **Li, C.P.**, **Jiang, J.**, **Chen, W.**, **Ji, T. and Smee, J.** (2017). 5G Ultra-reliable and Low-latency Systems Design, *2017 European Conference on Networks and Communications (EuCNC)*, IEEE, pp.1–5.

[111] **Gemici, O.F.**, **Hokelek, I. and Çırpan, H.A.** (2014). GA Based Multi-objective LTE Scheduler, *2014 1st International Workshop on Cognitive Cellular Systems (CCS)*, IEEE, pp.1–5.

[112] **Michalewicz, Z.** (2013). *Genetic Algorithms + Data Structures = Evolution Programs*, Springer Science & Business Media.

[113] **Davis, L.** (1991). Handbook of Genetic Algorithms.

[114] **Li, A.**, **Lan, Y.**, **Chen, X. and Jiang, H.** (2015). Non-orthogonal Multiple Access (NOMA) for Future Downlink Radio Access of 5G, *China Communications*, *12*(Supplement), 28–37.

[115] **Kountouris, M. and Gesbert, D.** (2005). Memory-based Opportunistic Multi-user Beamforming, *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, IEEE, pp.1426–1430.

[116] **Wang, B.**, **Wen, X.**, **Su, D. and Zheng, W.** (2010). User Satisfaction Based Resource Allocation for OFDMA Relay Networks in the Resource-constrained System, *2010 Second International Conference on Future Networks*, IEEE, pp.304–308.

[117] **Gemici, Ö.F.**, **Kara, F.**, **Hökelek, İ. and Çırpan, H.A.** (2019). User Scheduling and Power Allocation for Nonfull-buffer Traffic in NOMA Downlink Systems, *International Journal of Communication Systems*, *32*(1), e3834.

[118] **Gemici, Ö.F.**, **Kara, F.**, **Hokelek, I.**, **Kurt, G.K. and Çırpan, H.A.** (2017). Resource Allocation for NOMA Downlink Systems: Genetic Algorithm Approach, *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, pp.114–118.

[119] **Kara, F.**, **Gemici, Ö.F.**, **Hökelek, İ. and Çirpan, H.A.** (2017). Optimal Power Allocation for DL NOMA Systems, *2017 25th Signal Processing and Communications Applications Conference (SIU)*, IEEE, pp.1–4.

[120] **Lee, J. and Tepedelenlioğlu, C.** (2013). Stochastic Ordering of Interference in Large-scale Wireless Networks, *IEEE Transactions on Signal Processing*, *62*(3), 729–740.

[121] **Andrews, L.C.** (1998). *Special Functions of Mathematics for Engineers*, Spie Press, Washington.

[122] **Brémaud, P.** (2013). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer Science & Business Media, New York.

[123] **Chan, W.**, **Lu, T.C. and Chen, R.J.** (1997). Pollaczek-Khinchin Formula for the M/G/1 Queue in Discrete Time with Vacations, *IEE Proceedings-Computers and Digital Techniques*, *144*(4), 222–226.

[124] **Parkvall, S.**, **Dahlman, E.**, **Furuskar, A. and Frenne, M.** (2017). NR: The New 5G Radio Access Technology, *IEEE Communications Standards Magazine*, *1*(4), 24–30.

[125] **Gemici, Ö.F.**, **Hökelek, İ. and Çırpan, H.A.** (2019). NOMA Power Allocation for Minimizing System Outage under Rayleigh Fading Channel, *2019 IEEE 40th Sarnoff Symposium*, IEEE, pp.1–6.

[126] **Andrews, L.C. and Andrews, L.C.** (1992). *Special Functions of Mathematics for Engineers*, McGraw-Hill New York.

[127] **Egrioglu, E.**, **Aladag, C.H.**, **Basaran, M.A.**, **Yolcu, U. and Uslu, V.R.** (2011). A New Approach Based on the Optimization of the Length of Intervals in Fuzzy Time Series, *Journal of Intelligent & Fuzzy Systems*, *22*(1), 15–19.

