



M Ű E G Y E T E M 1 7 8 2

Budapest University of Technology and Economics

Faculty of Electrical Engineering and Informatics

Doctoral School of Informatics

Department of Telecommunications and Artificial Intelligence

Optimization of Power and Resource Allocation for 5G/6G Networks

Ph.D. Dissertation

Qusay Abdulameer Abdulhasan Alghazali

Scientific Supervisor

Prof. Tibor Cinkler DSc

Co-Supervisor

Dr. Husam Al-Amaireh

April 21, 2026

Declaration

I, the undersigned, **Qusay Alghazali**, hereby declare that the present Ph.D. dissertation is the result of my own original research and work. All sources of information, data, and ideas derived from the work of others have been duly acknowledged and cited in the references section. Any quotations or paraphrased material have been clearly identified and properly attributed.

Budapest, April 21, 2026

Qusay Alghazali

Abstract

The exponential growth of mobile data traffic, the proliferation of latency-sensitive applications, and the increasing density of user devices have collectively necessitated a transformative evolution in wireless network design. This thesis explores novel solutions for intelligent resource allocation in fifth-generation (5G) and beyond networks, focusing on enhancing spectral efficiency, minimizing energy consumption, and ensuring robust performance in dynamic and dense communication environments.

We begin by investigating Non-Orthogonal Multiple Access (NOMA), a technique that offers several benefits, including improved spectral efficiency particularly for heterogeneous and low-rate users, enhanced fairness, support for massive connectivity, efficient resource utilization, and flexible power allocation, where users share time-frequency resources by being assigned distinct power levels.

Recognizing that channel and power assignment play a pivotal role in the performance of Orthogonal Frequency Division Multiple Access (OFDMA) systems, we propose a hybrid algorithmic solution. The proposed Channel User Sorting and Filling (CUSF) algorithm efficiently allocates channels based on user channel conditions. Subsequently, a two-tiered power allocation mechanism is developed, combining water-filling techniques with Fractional Transmit Power Control (FTPC), to enhance user throughput while maintaining fairness and interference constraints.

Expanding the focus to energy sustainability, we address the rising energy demands in Mobile Edge Computing (MEC) environments. A novel energy-aware resource management framework is introduced, capable of dynamically adjusting computational and communication resources based on service type and network load. We derive optimized local processing energy models and implement adaptive strategies tailored for both NOMA and Massive Multiple-Input Multiple-Output (mMIMO) architectures. In NOMA scenarios, our proposed algorithm minimizes unnecessary energy consumption by fine-tuning the processing resource allocation, while in mMIMO networks, we optimize power control strategies to balance energy efficiency and performance. Simulation results demonstrate that our framework achieves significant reductions in energy consumption across various operational conditions.

The challenge of spectrum reuse in Device-to-Device (D2D) communication is then examined, where the coexistence of cellular and D2D users creates complex interference dynamics. To address the scalability issue in large networks, we introduce heuristic techniques involving user clustering via K-means and channel allocation through graph coloring based on proximity and interference levels. The system is further optimized by solving a max-min fairness power allocation problem, ensuring equitable resource distribution among users. The proposed schemes significantly outperform baseline solutions in terms of both user throughput and system fairness.

Finally, to address the limitations of static allocation strategies in highly dynamic network conditions, we present a machine learning-based framework for mobility-aware resource optimization in direct D2D-enabled cellular networks. By integrating user mobility prediction with the Non-Dominated Sorting Genetic Algorithm II (NSGA-II), our system jointly optimizes power control, spectrum reuse, and channel assignment in a multi-objective setting. Simulation results validate the superiority of our approach over static, greedy, and random strategies, showing enhanced throughput, minimized interference, and improved fairness under user mobility. The results also highlight the critical role of mobility awareness in real-time resource allocation, as it directly impacts SINR, energy efficiency, and Quality of Service (QoS). This framework sets the groundwork for further explorations in deep reinforcement learning, federated learning, and real-world deployments in next-generation wireless systems.

In summary, this thesis contributes to the design of intelligent, energy-efficient, and mobility-aware resource allocation mechanisms in modern cellular networks. By integrating optimization theory, heuristic algorithms, and machine learning, it provides a comprehensive toolkit for tackling the key challenges in 5GB wireless systems and paves the way for adaptive, scalable, and high-performance communication infrastructures.

Acknowledgements

I would like to express my deepest gratitude and sincere appreciation to my supervisor, Professor Tibor Cinkler, for his exceptional guidance, continuous support, and encouragement throughout my PhD journey. I am profoundly grateful for his academic mentorship, as well as for the knowledge, insight, and values he shared with me along the way. His wisdom and dedication greatly contributed to my development as a researcher and as a person. I also deeply appreciate the opportunities he provided for me to continue my studies and his support during difficult times, for which I will always remain sincerely thankful.

I would also like to express my heartfelt gratitude to my co-supervisor, Dr. Husam Al-Amaireh, for his continuous support, patience, dedication, and generous assistance throughout this work. I am especially grateful for his valuable advice, careful follow-up, and constant willingness to guide and support me whenever needed. His encouragement and sincere commitment were truly invaluable, particularly during challenging periods, and I deeply appreciate his contributions throughout this journey.

My sincere thanks also go to the Budapest University of Technology and Economics for providing a valuable academic environment where I was able to learn, grow, and complete my doctoral studies. I also warmly appreciate the Department of Telecommunications and Artificial Intelligence and the Doctoral School of Electrical Engineering for their support and encouragement throughout my studies. I am deeply grateful to the Stipendium Hungaricum Scholarship Programme for the opportunity and financial support that made this journey possible. I also sincerely thank the Ministry of Higher Education and Scientific Research in Iraq for its support and for helping students continue their education after difficult years of war and hardship.

With deep love and remembrance, I express my gratitude to my parents in heaven. Although they are no longer with me, their love, sacrifices, and prayers have always remained in my heart and continue to guide me every day.

I would also like to express my sincere gratitude to my family and loved ones for their constant support, patience, and encouragement. Their presence and belief in me were a lasting source of strength throughout this journey. My sincere thanks also go to my friends for their kindness, encouragement, and support over the years.

To everyone who stood by me and supported me during this journey, I offer my heartfelt thanks.

Contents

Declaration

Abstract	i
Acknowledgements	iii
List of Figures	ix
List of Tables	x
List of Acronyms	xi
List of Notations	xiv
1 Introduction	2
1.1 Background	3
1.1.1 Resource Allocation Challenges in 5G Networks	4
1.1.2 Non-Orthogonal Multiple Access (NOMA)	5
1.1.3 Massive Multiple-Input Multiple-Output (mMIMO)	6
1.1.4 Mobile Edge Computing (MEC)	7
1.1.5 Device-to-Device (D2D) Communication	8
1.2 Problem Statement	9
1.3 Research Questions	10
1.4 Research Objectives and Contributions	11
1.5 Research Methodology	12
1.6 Dissertation Structure	13
2 Joint power and channel allocation for NOMA in 5G networks and beyond	14
2.1 Introduction	14
2.2 Related Work	15
2.3 System Model	17

2.3.1	NOMA system	18
2.3.2	Joint Channel and Power Allocation	19
2.4	One-to-Many NOMA Algorithm	20
2.4.1	Stability assumption: the matching γ is stable if it is not blocked by any user–sub-channel pair	21
2.5	Power Allocation	22
2.6	Results and Discussion	23
2.7	Conclusion	27
3	Energy-Efficient Resource Allocation in Mobile Edge Computing Using NOMA and Massive MIMO	28
3.1	Introduction	28
3.2	Related Work	29
3.3	System Model	30
3.3.1	User local execution Energy	30
3.3.2	User data offloading Energy	31
3.4	Problem formulation	32
3.5	Offline processing	34
3.5.1	Tasks with known working load	34
3.5.2	Tasks with unknown working load	35
3.6	Data offloading	36
3.6.1	NOMA network configuration	36
3.6.2	NOMA Power allocation	38
3.6.3	Massive Multiple Input Multiple Output	42
3.7	Results and Discussion	48
3.8	Conclusion	54
4	Graph Coloring and User Clustering-Based Resource Allocation for Device-to-Device Communication in 5G Networks	55
4.1	Introduction	55
4.2	System Model	56
4.3	User Clustering	58
4.4	Graph Coloring	59
4.5	Power Allocation	60
4.6	Simulation	61
4.7	Conclusion	64
5	Mobility-Aware Resource Allocation in D2D Communications Using Ge- netic Algorithms	65

5.1	Introduction	65
5.2	Related Work	66
5.3	System Model	67
5.3.1	Zone-Based Spatial Division	68
5.3.2	Channel Model	68
5.3.3	POWER CONTROL CONSTRAINTS	70
5.3.4	MOBILITY MODEL	70
5.4	PROBLEM FORMULATION	71
5.4.1	Objective Function	71
5.4.2	System Constraints	72
5.4.3	Constraint Analysis	73
5.4.4	Mobility-Aware Optimization	74
5.5	Solution to the Optimization Problem	74
5.5.1	Non-dominated Sorting Genetic Algorithm II (NSGA-II) for Resource Allocation	75
5.5.2	Static Mobility-Aware NSGA-II Algorithm	76
5.5.3	Computational Complexity Analysis	79
5.5.4	Stability-Aware Optimization via Lyapunov Drift	80
5.5.5	Mobility Model	82
5.5.6	Integration of Mobility into NSGA-II	83
5.6	Performance Evaluation	86
5.6.1	Evaluation Objectives	86
5.6.2	Simulation Setup	86
5.6.3	Performance Metrics	87
5.6.4	Simulation Results	87
5.6.5	Power Allocation vs. Mobility Speed	90
5.7	Conclusion	93
6	Summary of Results and Future Work	94
6.1	Summary of Results	94
6.2	Future Work	95
	List of Publications	96
	Bibliography	97
	Appendices A–D	107
A	Water-Filling Power Allocation Derivation	108

B	KKT Derivation for Problem P_2	110
C	KKT-Based Convex Optimization for Problem P_5	112
D	Proof of Theorem 1: Lyapunov Stability	114

List of Figures

1.1	Capabilities of IMT-2030, highlighting new and enhanced features across AI, sensing, sustainability, latency, reliability, and capacity (Source: ITU-R WP5D [41]).	3
1.2	IMT-2030 standardization roadmap outlining system development and spectrum planning phases (Source: ITU-R WP5D [41]).	4
2.1	Capacity of the system versus different numbers of users.	24
2.2	System capacity distributed over channels.	25
2.3	Capacity per user for per sub-channel	26
2.4	System capacity for different alpha values.	27
3.1	Capacity of the system versus different numbers of users.	49
3.2	Local energy consumption for different delay times.	50
3.3	Energy consumed for remote processing vs. number of users.	51
3.4	Energy consumed for local and remote processing.	51
3.5	Impact of increasing PCs ($W_k = 1e9$).	52
3.6	CDF of Spectral Efficiency.	53
3.7	Spectral Efficiency Vs Number of Antenna.	53
3.8	Energy consumed when applying remote processing for different delay times.	54
4.1	A cellular network consists of a single BS, 6 CUs and 6 D2D users. Solid lines are communication channels, and dashed lines are interference channels.	56
4.2	Conflict graph with nodes colored to avoid interference	59
4.3	Initial user clustering based on proximity to the BS using K-means.	61
4.4	Channel distribution after applying graph coloring.	62
4.5	Users per channel after allocation.	62
4.6	Interference level experienced by each user.	63
4.7	CDF of power allocation with fairness consideration.	63
4.8	CDF of SINR achieved per user.	64

5.1	Enhanced concentric zone architecture in a cellular network. The base station (BS) is located at the center of the cell, surrounded by three radial zones. Examples of users include a cellular user (CU), a device-to-device (D2D) transmitter (Tx), and a receiver (Rx). Transition probabilities between zones are shown with arrows.	69
5.2	Total system throughput vs. user mobility speed.	88
5.3	Average interference power vs. user mobility speed.	89
5.4	SINR distribution for cellular and D2D users.	90
5.5	Impact of user mobility on fairness (Jain's Index).	91
5.6	Impact of user mobility on power allocation.	91
5.7	Pareto Front: Throughput vs. Interference for different allocation strategies.	92
5.8	Spatial distribution of interference power across the cellular coverage area. Interference is elevated near the cell edge and around dense user clusters (hotspots), highlighting the importance of spatially-aware resource allocation.	92

List of Tables

2.1	The list of simulation parameters.	24
2.2	Channel List of Users Sorted According to Their Gain	25
2.3	Channel User Assignment Using CUSF	26
3.1	Simulation Parameters	49
3.2	Simulation Parameters	52
5.1	Simulation Parameters	87
5.2	Fairness Comparison of Different Algorithms	89
5.3	Performance Comparison of Different Approaches	90

List of Acronyms

Acronym	Description
3GPP	3rd Generation Partnership Project
5G	Fifth Generation Mobile Networks
6G	Sixth Generation Mobile Networks
ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
AI-native	AI-driven network architecture concept for 6G
AIMD	Additive Increase Multiplicative Decrease
AoA	Angle of Arrival
B5G	Beyond 5G
BS	Base Station
BW	Bandwidth
CCCF	Complementary Cumulative Distribution Function
CCCP	Concave–Convex Procedure
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CPU	Central Processing Unit
CR	Computational Resource
CU	Cellular User
CUSF	Channel User Stable Fairness Algorithm
CSS-PA	Channel State Sorting–Pairing Algorithm
CSI	Channel State Information
D2D	Device-to-Device Communication
DPP	Drift-Plus-Penalty (Lyapunov Optimization Framework)
DRF	Dominant Resource Fairness
DVS	Dynamic Voltage Scaling
EECO	Energy-Efficient Computation Offloading
eMBB	Enhanced Mobile Broadband
FD	Full-Duplex
FG NET-2030	Focus Group on Network 2030 (ITU-T)
FTPC	Fractional Transmit Power Control
G(V,E)	Graph with Vertices and Edges
GADIA	Greedy Asynchronous Distributed Interference Avoidance
HD	Half-Duplex

Acronym	Description (continued)
Hexa-X	European 6G Flagship Research Project
ILP	Integer Linear Programming
IMT-2020	International Mobile Telecommunication Standard for 5G
IMT-2030	International Mobile Telecommunication Framework for 6G
IoT	Internet of Things
IPSBA	Iterative Power Scaling Bisection Algorithm
IRS	Intelligent Reflecting Surface
ITU-R	International Telecommunication Union – Radiocommunication Sector
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
JCAS	Joint Communication and Sensing
JFI	Jain’s Fairness Index
K-means	Clustering Algorithm for User Grouping
KKT	Karush–Kuhn–Tucker Conditions
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
mMIMO	Massive Multiple-Input Multiple-Output
ML	Machine Learning
MINLP	Mixed-Integer Nonlinear Programming
MMSE	Minimum Mean Square Error
MOEA/D	Multi-Objective Evolutionary Algorithm based on Decomposition
MOOP	Multi-Objective Optimization Problem
MU-LP	Multi-User Linear Precoding
NIMP	National Intern Matching Program
NOMA	Non-Orthogonal Multiple Access
NP	Non-deterministic Polynomial-time
NSGA-II	Non-dominated Sorting Genetic Algorithm II
NSGA-III	Non-dominated Sorting Genetic Algorithm III
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
$\text{opt}(\hat{G})$	Optimal Chromatic Number of the Conflict Graph
PC	Processing Center, which can be modeled as either a remote cloud or a nearby cloudlet.
PU	Processing Unit within the processing center.
PDF	Probability Density Function
QoE	Quality of Experience
QoS	Quality of Service
RIS	Reconfigurable Intelligent Surface
SCA	Successive Convex Approximation
SBX	Simulated Binary Crossover