APPENDICES

APPENDIX A: Derivation of the Service Time Statistics

APPENDIX A: Derivation of the Second Moment of the Service Time

The solution of equation (5.29) in Section 5.3 is provided by enabling the generating functions (g.f.) of random variables as described in [122]. The generating function of random variable Y is defined as:

$$g_Y(z) = \mathrm{E}\left[z^l\right] = \sum_{k=1}^{\infty} \mathrm{P}[l=k]\, z^k \tag{A.1}$$

where $z \in \bar{D}(0:1)$ while $\bar{D}(0:1)$ is the complex closed unit disk centered at 0. Similarly, the g.f. of S and R random variables represented as $g_S(z)$ and $g_R(z)$, respectively. By computing the generating function of the random sums of i.i.d. random variables problem defined in equation (5.29), the final representation of the problem becomes:

$$g_Y(z) = g_S(g_R(z)) \ . \tag{A.2}$$

It is also provided in [122] that the domain definition of $g_Y(z)$ contains the open unit disk and the differentiation is possible inside the open disk. Therefore, let $z \to 1$, the first and second derivation of the $g_Y(z)$ satisfy the following equations:

$$\begin{aligned} g'_Y(1) &= \mathrm{E}[Y] \\ g''_Y(1)) &= \mathrm{E}\left[Y^2 - Y\right] \ . \end{aligned} \tag{A.3}$$

The first derivation of $g_Y(z)$ can be calculated for equation (5.29) then it will be $g'_Y(z) = g'_S(g_R(z))\, g'_R(z)$. Let $z \to 1$, $g'_Y(1) = g'_S(1)\, g'_R(1)$ is obtained. Therefore, the first moments of the random variables Y, S, and R provide the equation of $\mathrm{E}[Y] = \mathrm{E}[S]\,\mathrm{E}[R]$. Then the expected value of service time ($\mathrm{E}[S]$) is:

$$\mathrm{E}[S] = \frac{\mathrm{E}[Y]}{\mathrm{E}[R]} \ . \tag{A.4}$$

Taking the derivation of equation (A.2), the second derivation of $g_Y(z)$ is derived as:

$$\begin{aligned} g''_Y(z) &= \left(g'_S(g_R(z)) \cdot g'_R(z)\right)' \\ &= g''_S(g_R(z)) \cdot g'^{1}_R(z) \cdot g'_R(z) + \\ &\quad g'_S(g_R(z)) \cdot g'_R(z) \cdot g''_R(z) \ . \end{aligned} \tag{A.5}$$

Let $z \to 1$, then the second derivation of $g_Y(z)$ can be represented as:

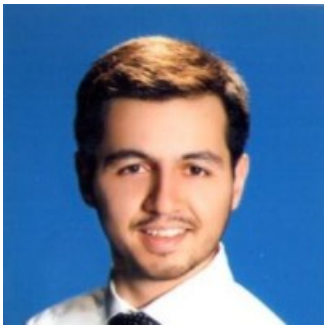
$$g''_Y(z) = g''_S(1) \cdot g'_R(1) \cdot g'_R(1) + g'_S(1)\, g''_R(1) \ . \tag{A.6}$$

Equations (A.3) and (A.6) are combined and simplified to obtain the following result:

$$\begin{aligned} &\mathrm{E}\left[Y^2 - Y\right] = \mathrm{E}\left[S^2 - S\right]\mathrm{E}[R]\,\mathrm{E}[R] + \mathrm{E}[S]\,\mathrm{E}\left[R^2 - R\right] \\ &= \mathrm{E}\left[S^2\right]\mathrm{E}[Y]^2 - \mathrm{E}[S]\,\mathrm{E}[R]^2 + \mathrm{E}[S]\,\mathrm{E}\left[R^2\right] - \mathrm{E}[S]\,\mathrm{E}[R] \\ &= \mathrm{E}\left[S^2\right]\mathrm{E}[R]^2 + \mathrm{E}[S]\left(\mathrm{E}\left[R^2\right] - \mathrm{E}[R]^2 - \mathrm{E}[R]\right) \\ &= \mathrm{E}\left[S^2\right]\mathrm{E}[R]^2 + \mathrm{E}[S]\left(Var(R) - \mathrm{E}[R]\right) \ . \end{aligned} \tag{A.7}$$

Substituting equation (A.4) into equation (A.7), the second moment of the service time ($\overline{S^2}$) obtained as:

$$\begin{aligned}
&\mathrm{E}\left[Y^2\right] - \mathrm{E}\left[Y\right] = \mathrm{E}\left[S^2\right]\mathrm{E}[R]^2 + \frac{\mathrm{E}\left[Y\right]}{\mathrm{E}\left[R\right]}\left(Var\left(R\right) - \mathrm{E}\left[R\right]\right) \\
&\mathrm{E}\left[Y^2\right] - \mathrm{E}\left[Y\right] - \mathrm{E}\left[Y\right]\left(\frac{Var\left(R\right) - \mathrm{E}\left[R\right]}{\mathrm{E}\left[R\right]}\right) = \mathrm{E}\left[S^2\right]\mathrm{E}[R]^2 \\
&\mathrm{E}\left[S^2\right] = \overline{S^2} = \frac{\mathrm{E}\left[Y^2\right] - \mathrm{E}\left[Y\right]\left(\frac{Var(R)}{\mathrm{E}[R]}\right)}{\mathrm{E}[R]^2} \,.
\end{aligned} \tag{A.8}$$