Acronym	Description (continued)
SC-NOMA	Single Carrier Non-Orthogonal Multiple Access
SE	Spectral Efficiency
SFC	Service Function Chaining
SIC	Successive Interference Cancellation
SIMO	Single-Input Multiple-Output
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
SMUSC	Stable Matching of Users to Sub-Channels
SINR	Signal-to-Interference-plus-Noise Ratio
THz	Terahertz
ULA	Uniform Linear Array
URLLC	Ultra-Reliable Low-Latency Communication
XR	Extended Reality

List of Notations

Symbol	Description
\mathcal{K}	Set of all users in the network (cellular and D2D).
\mathcal{U}_{CU}	Set of cellular users (CUs), indexed by $n = 1, \dots, N_{\text{CU}}$.
\mathcal{U}_{D2D}	Set of device-to-device (D2D) pairs, indexed by $i = 1, \dots, N_{\text{D2D}}$.
\mathcal{N}	Set of available orthogonal channels, $\mathcal{N} = \{1, 2, \dots, N_{\text{ch}}\}$.
R_{cell}	Radius of the cell coverage area.
N_{ch}	Total number of available channels.
B_j	Bandwidth of channel c_j .
N_0	Thermal noise power spectral density (AWGN).
α	Path-loss exponent.
β	Path-loss constant.
$d_{i,k}$	Euclidean distance between transmitter i and receiver k .
$h_{i,k}^{(j)}$	Small-scale Rayleigh fading coefficient on channel j .
$g_{i,k}^{(j)}$	Channel gain between transmitter i and receiver k on channel j , $g_{i,k}^{(j)} = h_{i,k}^{(j)} \frac{\beta}{d_{i,k}^\alpha}$.
P_{max}	Maximum transmit power allowed per user.
$P_i^{(j)}$	Transmit power of user i on channel j .
$x_{i,j}$	Binary channel assignment variable, $x_{i,j} = 1$ if D2D pair i uses channel j , otherwise 0.
$y_{n,j}$	Binary channel assignment variable for CU n on channel j .
z_i	Binary activation variable for D2D pair i .
$\gamma_{i,k}^{(j)}$	SINR of D2D receiver k from transmitter i on channel j .
$\gamma_n^{(j)}$	SINR of cellular user n on channel j .
$\gamma_{\text{D2D},\text{min}}$	Minimum SINR threshold required for D2D users.
$\gamma_{\text{CU},\text{min}}$	Minimum SINR threshold required for CUs.
T_u	Throughput achieved by user u .
T_{CU}	Aggregate throughput of all cellular users.
T_{D2D}	Aggregate throughput of all D2D pairs.
T_{total}	Total system throughput, $T_{\text{total}} = T_{\text{CU}} + T_{\text{D2D}}$.
R_j	Data rate of user j (bit/s/Hz).
$I_{\text{D2D}}^{(j)}$	Total intra-tier interference among D2D pairs on channel j .
$I_{\text{CU}}^{(j)}$	Interference from cellular users to D2D receivers on channel j .
$I_{\text{D2D} \rightarrow \text{BS}}^{(j)}$	Aggregate interference from D2D users to the BS on channel j .

Symbol	Description
I_{total}	Total interference (sum of all cross-tier and intra-tier components).
N_{reuse}	Maximum number of D2D pairs allowed to reuse a single CU channel.
M	Allowed reuse margin in constraint formulation.
Γ_j	Minimum target rate (QoS threshold) for user j .
$f_1(\mathbf{c})$	Throughput objective for candidate solution \mathbf{c} .
$f_2(\mathbf{c})$	Interference objective for candidate solution \mathbf{c} .
$f_3(\mathbf{c})$	Fairness objective for candidate solution \mathbf{c} .
$J(\mathbf{c})$	Jain's fairness index for candidate solution \mathbf{c} .
\mathbf{P}	Vector of users' transmission powers.
\mathbf{X}	Binary channel assignment matrix for D2D pairs.
\mathcal{L}_i	Set of users assigned to cluster i by K-means algorithm.
c_i	Centroid of cluster i .
$J(C)$	K-means clustering objective function: intra-cluster variance.
$G(V, E)$	Network connectivity graph (users as vertices, interference as edges).
$\hat{G}(\hat{V}, \hat{E})$	Conflict graph for interference modeling.
$\chi(\hat{G})$	Chromatic number: minimum number of colors (channels) needed.
N_{zones}	Number of concentric mobility zones within the cell.
ΔR	Radial width of each mobility zone, $\Delta R = R_{\text{cell}}/N_{\text{zones}}$.
R_z	Radius of the z -th zone boundary.
$s(t)$	Zone index of a user at time t .
$t_{z,z'}$	Transition probability from zone z to z' in Markov chain.
\mathbf{T}	Markov transition matrix of size $N_{\text{zones}} \times N_{\text{zones}}$.
v_i	Velocity of user i .
θ_i	Movement direction of user i .
$(x_i(t), y_i(t))$	2D position of user i at time t .
σ_z^2	Variance of intra-zone positional noise.
NSGA-II	Non-dominated Sorting Genetic Algorithm II (multi-objective evolutionary optimizer).
N_p	Population size in NSGA-II.
g_{max}	Maximum number of generations in NSGA-II.
P_c	Crossover probability in genetic operations.
P_m	Mutation probability.
η_c, η_m	Distribution indices for SBX crossover and polynomial mutation, respectively.
\mathcal{P}^*	Final Pareto-optimal set of solutions.
G_{max}	Maximum generation count (synonymous with g_{max}).
K	Optimization interval (time slots between successive updates).
$L(t)$	Lyapunov function representing system stability.
$\Delta(t)$	Conditional Lyapunov drift, $\mathbb{E}[L(t+1) - L(t) \mathbf{x}(t)]$.
V	Drift-plus-penalty control parameter.

Symbol	Description
$Q_u(t)$	Virtual queue for throughput deficit of user u .
T_u^{target}	Target throughput for user u .
B	Upper bound constant from drift analysis.
ϵ	Minimum Lyapunov decay rate (stability margin).
I_{\max}	Upper bound on interference.
T_{\max}	Upper bound on throughput.
\mathcal{H}	Hypervolume metric used in Pareto front convergence.

Chapter 1

Introduction

Sixth-generation (6G) wireless networks are envisioned as the next paradigm shift in mobile communication, aiming to establish a hyper-connected, intelligent, and sustainable digital ecosystem by 2030. Although the global rollout of 5G networks is ongoing, inherent architectural and performance constraints have motivated early research toward 6G. This next-generation network evolution is driven by emerging applications such as extended reality (XR), holographic telepresence, digital twins, precision industrial automation, and pervasive artificial intelligence (AI)—all of which demand exceptional levels of reliability, ultra-low latency, massive capacity, and energy efficiency [29].

Collaborative research initiatives from academia and industry—including contributions from IEEE, 3GPP, and international alliances such as the ITU-R IMT-2030 working group, Hexa-X, and the Next G Alliance—have identified key technological pillars for 6G. These include sub-terahertz and terahertz (THz) communication to support data rates exceeding 100 Gbps, AI-native network architectures for autonomous optimization, and joint communication and sensing (JCAS) for immersive, context-aware services [1, 7, 41]. Additionally, 6G is expected to integrate intelligent reconfigurable surfaces, satellite-terrestrial convergence, quantum communication, and extreme edge computing to fulfill the connectivity and performance demands of diverse industrial sectors [7].

Figure 1.1 summarizes the advanced features envisioned in IMT-2030, building upon the foundation laid by IMT-2020 (5G).

The overarching design principles for 6G also emphasize sustainability and digital inclusion. Current efforts focus on energy-efficient architectures, dynamic spectrum access, and expanded reach to underserved and remote areas. The Next G Alliance roadmap, in particular, reinforces this vision by aligning national priorities with industry-driven objectives—such as AI-native infrastructure, energy-conscious operations, and domain-specific innovation [7].

Figure 1.2 illustrates the IMT-2030 standardization process, beginning with the endorsement of Recommendation M.2160 in 2023. It outlines critical milestones including framework development, requirement definition, spectrum coordination, and evaluation of candidate technologies in preparation for global 6G deployment.

In summary, 6G is not a linear progression of 5G capabilities but a transformative reimagining of wireless communication. It envisions an intelligent and responsive physical-digital environment, enabled by ubiquitous, real-time, and sustainable connectivity to address the evolving needs of society and industry on a global scale.

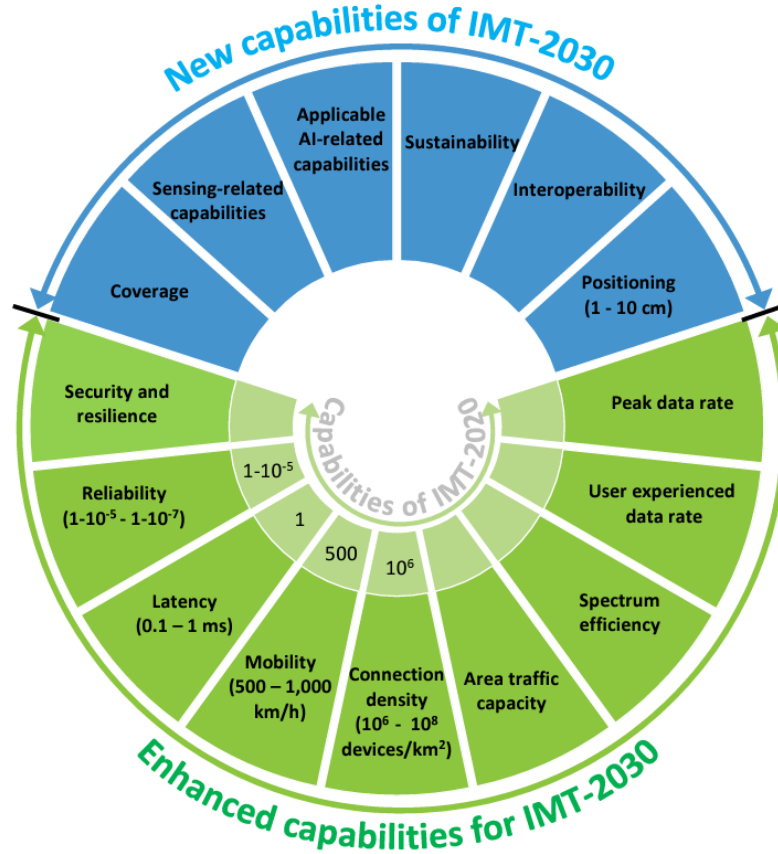


Figure 1.1: Capabilities of IMT-2030, highlighting new and enhanced features across AI, sensing, sustainability, latency, reliability, and capacity (Source: ITU-R WP5D [41]).

1.1 Background

The evolution of 5G networks has introduced unprecedented challenges in resource allocation, driven by diverse services, dense deployments, and limited spectral and computational resources. Traditional methods, designed for earlier generations, fail to address the complexity of dynamic and heterogeneous 5G networks [26].

Non-orthogonal multiple access (NOMA) has emerged as a key enabler for improving spectrum utilization by allowing multiple users to share frequency bands. This innovation reduces bandwidth limitations while increasing network capacity. However, implementing NOMA requires addressing inter-user interference and optimizing power allocation [123]. Similarly, mobile edge computing (MEC) reduces latency by processing computational tasks closer to end-users, while device-to-device (D2D) communication minimizes core network congestion by enabling direct connections [77].

This thesis proposes resource allocation frameworks that improve spectrum utilization, reduce energy consumption, mitigate interference, and enhance fairness and adaptability in dynamic 5G networks by exploiting NOMA, MEC, and D2D technologies.

Relationship and Timelines

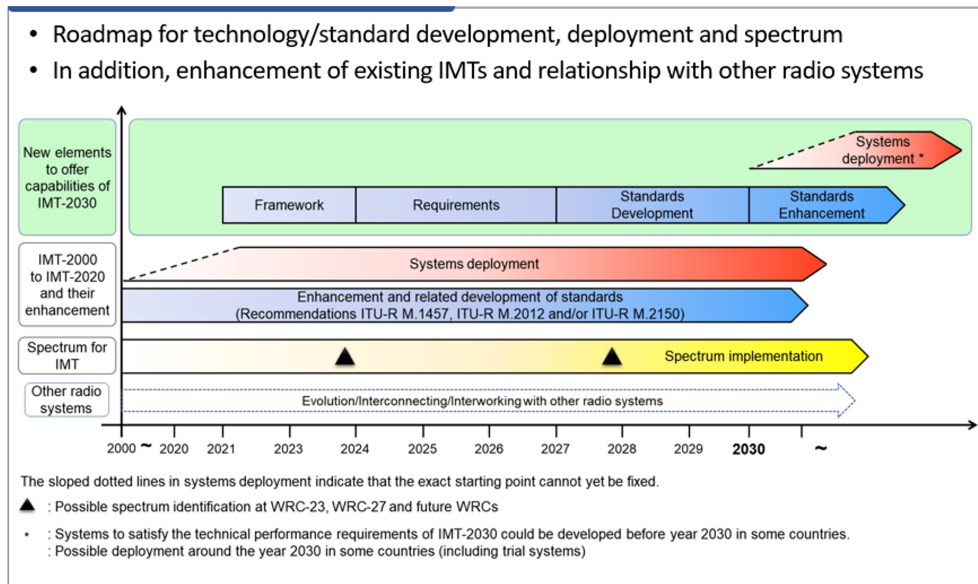


Figure 1.2: IMT-2030 standardization roadmap outlining system development and spectrum planning phases (Source: ITU-R WP5D [41]).

1.1.1 Resource Allocation Challenges in 5G Networks

Resource allocation in 5G networks involves optimizing the use of limited resources, such as spectrum, computational power, and energy, to meet the demands of diverse applications. The unique characteristics of 5G networks introduce several critical challenges:

- **Heterogeneous Service Requirements:** 5G supports a variety of services, including enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC). Each service has distinct requirements:
 - eMBB demands high data rates and large bandwidth for applications like streaming and augmented reality.
 - URLLC requires extremely low latency and high reliability for critical applications such as autonomous vehicles and remote surgery.
 - mMTC involves massive connectivity with low data rates, targeting IoT devices.

Balancing these conflicting requirements is a significant challenge for resource allocation [82].

- **Spectrum Scarcity and Efficiency:** The radio spectrum is a finite resource, and its allocation among a growing number of users and services is increasingly complex. Technologies like NOMA address this by enabling multiple users to share the same frequency band using power-domain multiplexing. However, managing inter-user interference and optimizing power allocation in real-time environments remain major obstacles [39].
- **Computational Resource Constraints:** With the proliferation of edge devices and latency-sensitive applications, computational resources are pushed to their limits. MEC reduces the computational burden on central servers by offloading tasks to

edge nodes near end-users. However, allocating computational resources dynamically while balancing workloads and ensuring energy efficiency is a persistent challenge [106].

- **Interference Management:** Dense deployments of 5G base stations and devices increase the risk of co-channel and inter-cell interference. D2D communication, while improving spectrum efficiency, adds another layer of complexity by introducing interference among directly communicating devices. Effective interference management is crucial to maintaining network performance [8].
- **Energy Efficiency:** Energy consumption in 5G networks is a critical concern due to the environmental and operational costs of dense deployments. Resource allocation must consider energy-efficient mechanisms, such as power control and energy harvesting, to ensure sustainability without compromising performance [34].
- **Dynamic and Adaptive Allocation:** The dynamic nature of 5G traffic, with varying user demands and mobility, requires adaptive resource allocation frameworks. These frameworks must respond to changing network conditions in real-time while optimizing spectrum, computation, and energy usage [97].

Efficient resource allocation frameworks must address these challenges holistically, leveraging advanced techniques to optimize spectrum utilization, computational resources, and energy efficiency.

1.1.2 Non-Orthogonal Multiple Access (NOMA)

Non-Orthogonal Multiple Access (NOMA) has emerged as a promising technology to address the spectral efficiency challenges in 5G networks. Unlike traditional orthogonal multiple access (OMA) techniques, which allocate separate time, frequency, or code resources to users, NOMA enables multiple users to share the same frequency band simultaneously by leveraging power-domain multiplexing. This approach significantly enhances spectrum utilization, making it suitable for the high user density and diverse service requirements of 5G networks [39].

The fundamental principle of NOMA is to assign different power levels to users based on their channel conditions. Users with better channel conditions are allocated lower power levels, while users with weaker channel conditions receive higher power levels. Successive interference cancellation (SIC) is then employed at the receiver to decode signals, starting with the strongest signal. This power-domain differentiation ensures that multiple users can access the same resources without causing excessive interference [26].

Despite its advantages, implementing NOMA introduces several challenges. Inter-user interference remains a significant concern, especially in scenarios with dense deployments or when users have closely matched channel gains. Managing power allocation in real-time to ensure fairness and maintain user Quality of Experience (QoE) is computationally demanding. Additionally, the effectiveness of SIC is highly dependent on accurate channel estimation, making it vulnerable to errors and delays in practical deployments [24, 39].

NOMA's integration with other enabling technologies, such as mobile edge computing (MEC) and device-to-device (D2D) communication, further complicates resource allocation. For instance, combining NOMA with MEC requires balancing the trade-offs between computational resources and spectral efficiency. Similarly, integrating NOMA with D2D communication introduces new interference management challenges due to the simultaneous transmission of multiple signals in close proximity [116].

NOMA has been recognized as a key component for beyond-5G (B5G) and sixth-generation (6G) networks. Its ability to support massive connectivity and high spectral efficiency makes it particularly suitable for applications such as massive machine-type communication (mMTC) and ultra-reliable low-latency communication (URLLC). NOMA supports URLLC by enabling multiple users to share the same time-frequency resources simultaneously, which reduces scheduling latency and improves spectral efficiency. Unlike orthogonal methods (OMA) that require waiting for available time slots, NOMA allows critical packets to be transmitted immediately by separating users in the power domain through superposition coding [79]. However, addressing its implementation challenges, including power optimization, interference management, and scalability, remains a critical area of ongoing research [39, 97].

In this thesis, the NOMA-related challenges are addressed through resource allocation frameworks that jointly optimize channel assignment and power control under fairness and interference constraints. The proposed methods aim to improve spectral efficiency, enhance user fairness, and reduce performance degradation in dynamic wireless environments.

1.1.3 Massive Multiple-Input Multiple-Output (mMIMO)

Massive Multiple-Input Multiple-Output (mMIMO) is a cornerstone technology for 5G and beyond, enabling significant improvements in spectral efficiency, energy efficiency, and capacity. Unlike traditional MIMO systems, mMIMO employs hundreds or even thousands of antennas at base stations to serve multiple users simultaneously. By exploiting spatial multiplexing and beamforming, mMIMO systems achieve unprecedented levels of throughput and connectivity [65].

The core concept of mMIMO lies in its ability to spatially separate user signals using highly directional beams. This is achieved through advanced beamforming techniques that focus energy toward specific users, minimizing interference and maximizing spectral reuse. The large number of antennas in mMIMO also facilitates robust interference suppression, making it particularly suited for ultra-dense 5G deployments [13].

Despite its advantages, implementing mMIMO introduces several challenges. One of the primary challenges is the need for accurate channel state information (CSI) at the base station. Acquiring and processing CSI becomes increasingly complex as the number of antennas and users grows, leading to high computational and signaling overheads. Pilot contamination, where non-orthogonal pilot signals interfere with each other during channel estimation, further degrades the system's performance [69]. Efficient pilot allocation and advanced signal processing algorithms are essential to address this issue.

Another critical challenge is the increased energy consumption of mMIMO systems. Although mMIMO is inherently energy-efficient due to its ability to focus energy toward users, the operation of a large number of antennas and the associated hardware introduces significant power demands. Research has focused on hardware-efficient designs, hybrid beamforming, and the use of low-resolution analog-to-digital converters (ADCs) to reduce energy consumption while maintaining performance [117].

The integration of mMIMO with other 5G enabling technologies, such as mobile edge computing (MEC) and non-orthogonal multiple access (NOMA), adds another layer of complexity. For instance, coordinating resource allocation between mMIMO and MEC systems requires advanced optimization frameworks that balance computational and spectral efficiency. Similarly, combining mMIMO with NOMA introduces interference management

challenges due to the simultaneous use of power-domain and spatial-domain multiplexing [105].

Massive MIMO is also expected to play a critical role in beyond-5G (B5G) and sixth-generation (6G) networks. Its scalability and ability to support massive connectivity make it an ideal candidate for advanced applications such as holographic communications, industrial automation, and ultra-reliable low-latency communication (URLLC). Addressing its challenges, including hardware limitations, energy efficiency, and integration with other technologies, remains a key focus of ongoing research [66].

In this thesis, the challenges of mMIMO are addressed through energy-aware resource allocation and power optimization frameworks. In particular, the proposed work focuses on reducing energy consumption, improving spectral efficiency, and managing interference under practical constraints such as imperfect CSI and large-scale user deployment. These contributions are developed in the context of MEC-enabled and high-density wireless scenarios, where efficient mMIMO operation is essential.

1.1.4 Mobile Edge Computing (MEC)

Mobile Edge Computing (MEC) is a transformative technology in 5G networks, designed to bring computation and storage resources closer to the end-users. By offloading computational tasks to edge servers located near base stations, MEC significantly reduces latency, enhances network efficiency, and alleviates the burden on centralized cloud infrastructures. This proximity to end-users makes MEC particularly valuable for latency-sensitive applications such as autonomous vehicles, augmented reality, and industrial automation [63].

One of the primary benefits of MEC is its ability to reduce end-to-end latency by processing data locally, rather than routing it to distant cloud servers. This reduction in latency is critical for enabling ultra-reliable low-latency communication (URLLC) in 5G networks. MEC also supports computational offloading, where resource-intensive tasks from devices with limited processing power, such as IoT devices, are delegated to edge servers [106]. However, the dynamic nature of 5G networks, characterized by fluctuating user demands and mobility, introduces significant challenges in the efficient allocation of MEC resources.

A key challenge in MEC is resource allocation, which involves distributing computational, storage, and bandwidth resources among multiple users. Balancing these resources dynamically to meet varying user demands while ensuring energy efficiency is computationally complex. Additionally, the limited capacity of edge servers, compared to centralized cloud infrastructures, makes optimizing resource allocation critical [21]. This challenge is further compounded by the need to integrate MEC with other 5G enabling technologies, such as massive MIMO and non-orthogonal multiple access (NOMA). For instance, the interplay between MEC and mMIMO requires joint optimization of computational and spectral resources, while MEC integration with NOMA must account for power and resource-sharing trade-offs [116].

Another significant challenge lies in managing user mobility. As users move across cells, maintaining seamless service delivery requires efficient handover mechanisms and resource reallocation strategies. Failure to manage these transitions effectively can result in service interruptions and degraded performance [95].

Energy efficiency is also a critical concern in MEC systems. While MEC reduces the energy consumption of end-user devices by offloading computational tasks, the operation of edge servers introduces additional energy demands. Techniques such as task scheduling,

energy-aware computation offloading, and the use of renewable energy sources are being explored to address this issue [63].

Looking ahead, MEC is expected to play a pivotal role in beyond-5G (B5G) and sixth-generation (6G) networks. Its ability to support emerging applications, such as tactile Internet, holographic communications, and real-time data analytics, positions MEC as a cornerstone technology for future networks. Addressing the challenges of resource allocation, user mobility, and energy efficiency will be critical to unlocking MEC's full potential [21].

In this thesis, the challenges of MEC are addressed through energy-aware resource allocation frameworks that jointly consider communication and computation aspects. The proposed work focuses on minimizing energy consumption for both local and remote processing, optimizing task offloading decisions, and improving resource efficiency under delay and channel constraints. These contributions are developed for NOMA- and mMIMO-enabled MEC scenarios, where dynamic allocation is essential for maintaining performance.

1.1.5 Device-to-Device (D2D) Communication

Device-to-Device (D2D) communication is a key enabling technology in 5G networks, allowing direct communication between nearby devices without routing data through a base station or core network. This approach significantly enhances spectrum utilization, reduces latency, and alleviates network congestion, making it ideal for applications such as proximity-based services, public safety communications, and the Internet of Things (IoT) [8].

The primary advantage of D2D communication lies in its ability to reuse spectrum resources. By enabling devices to communicate directly, D2D can operate as an underlay to traditional cellular networks, thereby improving spectral efficiency. This spectrum reuse, however, introduces new interference management challenges, particularly when multiple D2D links coexist with cellular users in the same frequency band. Efficient interference coordination mechanisms are critical to ensure that D2D transmissions do not degrade the performance of cellular networks [30].

D2D communication also offers energy efficiency benefits, as direct communication between devices typically requires lower transmission power compared to routing signals through a base station. However, energy-efficient resource allocation becomes increasingly complex in dense networks, where numerous D2D pairs compete for limited resources. Power control strategies and energy-aware algorithms are essential to address these challenges [102].

Another critical challenge in D2D communication is device discovery. Before initiating communication, devices must identify potential D2D peers within their proximity. This process, particularly in highly dynamic environments, can lead to signaling overhead and delays. Efficient and low-latency discovery protocols are necessary to support real-time applications such as autonomous vehicles and augmented reality [8].

The integration of D2D with other 5G technologies, such as Non-Orthogonal Multiple Access (NOMA) and Mobile Edge Computing (MEC), presents additional opportunities and challenges. For instance, combining D2D with NOMA can enhance spectral efficiency further but requires advanced interference management techniques due to the simultaneous use of power-domain and spatial-domain multiplexing. Similarly, integrating D2D with MEC enables efficient offloading of computational tasks between devices, reducing latency and improving network scalability, but introduces complexities in resource coordination and mobility management [125].

D2D communication is poised to play a significant role in beyond-5G (B5G) and sixth-generation (6G) networks. Its ability to support massive connectivity and enable decentralized architectures makes it an ideal solution for emerging applications such as smart cities, industrial automation, and disaster recovery. Addressing challenges related to interference, energy efficiency, and seamless integration will be critical to fully realize the potential of D2D communication in future networks [28].

In this thesis, the challenges of D2D communication are addressed through interference-aware resource allocation frameworks that combine channel assignment, clustering, graph coloring, and fairness-based power control. The proposed work focuses on improving spectrum reuse efficiency, reducing interference among cellular and D2D users, and enhancing fairness in dense and dynamic network environments. These contributions provide scalable and practical solutions for D2D-enabled wireless systems under realistic deployment constraints.

1.2 Problem Statement

The evolution of wireless communication systems toward fifth-generation (5G) and beyond has introduced enabling technologies such as Non-Orthogonal Multiple Access (NOMA), Massive Multiple-Input Multiple-Output (mMIMO), Device-to-Device (D2D) communication, and Mobile Edge Computing (MEC). Although these technologies improve spectral efficiency, latency, and energy performance, they also create tightly coupled resource allocation problems involving channel assignment, power control, interference coordination, computation offloading, and mobility adaptation.

In NOMA-based systems, multiple users share the same time-frequency resources through power-domain multiplexing. This creates a joint channel assignment and power allocation problem in which user pairing, interference management, and fairness must be carefully balanced. Existing static or heuristic solutions often face limitations in providing efficient and fair allocations under varying channel conditions, dense deployments, and heterogeneous QoS requirements.

In MEC-enabled systems, the main challenge is to jointly optimize communication and computation resources. Specifically, mobile users must decide between local processing and offloading under delay, energy, and channel constraints, while the network must allocate subchannels, transmit powers, and edge processing resources efficiently. The problem becomes more difficult when both NOMA and mMIMO are considered, since communication performance and energy consumption become strongly interdependent.

In D2D-enabled cellular systems, direct communication improves spectrum reuse but introduces severe cross-tier and intra-tier interference because D2D pairs share resources with cellular users. This gives rise to a combinatorial optimization problem involving interference-aware channel reuse, user grouping, clustering, and fairness-based power control. In dense deployments, exact optimization becomes computationally prohibitive, requiring scalable heuristic or graph-based methods.

When user mobility is incorporated, the resource allocation problem becomes time-varying and significantly more complex. Changes in user positions affect path loss, channel gains, interference patterns, and link quality, making previously optimal allocations quickly outdated. Therefore, static allocation strategies are insufficient, and mobility-aware frameworks are needed to update channel and power decisions in response to dynamic network conditions.

Despite extensive research, the existing literature still lacks an integrated and scalable framework that addresses these problems in a unified manner. Many works consider NOMA, MEC, D2D, mMIMO, or mobility separately, while practical wireless systems require joint treatment of spectral efficiency, energy efficiency, interference mitigation, fairness, and adaptability. Moreover, exact mathematical optimization techniques often become intractable in large-scale or real-time settings, whereas simple heuristics usually sacrifice performance or robustness.

Accordingly, the core problem addressed in this thesis is the design of resource allocation frameworks for 5G and beyond wireless networks that are both efficient and adaptive. More specifically, this thesis seeks to solve: (i) channel and power allocation in NOMA systems under fairness and interference constraints, (ii) energy-aware computation and communication optimization in MEC scenarios, (iii) scalable interference-aware channel reuse and power control in D2D-enabled networks, and (iv) mobility-aware resource allocation under dynamic user movement. To address these specific problems, the thesis develops a set of complementary solutions based on mathematical optimization, heuristic clustering, graph-based algorithms, and machine learning-driven multi-objective optimization, including NSGA-II.