CURRICULUM VITAE

Name Surname: Ömer Faruk Gemici

Place and Date of Birth: Konya, 26/11/1988

Adress: Emek Mh. Gürsu Sk. G-Marin A10/12 Darıca-KOCAELİ

E-Mail: omerfaruk.gemici@tubitak.gov.tr, omerfarukgemici@gmail.com

B.Sc.: Istanbul Technical University, Faculty of Electrical and Electronic Engineering, Department of Electronics Engineering

M.Sc.: Istanbul Technical University, Graduate School of Science, Engineering and Technology, Telecommunication Engineering Graduate Program

Professional Experience:

▪ Employed in Research Center for Advanced Technologies on Informatics and Information Security (TUBITAK BILGEM) as signal processing and communication system engineer with the responsibilities of design and implementation of both physical and MAC layer algorithms. (2011 - current)

PUBLICATIONS ON THE THESIS

Journal Papers:

▪ **Gemici, Ö.F.**, Kara, F, Hökelek, İ, Çırpan, H.A. User Scheduling and Power Allocation for Nonfull Buffer Traffic in NOMA Downlink Systems, *International Journal of Communication Systems (IJCS)* . 2019.

▪ **Gemici, Ö.F.**, Hökelek, İ. and Çırpan, H.A. On Queuing Delay Analysis of NOMA Downlink Systems in 5G NR, *ETRI Journal*, (submitted: 28 Jul 2020 [under review]).

▪ **Gemici, Ö.F.**, Hökelek, İ. and Çırpan, H.A. Modeling Queuing Delay of 5G NR with NOMA under SINR Outage Constraint, *IEEE Transactions on Vehicular Technology (TVT)* (submitted: 10 May 2020 [under review]).

International Conference Papers:

▪ **Gemici, Ö.F.**, Kara, F., Hökelek, İ., Kurt, G.K. and Çırpan, H.A. (2017). Resource Allocation for NOMA Downlink Systems: Genetic Algorithm Approach, 2017 *IEEE 40th International Conference on Telecommunications and Signal Processing (TSP)*, pp.114–118.

▪ **Gemici, Ö.F.,** Hökelek, İ. and Çırpan, H.A. (2019). NOMA Power Allocation for Minimizing System Outage under Rayleigh Fading Channel, 2019 *IEEE 40th Sarnoff Symposium*, pp.1–6.

National Conference Papers:

▪ Kara, F., **Gemici, Ö.F.,** Hökelek, İ. and Çırpan, H.A. (2017). Optimal Power Allocation for DL NOMA Systems, 2017 *IEEE 25th Signal Processing and Communications Applications Conference (SIU)*, pp.1–4.

OTHER PUBLICATIONS

▪ **Gemici, Ö. F.,** Hokelek, I. Cirpan, H.A., Trade-off Analysis of QoS-aware Configurable LTE Downlink Schedulers, (2013) *20th International Conference on Telecommunications (ICT)*, vol., no., pp.1,5, 6-8 May 2013.

▪ **Gemici, Ö. F.,** Hökelek, İ., Çırpan, H.A., GA based Multi- Objective LTE Scheduler (2014) *1th International Workshop on Cognitive Cellular Systems (CCS)*, vol., no., pp.1,5, 2-4 September 2014.

▪ Ozcan, B., Zorlu, E., Atay, M., **Gemici O.F.,** Digital Logarithmic Video Detector Design, (2014), *22th Signal Processing and Communications Applications Conference (SIU)* , April 2014.

▪ Kahraman, F., İmamoğlu, M., Özcan, B. Y., Hüroğlu, C., Alasağ, T., **Gemici, Ö. F.,** Ateş, H. F. VİSKON-RS: Rapid Damage Assessment Software with Remote Sensing, 2015 *IEEE 23nd Signal Processing and Communications Applications Conference (SIU)* (pp. 1773-1776), May 2015.

▪ **Ö. F. Gemici et al.,** Real-time 5G Technology Development Platform, (2017) *25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2017, pp. 1-4.

▪ Demir, M. S.; **Gemici Ö. F.,** Uysal, M. Genetic Algortihm Based Resource Allocation Technique for VLC Networks, (2017) *25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2017, pp. 1-4.

▪ E. Şapla, **Ö. F. Gemici**, I. Hökelek, Performance Analysis of Signal Processing Algorithms Using Multi-core DSP Platform, (2018) *26th Signal Processing and Communications Applications Conference (SIU),* Izmir, 2018, pp. 1-4.