1.3 Research Questions

1. How can spectral efficiency be maximized in Non-Orthogonal Multiple Access (NOMA)-enabled wireless systems through intelligent channel and power allocation schemes?
2. What strategies can be employed to minimize energy consumption in Mobile Edge Computing (MEC) architectures while maintaining Quality of Service (QoS) for varying service demands and network conditions?
3. How can resource allocation be formulated and optimized for Device-to-Device (D2D) communication scenarios in a way that balances fairness, power control, and channel reuse efficiency?
4. What scalable and heuristic algorithms can be developed to solve complex resource allocation problems—such as Integer Linear Programming (ILP) formulations—in large-scale cellular networks?
5. How can user mobility be predicted and incorporated into real-time resource allocation frameworks to improve throughput, reduce interference, and enhance fairness in dynamic 5G and beyond environments?
6. In what ways can machine learning techniques, particularly multi-objective evolutionary algorithms like NSGA-II, improve power control, spectrum reuse, and overall system performance under variable user mobility conditions?
7. What are the comparative benefits of integrating clustering, graph coloring, and mobility-aware optimization in addressing interference and scalability issues in dense cellular deployments?
8. How can a unified framework be developed to jointly address spectral efficiency, energy efficiency, and mobility adaptation across heterogeneous wireless network scenarios?

1.4 Research Objectives and Contributions

This dissertation addresses core challenges in resource allocation for 5G and beyond (5GB) networks, where achieving high spectral and energy efficiency, minimizing interference, and dynamically adapting to user mobility are essential for performance. The research spans several complementary areas—Non-Orthogonal Multiple Access (NOMA), Mobile Edge Computing (MEC), Device-to-Device (D2D) communications, and machine learning (ML)-based mobility-aware optimization. Each chapter presents a novel contribution aimed at advancing the state-of-the-art in one of these domains. The research objectives and corresponding contributions are outlined as follows:

- **Chapter 2:** The first objective is to improve spectral efficiency in OFDMA-NOMA systems through intelligent channel and power allocation. This chapter introduces a novel two-stage strategy: first, it proposes the Channel User Sorting and Filling (CUSF) algorithm to solve the one-to-many user-channel allocation problem; second, it decouples the power control into a convex optimization formulation followed by the use of the Fractional Transmit Power Control (FTPC) algorithm. Together, these techniques enable non-orthogonal resource reuse under fairness constraints. The contributions include algorithm design, complexity analysis, and performance benchmarking against conventional NOMA baselines.
- **Chapter 3:** The second objective is to develop energy-efficient computation and communication strategies for mobile users in a MEC environment. This chapter proposes a novel framework for modeling and minimizing energy consumption during both local and remote processing. In the NOMA case, it introduces a sub-channel assignment and power control method to reduce uplink energy expenditure. For mMIMO, it presents two power optimization algorithms suitable for large-scale user scenarios. The proposed models account for delay sensitivity and channel conditions, and simulation results confirm energy savings across network loads. This chapter also sets the stage for integrating the downlink phase in future MEC research.
- **Chapter 4:** The third objective focuses on optimizing spectrum reuse and fairness in D2D-enabled cellular systems. This chapter formulates the channel reuse problem as a quadratic binary optimization model and introduces a scalable heuristic that combines K-means user clustering and graph coloring for interference-aware channel assignment. Power control is integrated into the solution to maximize achievable rates while ensuring user fairness. The simulation study demonstrates that this method reduces interference and enhances fairness in dense networks. This contribution bridges combinatorial optimization with practical heuristics to manage complexity in large deployments.
- **Chapter 5:** The final objective is to address the limitations of static allocation under dynamic user mobility. This chapter presents a machine learning-based resource allocation framework that uses mobility prediction along with the Non-Dominated Sorting Genetic Algorithm II (NSGA-II). The framework supports joint optimization of channel assignment, power control, and fairness under varying mobility profiles. Simulation results show significant performance gains in throughput, SINR, spectral efficiency, and user fairness compared to traditional greedy and random schemes. This work opens future directions involving deep reinforcement learning, federated models, and field deployment to further enhance adaptability and scalability.

By achieving these objectives, the dissertation delivers a comprehensive, modular set of tools for resource allocation in heterogeneous 5G and beyond networks. Each chapter contributes new algorithms, theoretical models, and simulation-based validations that collectively advance the design of scalable, energy-aware, and mobility-adaptive wireless communication systems.

1.5 Research Methodology

This dissertation adopts a structured, multi-stage research methodology to address the research questions identified in Section 1.3 and fulfill the objectives outlined in Section 1.4. The methodology is built on a combination of mathematical modeling, non-convex and convex optimization, heuristic design, graph-based methods, stochastic mobility modeling, and multi-objective evolutionary optimization. These tools are used to formulate, analyze, and solve resource allocation problems in next-generation wireless networks. The process is divided into four major phases, each corresponding to a key thematic area of the dissertation, and concludes with simulation-based performance validation.

1. **Channel and Power Allocation for NOMA (Chapter 2):** The OFDMA-NOMA resource allocation problem is formulated as a non-convex optimization problem. To handle its complexity, the problem is decomposed into channel allocation and power control sub-problems. A heuristic matching-based algorithm, namely Channel User Sorting and Filling (CUSF), is developed for channel assignment, while the power allocation stage is handled using convex reformulation together with Fractional Transmit Power Control (FTPC). The proposed methods are evaluated in terms of fairness, throughput, and computational complexity.
2. **Energy-Aware Optimization in MEC with NOMA and mMIMO (Chapter 3):** A mathematical energy-consumption framework is developed for both local and remote processing in MEC systems. The methodology combines delay-aware energy modeling, sub-channel allocation, and power optimization. For the NOMA-based case, the problem is treated using communication-computation coupling and resource optimization. For the mMIMO scenario, scalable power optimization algorithms are developed to reduce energy consumption while preserving spectral efficiency. The analysis focuses on the trade-offs among delay, energy, and throughput.
3. **Clustering and Graph-Based Resource Reuse for D2D Communications (Chapter 4):** The D2D channel reuse problem is formulated as a quadratic binary optimization problem. Because exact optimization becomes computationally expensive in dense deployments, a scalable heuristic solution is designed using K-means clustering and graph coloring for interference-aware channel assignment. This is followed by fairness-aware power control to improve balanced throughput and reduce interference. Simulation analysis is used to study performance under different user densities and topologies.
4. **Mobility-Aware Optimization Using Machine Learning (Chapter 5):** A dynamic resource allocation framework is developed by combining stochastic mobility modeling with multi-objective evolutionary optimization. User mobility is represented using Markov-based zone transitions, and the resource allocation problem is solved using the Non-Dominated Sorting Genetic Algorithm II (NSGA-II). The optimization simultaneously considers throughput, interference, and fairness under

dynamic user movement. Comparative analysis with static and greedy approaches is carried out using SINR, spectral efficiency, throughput, and fairness metrics.

The numerical results reported in this dissertation were obtained using custom-developed MATLAB and Python codes . The implementation was carried out using MATLAB together with standard toolboxes, including the Optimization Toolbox and Signal Processing Toolbox, as well as selected Python libraries for numerical computation, data processing, and visualization. The proposed algorithms and simulation scenarios were implemented directly based on the mathematical models and optimization formulations presented in the thesis, without relying on external network simulation frameworks.

Overall, the research methodology emphasizes the use of mathematical tools for problem formulation and solution design, including non-convex optimization, convex approximation, decomposition, clustering, graph theory, stochastic modeling, and evolutionary multi-objective optimization. These methods are combined with simulation-based validation to provide both theoretical insight and practical design guidance for scalable 5G and beyond wireless networks.

1.6 Dissertation Structure

The remainder of this dissertation is organized as follows:

Chapter 2 introduces a spectral efficiency enhancement framework for OFDMA-NOMA systems. It presents the CUSF algorithm for one-to-many channel user allocation and applies a two-step power control method using water-filling and FTPC. The proposed approach addresses the non-convexity of the joint allocation problem and improves overall resource utilization. [J1]

Chapter 3 presents an energy optimization framework for MEC environments. It proposes algorithms for minimizing energy consumption in both NOMA and mMIMO scenarios by adapting resource allocation based on service type and load. The chapter validates the approach through simulations, showing notable energy savings. [J2]

Chapter 4 addresses resource allocation for D2D communications in 5G and Beyond networks. The channel reuse problem is modeled as a binary ILP and solved through both exact and heuristic methods. Clustering and graph coloring are applied for channel assignment, followed by power allocation to ensure fairness. [C1]

Chapter 5 proposes a mobility-aware resource allocation framework for D2D-enabled networks. By combining zone-based mobility prediction with NSGA-II, the method optimizes power control and channel assignment while ensuring fairness. Simulation results highlight gains in throughput, interference reduction, and SINR stability. [J3]

Chapter 6 concludes the dissertation by summarizing the research findings and discussing the key improvements achieved through the proposed methodologies. This chapter also outlines potential avenues for future work in congestion control and real-time multimedia communications.

Appendices provide the complete mathematical and analytical foundations supporting the main chapters. They include water-filling and KKT-based derivations for power allocation, convex optimization formulations for resource management, and Lyapunov stability proofs. Collectively, these analyses support the theoretical validity of the proposed models and provide mathematical insight into the behavior, efficiency, and stability of the developed algorithms where applicable.

Chapter 2

Joint power and channel allocation for NOMA in 5G networks and beyond

2.1 Introduction

Numerous 5G business models and applications are continuously emerging and developing due to the rapid growth of the IoT, cloud services, and pervasive mobile devices and applications. To efficiently address the rising demand for BW and services in this expanding environment, cutting-edge wireless technologies must be introduced. [55].

The superiority of NOMA systems was theoretically shown in the straightforward scenario of a single Base Station (BS) and two users [87]. Since then, other works using various radio techniques and scenarios have addressed the assignment problem. The BS receives all signals in the up-link direction and can easily decode and cancel the individual data streams in a specific order. Each user must contend with this interference in the down-link direction since each user receives both his signal and all of the signals meant for the other users assigned to the same channel. Therefore, the receiver attempts to eliminate the signals interfering with its signal reception whenever it is practical. This results in a set of restrictions on the minimum rate that the receiver observes for each data stream to be terminated, which makes the receiver's job substantially more complex. The advantages of NOMA compared to traditional systems can be summarized as enhanced flexibility, more equitable access, and more significant spectral efficiency.

Since interference significantly impacts NOMA, channel, and power optimization are crucial to its performance. Over the past 20 years, much research work has been carried out on power and sub-channel allocation [70]. The challenge of joint optimization of user association and power regulation to maximize the overall spectral efficiency is proposed assuming that user-specific quality-of-service and total power transfer are guaranteed. The mixed-integer non-convex programming issue is addressed using a new transformation technique, which initially demonstrated that the non-convex channel allocation issue could be resolved nearly optimally in the Lagrangian dual domain [12].

In NOMA systems, attempting to overlay all users onto a single resource block is inefficient and unpractical in real systems and scenarios, due to the extensive decoding delay and the potential for severe error propagation in the Successive Interference Cancellation (SIC) process at the receiver. Hence, it becomes essential to decrease the count of users being superimposed on a given channel by distributing users considering the conditions of the channels and according to the complexity and delay requirements of the SICs. This process will result in groups of users, and each group can be handled as a NOMA group. By em-

ploying effective algorithms for user grouping and power allocation, the signal interference can be mitigated, which leads to overall system capacity enhancement.

In this work, we show that the joint resource and power allocation is a non-convex and NP-hard problem. Furthermore, we decouple the sub-channel and power allocation problems. We show that sub-channel allocation can be viewed as a matching procedure, where users and sub-channels are two sets that need to be paired together. This pairing aims to maximize the achievable data rate between them. We propose a novel resource allocation scheme by dividing the non-convex problem; first, we introduce a new one-to-many resource allocation by assuming equal power allocation over sub-channels. The main idea is to use one-to-many heuristics to predict the optimal resource allocation. Then we tackle the power allocation by deriving an iterative water-filling scheme for a total rate maximization for down-link NOMA systems. The optimal scheme can be obtained by formulating a power allocation as a sum rate maximization problem and then exhaustively searching for the solution to the formulated problem.

2.2 Related Work

Numerous studies have gone into great detail concerning optimizing resource management for NOMA transmission. The subject has attracted a great deal of interest in the literature. The discipline of wireless communications optimization has made outstanding strides in recent years as wireless communications technology continues to develop and grow. One of the key developments in this area was published in [12], which showed that the non-convex channel allocation problem can be solved in the Lagrangian dual domain with near-optimal performance. Since then, numerous studies have focused on resolving allocation issues in various contexts and across different radio technologies.

For instance, [86] proposed cross-layer solutions, [100] investigated cognitive radio, [114] analyzed small cell and heterogeneous networks, the authors in [74] explored cloud radio access networks, and in [68] they investigated Multiple-Input Multiple-Output (MIMO) systems. These studies exemplify the extensive research efforts to optimize wireless communications systems through effective resource allocation strategies. By tackling the challenges posed by non-convex optimization problems, researchers have made significant progress toward developing efficient and practical optimization techniques for wireless networks. These efforts are essential for meeting the ever-increasing demands of modern society for reliable and high-speed wireless connectivity. In [11], a straightforward case study using one BS and two users illustrated the NOMA system's theoretical superiority. The findings of this study provide compelling evidence that NOMA may be preferable to other multiple-access systems, even in critical situations. The authors in [57] studied different NOMA techniques, challenges and its implementations in 5G and beyond networks.

As part of the extensive research on optimizing resource management for NOMA transmission, the work in [93] presents a proposed framework to address a resource management problem in two-user NOMA systems. The authors aim to enhance the sum capacity of the system by first providing the minimum Quality of Service (QoS) for one mobile user and then allocating the remaining power to the other user to maximize overall capacity. This approach provides a potential solution to the challenge of resource allocation in NOMA transmission and contributes to the ongoing efforts to improve the performance of these systems. The key challenge in managing resource allocation for NOMA transmission is addressing the multi-user interference stemming from non-orthogonal channel access. However, this challenge is complicated by the non-convex essence of the allocation

problem, which requires advanced and complex algorithms to solve. A potential solution to this challenge is presented in [11], where the authors demonstrate that the NOMA Full-Duplex (NOMA-FD) mode is theoretically feasible and can provide substantial gains over NOMA Half-Duplex (NOMA-HD) and orthogonal multiple access. However, this approach demands proper co-channel multi-user interference management for optimal implementation. The greedy asynchronous distributed interference avoidance algorithm (GADIA) based power allocation strategy for NOMA-based communications was discussed in [76].

The authors in [50] presented a price-based power optimization scheme for down-link wireless networks to maximize both revenues and the average achievable rate of the network. To achieve this objective, the authors adopted a game-theoretic approach. Since the resulting optimization problem was non-convex, they decoupled it into more manageable sub-problems and utilized an alternating optimization algorithm to obtain an efficient solution. By doing so, they were able to effectively address the complexity of the problem and provide a viable approach for optimizing power allocation in wireless networks.

Several power control algorithms have been proposed for NOMA systems, such as the algorithm proposed in [67], which aims to maximize the transmission rate of users while minimizing the transmission power. The authors in [67] proposed a distributed power adaptive algorithm that adjusts the transmission power of a user based on the signal quality of adjacent users. These algorithms can help improve the performance of NOMA systems by optimizing power allocation. A low-complexity power allocation method was put forth in [84] to enhance the weighted sum capacity in down-link NOMA systems. The method took into account both a two-user case and a multi-user one. It employed closed-form solutions to tackle the non-convex optimization issue effectively.

The effect of power distribution on the equity of the down-link NOMA system was examined in [89]. The authors of [124] suggested an energy-efficient Power distribution plan that addressed the Single Carrier NOMA (SC-NOMA) system's sum rate maximization problem. Energy-efficient power allocation for a hybrid system with NOMA connected to OMA was researched in [111]. For unsatisfactory NOMA-based down-link heterogeneous networks, the problem of cluster formation and power-BW allocation is addressed in [17]. As a function of QoS requirements, SIC efficiency, and allotted BW, [18] and [16] concentrate on the up-link resource allocation and user pairing and determine the greatest practical NOMA cluster size.

In [27], the authors accomplished energy-efficient resource management in NOMA Heterogeneous Networks (HetNets) by employing a transformation technique that converted the original non-convex optimization problem into a convex problem. Subsequently, they employed a dual method for effective sub-channel and power allocation, enabling efficient utilization of network resources while maintaining energy efficiency. A plan to optimize user association and spectrum allocation was put out in reference [96]. The strategy is intended to boost system performance while considering the fairness restriction. However, NOMA networks are not covered by the present approaches, which exclusively deal with the problem of user association in conventional heterogeneous networks.

The performance of NOMA for massive MIMO (mMIMO) networks, which is dependent on beam-forming and user clustering, was studied by the authors in [10]. The work in [48] presents a distributed approach for resource allocation and interference management in wireless networks, focusing on energy efficiency. The proposed solution allows for flexible and dynamic resource partitioning between macro and small cells, enabling energy-saving resource allocation. In their study, the authors in [60] presented a straightforward NOMA system configuration involving a single BS and two users. The analysis considered the Nakagami-M fading channel model, accounting for the statistical characteristics of the

channel state information. Specifically, the authors formulated the outage probability of the NOMA network, offering insights into the system's performance under the influence of fading channels. While NOMA offers several advantages, such as high spectral efficiency and support for massive connectivity, its successful implementation is not without challenges. One significant challenge arises from the large number of users sharing the same system resources, which enormously increases the complexity of the SIC process at the receivers. To address this issue, the performance of NOMA-FD systems has been investigated under the presumptions of imperfect SIC and channel state information (CSI) errors in studies such as [18, 98, 113]. Efficient utilization of network resources through optimal resource allocation is necessary to overcome limitations and improve the performance of NOMA-based systems [42].

2.3 System Model

In this section, we assume a down-link BS serving a set of users denoted by \mathcal{K} , where $\mathcal{K} = \{1, 2, \dots, K\}$. The available BW B is equally divided into \mathcal{N} sub-channels, where $\mathcal{N} = \{1, 2, \dots, N\}$, and each sub-channel with a BW $b = B/N$. Also, Channel State Information (CSI) is fully available at BS. The users are assigned, according to NOMA, to the sub-channels based on their CSI in a manner that each sub-channel serves a sub-group of users.

The total power transmitted is denoted by P_T , and each sub-channel $c_n \in \mathcal{N}$ is assigned a power P_n such that $0 \leq P_n \leq P_T$. Also, users are assigned power p_{jn} where $u_j \in \mathcal{K}$ and $c_n \in \mathcal{N}$ such that $\sum_{u_j \in \mathcal{K}} p_{jn} \leq P_n$.

Let the channel between a user u_j and BS on sub-channel c_n be h_{jn} . The channel matrix \mathbf{H} between user u_j and BS can be seen as $\mathbf{H}_{jn} \in \mathbb{C}^{M_R \times M_T}$ with M_R, M_T being the number of received and transmitted antenna respectively.

Without loss of generality, we assume $M_R = 1$ (this can be seen as a single antenna user or a single link between BS and one received antenna).

In a rich multi-path environment (as is usually the case in cellular systems) and benefiting from the central limit theorem [22], the channel vector can be modeled as complex Gaussian distribution with $\mathbf{h}_{jn} \sim CN(\mu_{jn}, \mathbf{R}_{jn})$ where $\mu_{jn} \in \mathbb{C}^{M_T}$ represents the line of sight propagation, and the covariance matrix $\mathbf{R}_{jn} \in \mathbb{C}^{M_T \times M_T}$ represents the variable nature of the channel. This model is called Rayleigh fading in case $\mu_{jn} = 0$, otherwise it is a Rician channel. The off-diagonal elements in \mathbf{R}_{jn} represent the spatial directivity. 3GPP has modeled the channel attenuation as [36]:

$$\beta = -128.1 - 37.6 \log_{10} d \quad (2.1)$$

Where d is the separation in kilometer. Furthermore, the noise power can be represented as:

$$\sigma^2 \text{ (dBm)} = -174 + 10 \log_{10}(b) + n_f \quad (2.2)$$

Where b is in Hertz, and n_f is the hardware noise figure in dB. The data rate depends on Signal to Noise Ratio (SNR) and hence, by assuming that the transmitted signals from each antenna are independent and identically distributed (i.i.d.) with a total power p_{jn}

we get:

$$\begin{aligned}\text{SNR}_{jn} &= \frac{p_{jn} \text{tr}(\mathbf{R}_{jn})}{M_T \sigma^2}, \\ &= \frac{p_{jn} g_{jn}}{\sigma^2}\end{aligned}\tag{2.3}$$

where $\text{tr}(\cdot)$ is the trace of a matrix, and $g_{jn} = \frac{\text{tr}(\mathbf{R}_{jn})}{M_T}$ is the average channel gain. From (2.3) we can notice that the SNR for a single user with the optimal preprocessing (like matched filtering) transforms the Multiple-Input Single-Output (MISO) channel into an equivalent Single-Input Single-Output (SISO) channel.

2.3.1 NOMA system

A significant classification for multiple access systems has surfaced that is based on orthogonality. Various techniques such as time, frequency, coding, and space can be utilized to achieve the orthogonality of communication resources. When these resources, or their combination, achieve orthogonality, the communication schemes can be classified as Orthogonal Multiple Access (OMA). In contrast, NOMA is gaining popularity due to its potential to enhance spectral efficiency, user fairness, and reliability and accommodate more users. Numerous options, including coding and power, can be used to implement NOMA. Coding achieves multiplexing through the code domain. The code domain shares time and frequency, much like Code-Division Multiple Access (CDMA). In contrast, user-specific spreading sequences that are either sparse or non-orthogonal cross-correlation sequences with low correlation coefficients are used by code-domain NOMA [40]. On the other hand, NOMA power-domain multiplexing is generally regarded as less complex than code-domain. In power-domain NOMA, fractions of power (that sum up to a total of P_T) are allocated to users, thereby increasing the rate of OMA. The receiver differentiates users based on channel strength, with stronger channel users using the SIC method and weaker channel users treating the other signals as noise and decoding the correct signal. It is worth noting that the user's power fraction is not solely dependent on the user's channel condition; it can be controlled to regulate the rate per user to make the system fair.

The number of users served by the NOMA system, or more specifically, SIC should be limited to a certain threshold mainly for two reasons: reducing complexity and minimizing error propagation. If we denote the sub-group of NOMA users on sub-channel $c_n \in \mathcal{N}$ as $\mathcal{S}_n \subseteq \mathcal{K}$ such that $1 \leq |\mathcal{S}_n| \leq S$, then we can express users of the main group as $u_j \in \mathcal{K}$ and users belonging to the same NOMA sub-group on c_n as $u_{jn} \in \mathcal{S}_n$.

Consider that the users u_{jn} are arranged in descending order according to their gains such that:

$$g_{1n} \geq g_{2n} \geq g_{3n} \geq \dots \geq g_{Sn}\tag{2.4}$$

Then according to the NOMA principle, the power will be assigned as:

$$p_{1n} \leq p_{2n} \leq p_{3n} \leq \dots \leq p_{Sn}\tag{2.5}$$

The input signal u_{jn} is received and subjected to SIC by subtracting the signal intended for a later receiver from the composite signal. This leads to an improvement in the signal-to-interference-plus-noise ratio (SINR). To decode its own signal, u_{jn} first decodes the

interfering signals intended for the later receivers u_{in} , where $i > j$ and $i, j \in \mathcal{S}_n$. The interfering signals with lower order are not decoded and are treated as noise. Thus, the SINR prior to SIC can be expressed as follows:

$$\text{SINR}_{jn} = \frac{p_{jn}g_{jn}}{\sum_{i=1, i \neq j}^S p_{in}g_{jn} + \sigma^2}, \quad (2.6)$$

Then after SIC, the estimated SINR is:

$$\text{SINR}_{jn} = \frac{p_{jn}g_{jn}}{\sum_{i=1}^{j-1} p_{in}g_{jn} + \sigma^2} \quad (2.7)$$

Based on the SINR from (3.5), the user rate can be expressed as:

$$R_{jn} = \log_2 \left(1 + \frac{p_{jn}g_{jn}}{\sum_{i=1}^{j-1} p_{in}g_{jn} + \sigma^2} \right) \quad (2.8)$$

2.3.2 Joint Channel and Power Allocation

In our work, we consider the system performance as the total sum rate of all users. Then the optimization problem writes:

Objective: maximize the weighted sum rate of all users.

$$P2.1 : \quad \max_{x,p} \sum_{u_j \in \mathcal{K}} \alpha_j \sum_{c_n \in \mathcal{N}} R_{jn} x_{jn} \quad (2.9a)$$

Subject to:

$$\sum_{u_j \in \mathcal{K}} \sum_{c_n \in \mathcal{N}} p_{jn} \leq P_T, \quad (2.10a)$$

$$\sum_{u_j \in \mathcal{K}} p_{jn} \leq P_n, \quad \forall c_n \in \mathcal{N}, \quad (2.10b)$$

$$s_\ell \leq \sum_{u_j \in \mathcal{K}} x_{jn} \leq S, \quad \forall c_n \in \mathcal{N}, \quad (2.10c)$$

$$x_{jn} \in \{0, 1\}, \quad \forall u_j \in \mathcal{K}, c_n \in \mathcal{N}, \quad (2.10d)$$

$$p_{jn} \geq 0, \quad \forall u_j \in \mathcal{K}, c_n \in \mathcal{N}. \quad (2.10e)$$

Constants and parameters:

- α_j : weight coefficient associated with user u_j
- R_{jn} : achievable rate of user u_j on sub-channel c_n
- P_T : total transmit power budget
- P_n : maximum power allocated to sub-channel c_n
- s_ℓ : minimum number of multiplexed users per sub-channel
- S : maximum number of multiplexed users per sub-channel

Optimization variables:

- x_{jn} : binary channel assignment variable, where $x_{jn} = 1$ if user u_j is assigned to sub-channel c_n , and $x_{jn} = 0$ otherwise
- p_{jn} : power allocated to user u_j on sub-channel c_n

The objective is to maximize the weighted utility function (2.9a), where R_{jn} can be found in (2.8). Here α_j is the weight coefficient of u_j , $u_j \in \mathcal{K}$. The choice of these weights significantly impacts how resources are distributed among the users, and these weights can be employed to guide the resource distribution towards different objectives, such as giving priority to specific users and ensuring fairness by assigning a higher weight to a user with a relatively weaker channel. Constraint (2.10a) guarantees that the power budget will not be exceeded. Constraint (2.10b) ensures that the total sub-channel allocated power does not exceed a certain threshold. Constraints (2.10c) and (2.10d) ensure respectively that each sub-channel is both upper and lower bounded by the number of users, and u_j is multiplexed in c_n . Constraint (2.10e) ensures that power values are nonnegative.

The problem P2.1 is non-convex due to the existence of the binary variable x_{jn} and the interference term in the objective function. Furthermore, it is NP-hard, which can be seen if we make $S = 1$, reducing the problem to Orthogonal Frequency Division Multiple Access (OFDMA), where its NP-hardness is established in [56]. Because of the NP-hardness of this problem, we can no longer insist on having an efficient algorithm that can find its global optimum in polynomial time. Instead, we have to settle for less ambitious goals and find approximate solutions. In the next section, we address the sub-channel and power allocation problems independently.

2.4 One-to-Many NOMA Algorithm

In this algorithm, we are looking for matching between sub-channels and users without considering power assignment. We assume that channels have preferences over users in a way that they are able to rank order them based on their channel gain, and the same goes for users. However, each channel will be able to choose more than one user, and users will be able to choose one channel in a one-to-many scenario.

As mentioned above, we consider two finite and disjoint sets \mathcal{K}, \mathcal{N} for users and sub-channels respectively. Each sub-channel has preferences over users and each user has preferences over sub-channels. In our model, the preferences are considered transitive such that, if a user $u_j \in \mathcal{K}$ prefers sub-channel $c_a \in \mathcal{N}$ over $c_b \in \mathcal{N}$, and prefers c_b over c_c , then it definitely prefers c_a over c_c . The preferences (ordered from best to worst) for both users and sub-channels can be expressed as follows:

$$F(u_j) = c_a, c_b, c_c, \dots \tag{2.11}$$

$$F(c_n) = u_a, u_b, u_c, \dots \tag{2.12}$$

It can be noticed from (3.27) that u_j finds acceptable both c_a , c_b and discards other sub-channels below a certain channel gain threshold. Also the same applies for (3.28), in which sub-channel c_n accepts only users u_a, u_b, u_c and discards others with insufficient channel gain. We also denote $c_a \succeq_{u_j} c_b$ as: user u_j prefers c_a at least as c_b , and $u_a \succeq_{c_n} u_b$ as: sub-channel c_n prefers u_a at least as u_b .

The expected outcome of such a model is that each user is matched to at most one sub-channel and each sub-channel is matched to a specific quota (based on NOMA complexity and error propagation requirements), and the matching is bilateral in a way that a user is paired with a sub-channel if and only if the sub-channel is paired with the user. Based on the above, let us define the matching process γ as follows:

1. $|\gamma(u_j)| = 1$ for every user $u_j \in \mathcal{K}$, and $\gamma(u_j) = u_j$ if $\gamma(u_j) \notin \mathcal{N}$
2. $|\gamma(c_n)| = S$ for every sub-channel $c_n \in \mathcal{N}$; any unfilled position with users will be filled with c_n
3. $|\gamma(u_j)| = c_n \iff u_j \in \gamma(c_n)$

Since sub-channels in the matching algorithms serve a specific quota of users, the matching algorithm should allow the sub-channel to compare groups of users and compare different matching.

Next, inspired by the National Intern Matching Program (NIMP) [78], we present our modified algorithm which we name the Channel User Sorting and Filling (CUSF) algorithm. The functioning of the algorithm is outlined as follows:

- **Entry stage:**
First, the BS orders the users who have applied based on each sub-channel's ranking. Any user with an unacceptable channel gain is eliminated, and each user ranks the sub-channels, and any sub-channel with an unacceptable channel gain is eliminated from the users' lists. Next, the lists enter a list processing beginning with the matching stage.
- **Matching stage:**
In the matching stage, the algorithm searches for user-sub-channel pairs that are top-ranked in each other's ranking. If no matches are found, it proceeds to the next step, where the second-ranked sub-channel on each user's ranking is compared with the top-ranked user in that sub-channel's ranking. The generic step seeks to find user-sub-channel pairs such that the user is top-ranked on the sub-channel's ranking and the sub-channel is ranked k_{th} by the user. If matches are found, the algorithm goes to the next stage:
- **Provisional assignment and update stage**
In this stage, $k:1$ matches are tentatively made, and each user who is the top-ranked choice of their k_{th} choice sub-channel is tentatively assigned to that sub-channel. The rankings of users and sub-channels are then updated, and the algorithm returns to the start of the matching stage, which examines the updated ranking for new matches. The algorithm continues until no new tentative matches are found, at which point tentative matches become final.

The CUSF algorithm is shown in Algorithm 1.

2.4.1 Stability assumption: the matching γ is stable if it is not blocked by any user-sub-channel pair

A matching γ is said to be blocked by a sub-channel c_n and a user u_j if they are currently not matched to each other in γ , but both would prefer to be matched with each other

Algorithm 1 Channel User Sorting and Filling (CUSF) Algorithm

Input: Channel gain matrix between users $u_j \in \mathcal{K}$ and sub-channels $c_n \in \mathcal{N}$

Output: Stable and fair user–sub-channel assignment

- 1: Rank order both sub-channel and user lists based on channel gains as ζ and η
 - 2: **while** there exist unassigned users requesting one-to-one matches **do**
 - 3: Assign all items marked as tentative
 - 4: Remove sub-channels with lower ranks from users already assigned
 - 5: Eliminate users (tentatively assigned) from sub-channels they ranked lower than their current assignment
 - 6: **end while**
-

rather than their current assigned matches, where we consider (c_n, u_j) as a blocking pair because of the simultaneous occurrence of $\gamma(u_j) \neq c_n$, $c_n >_{u_j} \gamma(u_j)$, and $u_j >_{c_n} \gamma(c_n)$.

Theorem: CUSF is a stable algorithm.

Proof: After the algorithm stops, each sub-channel c_n is paired with the best S options from its updated rank-ordered list. This is because the algorithm only stops when it is no longer possible to find tentative $k : 1$ matches. The resulting matching is considered stable, as any user u_j that was initially ranked higher by sub-channel c_n than one of its final choices has been removed from c_n 's list due to being provisionally assigned to a higher-ranked sub-channel on u_j 's list. Therefore, the final assignment provides u_j with a position that they prefer over c_n . Consequently, the final matching is not unstable in terms of any c_n or u_j of this kind.

At this stage, we should have a new user–sub-channel assignment. Next, we move to power assignment–based CUSF algorithm outcome.

2.5 Power Allocation

After assigning sub-channels to users, we now look into the power allocation. For this purpose, we divide the problem into two parts: allocating power per sub-channel and then allocating power to users that are superimposed on a single sub-channel.

To allocate power per sub-channel, we consider sub-channels with only one user per sub-channel. Next, we establish an optimization problem based on maximizing the total sum rate as follows:

$$P2.2 : \max_{p_n} \sum_{n \in \mathcal{N}} R_n \quad (2.13a)$$

$$\text{s.t.} \quad \sum_{n \in \mathcal{N}} p_n \leq P_T, \quad (2.13b)$$

$$0 \leq p_n \quad (2.13c)$$

Problem $P2.2$ is strictly convex with respect to p_n , and so it has a unique solution. We can solve $P2.2$ by Karush–Kuhn–Tucker (KKT) conditions. To simplify the notations, we assume that

$$\frac{g_n}{\sigma^2} = H_n.$$

Let us establish the Lagrangian function as follows:

$$L(p, \lambda) = \sum_{n \in \mathcal{N}} \log_2(1 + p_n H_n) - \lambda \left(\sum_{n \in \mathcal{N}} p_n - P_T \right) \quad (2.14)$$

When we solve for p_n and λ (see Appendix A for derivations), then (2.14) gives:

$$p_n = \frac{1}{\ln 2 \lambda} - \frac{1}{H_n}, \quad (2.15)$$

The power allocation in (2.15) is called water-filling. From (2.14) and (2.15), we derive the procedure to get the optimal values for p_n as follows:

$$\lambda_{i+1} = \lambda_i - \mathcal{A} \left[P_T - \sum_{n \in \mathcal{N}} p_n \right]^+ \quad (2.16)$$

where \mathcal{A} is the step size, and $[\cdot]^+$ denotes a non-negative number. So far, the results give us power allocations considering only one user per sub-channel. The next step is to allocate power for each sub-channel with superimposed NOMA users. For this purpose, we benefit from Fractional Transmit Power Control (FTPC) [80] as follows:

$$p_{j,n} = \frac{p_n}{\sum_{u_i \in \mathcal{S}_n} H_{i,n}^{-\alpha_{\text{FTPC}}}} H_{j,n}^{-\alpha_{\text{FTPC}}} \quad (2.17)$$

where \mathcal{S}_n is the set of NOMA users multiplexed on sub-channel n , and $0 \leq \alpha_{\text{FTPC}} \leq 1$ is the decay factor. The case of $\alpha_{\text{FTPC}} = 0$ corresponds to equal transmit power allocation among the users. The more α_{FTPC} is increased, the more power is allocated to the user with lower channel gain $H_{j,n}$. The power assignment algorithm is shown in Algorithm 2.

Algorithm 2 Iterative Power Allocation

- Input:** Initial value $\lambda_{(0)} > 0$, tolerance threshold ξ
Output: Optimal power allocations p_n and $p_{j,n}$
- 1: Set iteration index $i = 0$
 - 2: **while** $|\lambda_{i+1} - \lambda_i| > \xi$ **do**
 - 3: Compute p_n using (2.15)
 - 4: Update λ_{i+1} using (2.16)
 - 5: $i \leftarrow i + 1$
 - 6: **end while**
 - 7: Compute $p_{j,n}$ using (2.17)
-

2.6 Results and Discussion

In the simulation section, we study the capacity of the system with the parameters mentioned in Table 2.1. Furthermore, we compare the scheme with another algorithm which is the Channel State Sorting-Pairing Algorithm (CSS-PA) [115]. The use of SIC in wireless communication requires a significant difference in SINR between paired users to prevent error propagation, and CSS-PA addresses this by pairing a user with a good channel condition with one who has a bad channel condition. This improves fairness and increases the capacity of the system. Additionally, OFDMA is used to evaluate the

Table 2.1: The list of simulation parameters.

Simulation Parameters	Parameter Value
Cell radius	500 m
The minimum distance between BS and UEs	50 m
The minimum distance between user and user	40 m
System bandwidth	5 MHz
The maximum number of UTs	60
Noise power spectral density	-174 dBm/Hz
Difference tolerance in algorithm 2	0.01
The base station peak power P_{BS}	30 dBm

impact of NOMA on the overall system performance.

Figure 2.1 illustrates the network capacity curve as the number of users in a cell increases from 10 to 60. The channel capacity is calculated by multiplying (2.8) with the BW of each sub-channel, then we summed the capacities of all sub-channels. We assumed the BW of each sub-channel equals the total BW divided by the number of sub-channels. The capacity of the cell system also increases with the number of users.

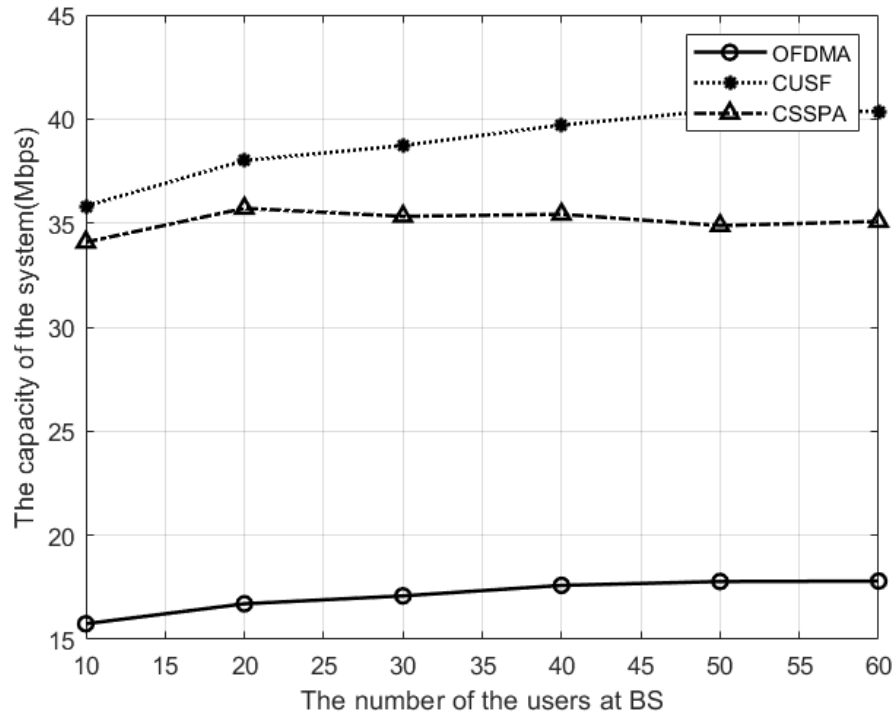


Figure 2.1: Capacity of the system versus different numbers of users.

CUSF provides the highest system capacity for the NOMA system compared to other algorithms studied. At 40 users, the CUSF algorithm outperforms CSS-PA by up to 15% and OFDMA by up to 60%. This is due to the limitations of OFDMA for using only one

user per sub-channel, which results in the BS being unable to fully utilize the spectrum resources.

In Figure 2.2, we show the capacity of the system distributed over channels. By studying the figure, despite different channel conditions (we have assumed fading Rayleigh channel condition), we can notice that the algorithm manages to provide good fairness of capacity between channels. In other words, the system has distributed the users per channel in an optimized way based on their channel conditions. Furthermore, a detailed capacity distribution per user per sub-channel is shown in Figure 2.3. In the figure, we can see that the algorithm has optimally distributed the users on sub-channels in a way that each sub-channel serves a user with a good channel condition and a user with a severe channel condition, and this scenario will maximize the capacity per sub-channel and hence will increase the overall system capacity.

Table 2.2: Channel List of Users Sorted According to Their Gain

Channel	User Number (Highest → Lowest Gain)									
1	3	1	2	5	6	8	4	7	10	9
2	3	4	1	6	10	2	8	5	7	9
3	1	4	3	5	2	8	6	9	10	7
4	3	1	5	9	2	4	10	8	6	7
5	1	3	4	2	8	9	10	5	7	6

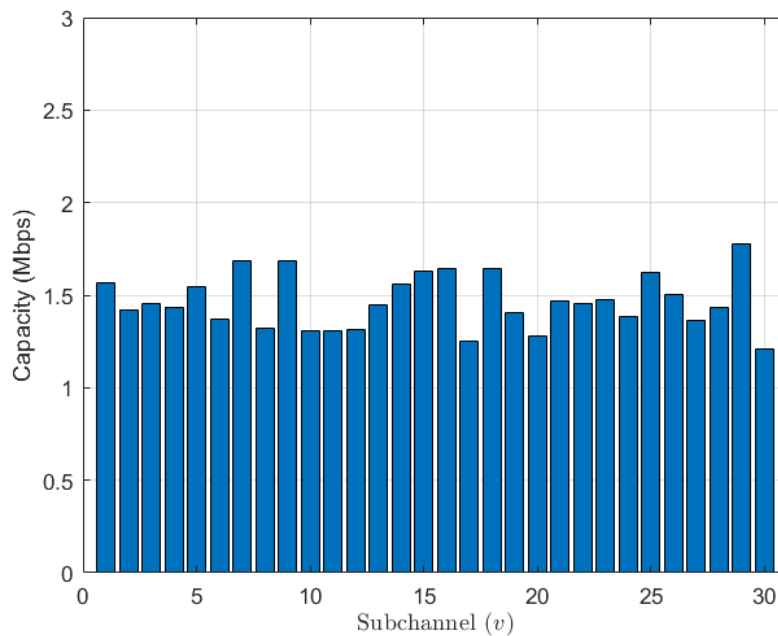


Figure 2.2: System capacity distributed over channels.

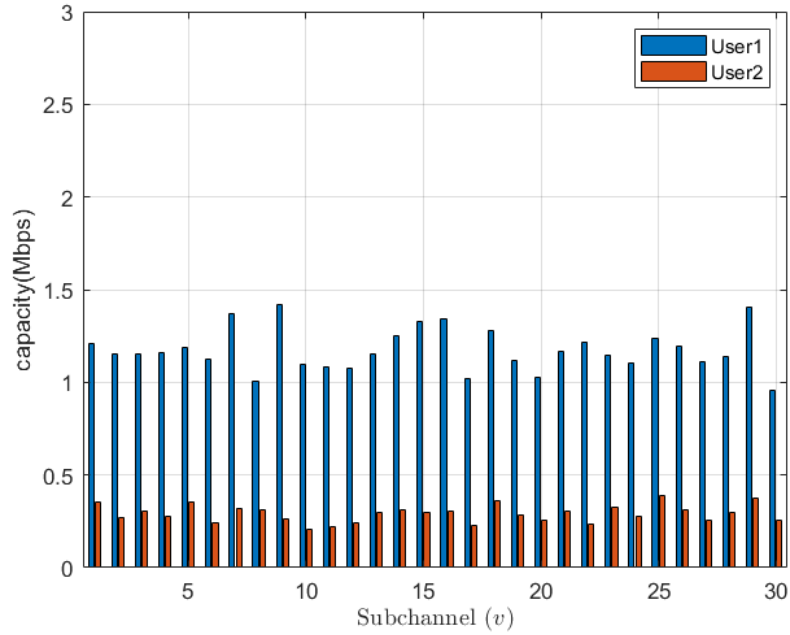


Figure 2.3: Capacity per user for per sub-channel

Table 2.3: Channel User Assignment Using CUSF

Sorted User	Assigned Channel
1	3
3	1
4	3
2	1
5	4
9	4
6	2
8	5
10	2
7	5

The impact of different power assignments is shown in Figure 2.4. It can be seen that the capacity increases inversely proportional to alpha. Furthermore, for the case of 5 channels and 10 users, we show in Table 2.2 a list of users sorted by each channel according to their respective gains before applying the CUSF algorithm. Then after applying the CUSF algorithm (before power assignment), we see the channel-user assignment in Table 2.3. According to the distribution of users, we see that the algorithm has, to a good degree, fairly distributed the users on channels, which helps in optimizing the overall system performance.

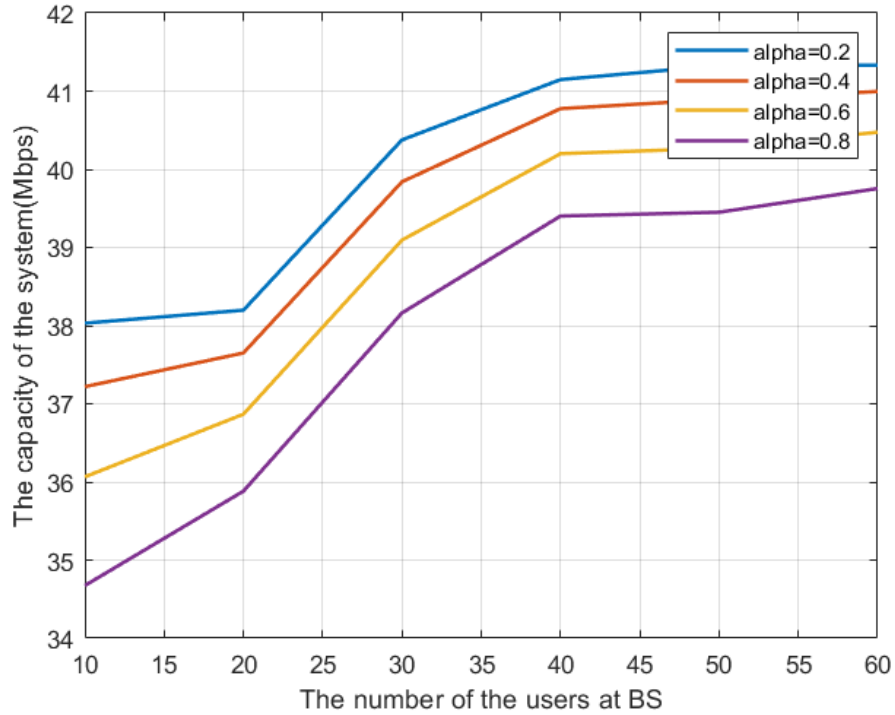


Figure 2.4: System capacity for different alpha values.

2.7 Conclusion

This chapter investigated the joint channel and power allocation problem in down-link NOMA systems for 5G and beyond networks. The original optimization problem was formulated as a weighted sum-rate maximization problem and shown to be non-convex and NP-hard. To obtain a practical solution, the problem was decoupled into two stages. First, a novel one-to-many matching-based channel allocation algorithm, namely CUSF, was proposed to assign users to sub-channels in a stable and fair manner. Second, power allocation was performed through water-filling across sub-channels and FTPC within each NOMA group. The simulation results demonstrated that the proposed framework improves system capacity and achieves better performance than the considered benchmark schemes. Overall, the chapter confirms that combining efficient channel assignment with structured power control provides a practical and effective solution for NOMA resource allocation.

Chapter 3

Energy-Efficient Resource Allocation in Mobile Edge Computing Using NOMA and Massive MIMO

3.1 Introduction

Since the 1980s, mobile networks have been undergoing continuous evolution, ushering in a new mobile network generation approximately every ten years. 5G technology, which supports enhanced mobile broadband, can achieve peak data rates of up to 10 Gbps. The 6G networks are projected to deliver extreme peak data rates surpassing 1 Tbps and introduce access to advanced edge computing capabilities with ultra low latency of less than ten nanoseconds [25].

Data collection and processing demands have increased unprecedentedly in the digital age due to the widespread use of mobile devices and the Internet of Things (IoT). In response to these needs, Mobile Edge Computing (MEC) has become a key technology that aims to move processing power closer to the data generation location. By processing data closer to the user at the edge rather than in remote centralized cloud data centers, MEC can achieve more processing speed, lower latency, and improved user experiences. MEC infrastructure installation and management are not without difficulties. Among these, energy consumption stands out as particularly important and has a significant impact on the environmental footprint, operational costs, and sustainability of MEC networks.

Deploying MEC nodes in a distributed manner, often in resource-constrained environments, necessitates optimizing their energy usage to ensure the long-term viability of edge computing services. In the context of processing the data on the user's mobile device (local processing), the energy consumption mainly depends on the CPU voltage supply, which proportionally controls the CPU speed. For this purpose, chip companies have implemented Dynamic Voltage Scaling (DVS), which allows the software to control the CPU voltage to save energy [129]. Furthermore, studies on user interfaces have indicated that response times ranging from 50 to 100 milliseconds have no significant impact on user thinking time [59]. Thus, unlike hard real-time applications, soft real-time applications (like multimedia) require only statistical performance guarantees. These arguments imply that running the CPU at maximum speed and consuming more energy has no benefit, and one should impose an acceptable delay time that satisfies the user's thinking on the one hand and reduces energy consumption on the other.

Energy consumption for data processing is inevitable. For cases when heavy processing is required, and mobile devices with limited energy resources (like rechargeable batteries) cannot provide such energy requirements for a specific duration, MEC provides the so-

lutions by introducing small processing centers (PC) located close to the BS and, hence, bringing computational resources closer to end-users. These PCs process the user’s data and return the results through the BS. However, data offloading requires a certain amount of energy from the user’s device. The key question addressed in this work is choosing between processing data locally or offloading it to a PC based on energy consumption. To the best of our knowledge, it introduces a comprehensive solution between local processing and data offloading in both NOMA and mMIMO channel structures. It is important to note that this chapter does not compare NOMA and mMIMO. This topic requires numerous considerations that fall beyond the scope of our work. Our work introduces the following novel achievements:

- **Energy Consumption Modeling:** This chapter presents a detailed analysis of energy consumption for local data processing, incorporating both deterministic and probabilistic models to provide a more robust understanding of energy dynamics.
- **Optimized Local Processing Model:** We propose an optimized energy consumption model for local processing, grounded in probability distribution, which offers a more efficient approach to energy utilization.
- **NOMA Channel Assignment:** A novel model for optimal channel assignment in NOMA is introduced, contributing to more effective management of communication resources.
- **NOMA Power Allocation:** We develop a power allocation strategy for NOMA, determined by the interaction between channel conditions and the computational resources of cloudlets, which are small edge servers located near the base station.
- **mMIMO Power Allocation:** Similarly, we provide a framework for power allocation in mMIMO, where channel conditions and cloudlet processing units jointly influence the power distribution.

3.2 Related Work

In this section, we review related works to provide insight into the topic. Many studies have addressed energy-efficient strategies in MEC, particularly through task offloading and resource allocation. In [90], the authors proposed reduced energy consumption using heuristic methods, while [5] introduced energy and resource allocation in a NOMA-based system. Similarly, [91] focused on optimizing power and resource allocation to reduce transaction time and improve throughput and spectral efficiency. Zhang et al. in [124] proposed an Energy-Efficient Computation Offloading (EECO) mechanism that optimizes radio resource allocation and offloading decisions to minimize energy consumption. In [118], they further explored the relationship between task execution latency and energy consumption.

Several other studies focused on minimizing energy and latency by improving task distribution between edge servers and user devices. In [71], the authors proposed splitting large tasks into smaller sub-tasks, distributed between the edge and devices. While [54] introduced an optimization problem for offloading tasks from user devices to fog nodes, aiming to reduce execution time and energy usage. Resource allocation efficiency was addressed by Wu et al. in [99], where they proposed the use of weighted Dominant Resource Fairness (DRF) for more effective resource management.

Further work in NOMA-based MEC systems emphasized resource optimization. In [107],

resource distribution between edge servers and user devices was optimized to reduce energy usage. Wan et al. [92] tackled computation offloading and resource allocation jointly in multi-access MEC systems using NOMA. Baidas in [9] explored resource allocation in clustered MEC networks using NOMA to maximize offloading efficiency. Zhang et al. in [121] addressed the issue of resource allocation in MEC networks through Service Function Chaining (SFC) and deep reinforcement learning algorithms, while Liu et al. [53] proposed a multi-user offloading and resource allocation strategy using competition algorithms and game model-based approaches. In contrast to these works, our study introduces a novel deterministic and probabilistic model for local processing, along with a new combined resource and power allocation method in NOMA.

Previous studies have investigated single-antenna MEC systems integrated with NOMA technology, focusing on sub-channel scheduling, task assignment, and power allocation. In contrast, multi-antenna NOMA systems have been explored for their potential to enhance the performance of multi-user MEC systems [94]. Additionally, the application of mMIMO in MEC systems has led to significant improvements in energy optimization and latency reduction under various constraints [32, 51, 112]. In our work, we aim to achieve optimal resource allocation by incorporating innovative local processing within mMIMO networks. We found that [110] has introduced NOMA and mMIMO in MEC networks, but our work differs in that we address not only delay minimization but also energy efficiency, which is crucial for mobile users with limited power resources. Moreover, our approach includes both NOMA and mMIMO, with a comprehensive framework that optimizes power and resource allocation for both local and remote processing, enhancing system performance and sustainability. Finally, unlike previous researches, which have primarily concentrated on improving offloading efficiency, we focus on optimizing local processing with novel mMIMO power allocation schemes. In summary, our work introduces an innovative approach to local processing and integrates it with two novel resource allocation scenarios: NOMA and mMIMO.

3.3 System Model

In this section, we assume a single-cell scenario where a BS serves a set of users \mathcal{K} . The available BW is divided into \mathcal{N} sub-channels, and each sub-channel has a BW b , where \mathcal{K} , \mathcal{N} , and b are defined as above. Each user $k \in \mathcal{K}$ is supposed to process a specific task locally or offload that task through BS to a processing center (PC). Generally, a PC can be a cloud with unlimited resources and processing units or a small cloudlet with limited resources. We find the latter more appropriate for our model and practical, as it can provide shorter latency than cloud computing, which requires uncontrolled latency for several reasons, one of which is the propagation delay.

For this reason, we consider the PC as a cloudlet with limited capacity and processing units and is typically colocated with the BS. A user's task S_k is characterized by the working load W_k , where W_k represents the number of Central Processing Unit (CPU) cycles required to accomplish the task S_k . Furthermore, in the case of data offloading, a user k is required to offload D_k bits to the cloudlet to accomplish the task S_k .

3.3.1 User local execution Energy

When executing a task on a mobile device, the CPU consumes most of the energy. The clock frequency of the chip scales almost linearly to the voltage, and the energy consumed per cycle is proportional to the square of the voltage, [47]. This relation allows us to

express the energy per CPU cycle consumed as a function of frequency f as follows:

$$\mathcal{E}_k(f) = \tau f^2 \quad (3.1)$$

Where τ is the effective switched capacitance that depends on the chip architecture of the mobile device [20], from (3.1), it can be seen that the CPU can save significant energy by running more slowly; for example, by running at half speed, it can reduce to quarter of the energy consumption. Generally, two factors restrict energy saving: the user experience, which dictates that the performance must meet the user's requirements, and the energy consumption of other components, such as memory storage and backlight, which increases with longer run times and may result in worse outcomes. Let C_k denote the number of CPU cycles required to process a single bit. Then:

$$W_k = D_k C_k \quad (3.2)$$

Dynamic Voltage Scaling (DVS), which enables software to adjust the processor's voltage, states that each cycle requires a specific frequency. As a result, we can express the total energy consumption of the mobile device for local processing as follows:

$$E_k^l = \sum_{i=1}^{W_k} \tau f_{k,i}^2 \quad (3.3)$$

Later we will elaborate more on (3.3) and find its optimized values.

3.3.2 User data offloading Energy

In the case of data offloading, we consider the Non-Orthogonal Multiple Access (NOMA) channel. In power-domain NOMA, each user sends a signal with power p_k , thereby increasing the rate of OMA. The receiver separates users based on signal strengths by implementing the Successive Interference Cancellation (SIC) method. SIC decodes the stronger signals, considers weaker signals as noise, and then decodes the correct signal. The NOMA system sets a threshold to limit the number of users it serves, primarily to reduce complexity and minimize error propagation. We denote the sub-group (cluster) of NOMA users on sub-channel c_n as $\mathcal{S}_n \subseteq \mathcal{K}$ such that $n \in \mathcal{N}$ and $1 \leq |\mathcal{S}_n| \leq S$. We distribute users in NOMA clusters, and within each cluster, we order users in descending order based on their signal strength.

The user's signal u_{jn} is received and subjected to SIC by subtracting the signal intended for a later receiver from the composite signal. This leads to an improvement in the signal-to-interference-plus-noise ratio (SINR). To decode the intended signal for u_{jn} , the interfering signals intended for other users u_{in} are decoded first, where $i > j$ and $i, j \in \mathcal{S}_n$. We do not decode the interfering signals of a lower order, instead treat them as noise. Therefore, we can express the SINR before SIC in the following way:

$$\text{SINR}_{jn} = \frac{p_{jn} g_{jn}}{\sum_{i=1, i \neq j}^S p_{in} g_{jn} + \sigma_n^2}, \quad (3.4)$$

Where p_{jn} and g_{jn} are respectively the transmitted power and channel gain of u_{jn} , σ is the noise power. Then, after SIC, the estimated SINR writes:

$$\text{SINR}_{jn} = \frac{p_{jn}g_{jn}}{\sum_{i=1}^{j-1} p_{in}g_{jn} + \sigma_n^2} \quad (3.5)$$

Based on the SINR from (3.5), the user rate can be expressed as:

$$R_{jn} = \log_2 \left(1 + \frac{p_{jn}g_{jn}}{\sum_{i=1}^{j-1} p_{in}g_{jn} + \sigma_n^2} \right) \quad (3.6)$$

Let us introduce the binary variable $x_{jn} \in \{0, 1\}$ such that $x_{jn} = 1$ if user u_{jn} is multiplexed on sub-channel n and $x_{jn} = 0$ otherwise. We then express the total rate that user u_j can achieve as follows:

$$R_j = \sum_{n=1}^N x_{jn} R_{jn} \quad (3.7)$$

The energy consumed by u_j can then be expressed as follows:

$$E_j^r = T_j^t P_j \quad (3.8)$$

Where T_j^t and P_j are respectively the transmission time (in seconds) and total transmission power (in watts). both can be expressed as:

$$T_j^t = \frac{D_j}{R_j} \quad (3.9)$$

$$P_j = \sum_{n=1}^N x_{jn} p_{jn} \quad (3.10)$$

The cloudlet introduces another time delay for data processing. Consider that the cloudlet consists of multiple computational resources (CR) of total number \mathcal{M} , each with a capacity of c . Assigning q CRs to user u_j results in the following processing time:

$$T_j^c = \frac{W_j}{q_j c}. \quad (3.11)$$

Both (3.9) and (3.11) are restricted by the total deadline time T_j^D such that:

$$\begin{aligned} T_j^t + T_j^c &\leq T_j^D \\ T_j^t &\leq T_j^D - T_j^c \end{aligned} \quad (3.12)$$

From (3.9) and (3.12) we reach the following important conclusion:

$$R_j \geq \frac{D_j}{T_j^D - T_j^c} \quad (3.13)$$

3.4 Problem formulation

In this section, we formulate the optimization problem to reduce total energy consumption. We must resolve the primary challenges by choosing either local or remote processing and then manage the remote processing to allocate devices in the NOMA clusters opti-

mally, thereby boosting their data rates. Given these considerations, we can formulate the optimization problem as follows:

Objective: minimize the total energy consumption by jointly deciding local execution or remote offloading, transmit power allocation, channel assignment, CPU frequency, and cloudlet resource allocation.

$$P3.1 : \min_{\mathbf{x}, \mathbf{p}, \mathbf{f}, \mathbf{q}} \sum_{k \in \mathcal{K}} (\alpha_k E_k^l + (1 - \alpha_k) E_k^r) \quad (3.14a)$$

Subject to:

$$R_j \geq \frac{D_j}{T_j^D - T_j^c}, \quad \forall j \in \mathcal{K}, \quad (3.15a)$$

$$\sum_{n \in \mathcal{N}} p_{kn} \leq P_k^{\max}, \quad \forall k \in \mathcal{K}, \quad (3.15b)$$

$$\sum_{n \in \mathcal{N}} x_{kn} \leq 1, \quad \forall k \in \mathcal{K}, \quad (3.15c)$$

$$s_l \leq \sum_{k \in \mathcal{K}} x_{kn} \leq S, \quad \forall n \in \mathcal{N}, \quad (3.15d)$$

$$x_{kn} \in \{0, 1\}, \quad \forall k \in \mathcal{K}, n \in \mathcal{N}, \quad (3.15e)$$

$$\alpha_k \in \{0, 1\}, \quad \forall k \in \mathcal{K}, \quad (3.15f)$$

$$p_{kn} \geq 0, \quad \forall k \in \mathcal{K}, n \in \mathcal{N}, \quad (3.15g)$$

$$q_k \neq q_{k'}, \quad \forall k, k' \in \mathcal{K}, k \neq k', \quad (3.15h)$$

$$\sum_{k \in \mathcal{K}} q_k \leq \mathcal{M}. \quad (3.15i)$$

Constants and parameters:

- \mathcal{K} : set of users
- \mathcal{N} : set of sub-channels
- E_k^l : energy consumed by user k for local processing
- E_k^r : energy consumed by user k for remote offloading
- R_j : achievable rate of user j
- D_j : data size of user j
- T_j^D : delay deadline of user j
- T_j^c : cloudlet processing time of user j
- P_k^{\max} : maximum transmit power of user k
- s_l : minimum number of users multiplexed per sub-channel
- S : maximum number of users multiplexed per sub-channel
- \mathcal{M} : total number of available cloudlet computational resources

Optimization variables:

- x_{kn} : binary channel assignment variable, where $x_{kn} = 1$ if user k is assigned to sub-channel n , and $x_{kn} = 0$ otherwise
- p_{kn} : transmit power allocated to user k on sub-channel n
- α_k : binary decision variable, where $\alpha_k = 1$ indicates local processing and $\alpha_k = 0$ indicates remote offloading
- q_k : computational resources allocated to user k in the cloudlet

In problem $P3.1$, the constraint (4.1c) ensures the minimum rate required to satisfy user needs based on wireless channel conditions and cloudlet limitations. Constraint (4.1d), to satisfy each user's maximum transmission. In a one-to-many scenario, constraint (4.1e) ensures that each user acquires only one sub-channel. Constraint (4.1f) to guarantee an upper and lower limit for NOMA clusters based on the limitations of receiver complexity and to limit receiver error propagation. Constraint (3.15h) to guarantee that resources used by each user in the cloudlet remain during processing and no preemptive processing is applied. Finally, constraint (3.15i) ensures that the number of PCs in the cloudlet is not exceeded.

Problem ($P3.1$) is a mixed-integer nonlinear programming (MINLP) problem. It is also NP-hard. Additionally, the problem's limitations prevent it from offering the best solution for local processing. We have chosen to address these challenges in multiple steps, heuristically solving the complex parts of $P3.1$ to simplify and make it solvable, as we will demonstrate next.

3.5 Offline processing

This section discusses the price (energy consumption) the user pays to execute the task locally. The processor's speed depends on the degree of knowledge; if the working load is already known, the best scenario is to assign a fixed constant speed for the entire task. Conversely, if the precise working load remains unknown, we will assign the speed according to the statistical distribution.

3.5.1 Tasks with known working load

Given that each cycle will require a specific CPU frequency and the task must meet the deadline, we can assume the following:

$$\sum_{i=1}^{W_k} \frac{1}{f_{k,i}} \leq T_k^D \quad (3.16)$$

Assuming a fixed frequency through all cycles for the known working load, then (3.16) writes:

$$\begin{aligned} \frac{W_k}{f_k} &\leq T_k^D, \\ f_k &\geq \frac{W_k}{T_k^D} \end{aligned} \quad (3.17)$$

From (3.3) and enforcing the lower bound on (3.17), we reach the optimum minimum energy consumption as follows:

$$E_{k_{\min}}^c = \frac{\tau (W_k)^3}{(T_k^D)^2} \quad (3.18)$$

3.5.2 Tasks with unknown working load

In this section, we consider tasks with soft real-time resource demands. Soft real-time demands, such as those found in multimedia applications, differ from hard real-time demands in that they solely necessitate statistical performance guarantees [59]. DVS is a well-known operating system principle that saves CPU energy by controlling the voltage linearly proportional to the frequency and its squared value proportional to the energy. Predicting CPU demands for a task is crucial because overestimating them will waste energy, while underestimating them will degrade the application's performance. For tasks with unknown work requirements but only a known distribution, the best scenario is to apply variable speed by running slowly, as it might require little work, then increasing the speed gradually until reaching the maximum if the task exceeds a specific deadline.

Because of the unknown cycles required, we modify (3.2) to become:

$$W_k = D_k \widehat{C}_k \quad (3.19)$$

Where \widehat{C}_k is a random variable drawn from a known distribution. Let us denote the Cumulative Distribution Function (CDF) of the work by F . Then:

$$F(w) = P[W \leq w] \quad (3.20)$$

Where (3.20) represents the probability that the task requires no more than w cycles. Our goal is to predict the minimum expected energy consumption. To achieve this goal, we formulate the following optimization problem:

$$P3.2 : \min_{f(w)} \sum_{w=1}^W \tau F^c(w) [f(w)]^2 \quad (3.21a)$$

$$\text{s.t.} \sum_{w=1}^W \frac{1}{f(w)} \leq T^D, \quad (3.21b)$$

$$0 \leq f(w) \quad (3.21c)$$

Where $F^c(w)$ is the complementary cumulative distribution function (CCDF) and represents the probability that the work gets done. It is related to $F(w)$ as: $F^c(w) = 1 - F(w)$. Problem $P3.2$ represents the minimum amount of energy consumed, taking into account all CPU cycles. Constraint 3.21b imposes a deadline that should be met.

We first notice that (3.21c) is implicit in the objective function, so we tackle the problem $P3.2$ by implementing Karush-Kuhn-Tucker (KKT) conditions as shown in Appendix B. Based on the result achieved for $f(w)$ in (B-8), we can express the optimal statistical energy in $P3.2$ as follows:

$$E^s = \frac{\tau}{(T^d)^2} \left[\sum_{w=1}^W (F^c(w))^{1/3} \right]^3 \quad (3.22)$$

Next, we need to decide on the distribution of task work. For this purpose, there are two approaches: parametric (known distributions) and non-parametric (distribution driven from samples). Based on the difference between the two methods, it is clear that the non-parametric approach is more accurate. This can be crucial in specific situations, like optimally determining the CPU frequency. However, our requirement is less strict, and we can revert to an estimation with accepted accuracy to help us decide which method to choose. In [59], it is shown that normal distribution is an acceptable approximation of non-parametric methods, so we consider this a valid assumption to represent the required distribution. Therefore, the CCDF can be expressed as follows:

$$F^c(w) = \frac{1}{2} - \frac{1}{2} \operatorname{erf} \left(\frac{w - \mu}{\sigma\sqrt{2}} \right) \quad (3.23)$$

Then (3.22) can be expressed as:

$$E^s = \frac{\tau}{(T^d)^2} \left[\sum_{w=1}^W \left(\frac{1}{2} - \frac{1}{2} \operatorname{erf} \left(\frac{w - \mu}{\sigma\sqrt{2}} \right) \right)^{1/3} \right]^3 \quad (3.24)$$

Where erf is the error function, τ and T^d are constants, μ is the mean, and σ is the standard deviation. Next, we wish to find an acceptable estimation for W expressed in (3.24). To achieve this, we want $F(w)$ in (3.20) to satisfy a certain degree of confidence, such as:

$$F(W) \geq \delta, \quad (3.25)$$

Where δ is a fraction close to 1 and can be chosen according to the accuracy required, for example, between (0.95-0.98). From (3.23) and (3.25) we can derive an expression for W as follows:

$$\begin{aligned} \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{W - \mu}{\sigma\sqrt{2}} \right) \right] &\geq \delta \\ \frac{W - \mu}{\sigma\sqrt{2}} &= \operatorname{erf}^{-1}(2\delta - 1) \\ W &= \mu + \sigma\sqrt{2} \cdot \operatorname{erf}^{-1}(2\delta - 1) \end{aligned} \quad (3.26)$$

3.6 Data offloading

This section discusses when the device decides to offload data to the PC. Later, we will enforce a condition for deciding whether to offload or process data locally. To reduce the complexity of problem P3.1, we work on reducing constraints and eliminating the integer variables. For this purpose, we resort to applying a heuristic algorithm that tackles the problem of clustering and assigning users to their optimum respective channels.

We will study two different cases for network configuration: NOMA and mMIMO.

3.6.1 NOMA network configuration

We find that the Gale-Shapely marriage algorithm is most appropriate for our model. The algorithm optimizes for the proposing party and selects their preferred option, which aligns with our search for the most efficient up-link channels for data offloading. Generally, the

algorithm operates on a one-to-one basis, but we modify it to operate on a many-to-one basis to align with our NOMA model. Furthermore, we assume that BS centralizes and controls the entire process, as expected.

We start the matching process by acknowledging that the two sets: \mathcal{K} and \mathcal{N} are two disjoint and finite sets, and each user has preferences over sub-channels, and each sub-channel has preferences over users, and we use channel gain as the decision factor between the sets. We assume that the preference is transitive, meaning that if u_k prefers n_i over n_j and n_j over n_l , then it prefers n_i over n_l . The ordered lists can be represented as follows:

$$O(u_k) = \{n_i, n_j, u_k, n_l, \dots\}, \quad (3.27)$$

$$O(n_k) = \{u_i, u_j, n_k, u_l, \dots\} \quad (3.28)$$

Where (3.27) shows that u_k prefers n_i over n_j . We introduced u_k in $O(u_k)$ to show that after n_j all sub-channels are of unacceptable channel gains and can't be considered. The same goes for (3.28). Also, let us denote the preference process as $n_i \succeq_{u_k} n_j$, which means user u_k prefers n_i at least as n_j , and $u_k \preceq_{n_i} u_l$ as channel n_i prefers u_k at least as u_l . Furthermore, as expressed in (4.1e) and (4.1f), each sub-channel can be matched up to S users, while each user can only be matched to one sub-channel in a many-to-one scenario. Then, the matching function (μ_p) can be applied such that:

1. $|\mu_p(u_k)| = 1, \forall u_k \in \mathcal{K}$
2. $|\mu_p(n_i)| = S, \forall n_i \in \mathcal{N}$. If the number of acceptable users is less than S , then the rest of the positions will be filled by n_i .
3. $|\mu_p(u_k)| = n_i \iff |\mu_p(n_i)| = u_k$

The modified algorithm then operates as follows: The two groups (users and sub-channels) have preferences over other group members based on their channel gains. It operates in iterations; each unassigned user proposes to its most preferred sub-channel. Sub-channels, upon receiving proposals, tentatively accept users based on their channel gains and available slots or retain the top preferred users if their slots are full. If a sub-channel receives multiple proposals, it keeps the most preferred users and rejects others. The process persists until sub-channels assign all users. The whole process is shown in algorithm 1.

Algorithm 3 Stable Matching of Users to Sub-channels (SMUSC)

Input:

- $O(u_k)$: List of users and their preferences for sub-channels
- $O(n_i)$: List of sub-channels and their preferences for users

Output: Stable matching of users to sub-channels

```

1: Step 1: Initialization
2: for each user  $u_k \in \mathcal{K}$  do
3:   Set  $\mu_p(u_k) \leftarrow u_k$  (initially unassigned)
4: end for
5: for each sub-channel  $n_i \in \mathcal{N}$  do
6:   Set  $\mu_p(n_i) \leftarrow \{n_i, \dots, n_i\}$  (initialize available slots)
7: end for
8: Step 2: Proposal Loop
9: while there exists any user  $u_k$  such that  $|\mu_p(u_k)| = 0$  do
10:  Choose an unassigned user  $u_k$  who hasn't proposed yet
11:   $u_k$  proposes to its most preferred sub-channel  $n_i$ 
12:  if sub-channel  $n_i$  has available slots then
13:    Tentatively accept  $u_k$ 
14:  else
15:    Compare  $u_k$  with the least preferred user in  $\mu_p(n_i)$ 
16:    if  $u_k$  is more preferred than the least preferred user then
17:      Replace the least preferred user with  $u_k$ 
18:    end if
19:  end if
20:  Update  $u_k$ 's status to assigned
21: end while
22: Step 3: Finalization
23: Output the final stable matching  $\mu_p$ 

```

3.6.2 NOMA Power allocation

Based on the above information, we have already determined several constraints in problem P3.1. Next, we will focus on the optimal power allocation for users based on their NOMA clustering. Let us introduce the required optimization problem to address this issue:

$$P3.3 : \min_{\mathbf{p}, \mathbf{q}} \sum_{k \in \mathcal{K}} E_k^r \quad (3.29a)$$

$$\text{s.t. } R_k \geq \frac{D_k}{T_k^D - T_k^c} \quad (3.29b)$$

$$\sum_{n \in \mathcal{N}} p_{kn} \leq P_k^{\max}, k \in \mathcal{K} \quad (3.29c)$$

$$0 \leq p_{kn}, k \in \mathcal{K} \quad (3.29d)$$

$$q_k \neq q_{k'}, k, k' \in \mathcal{K} \quad (3.29e)$$

$$\sum_{k \in \mathcal{K}} q_k \leq \mathcal{M} \quad (3.29f)$$

such that: D_k, T_k^D are known constants,

$$R_k = \log_2 \left(1 + \frac{p_k g_k}{\sum_{i=1}^{j-1} p_i g_k + \sigma_n^2} \right) \quad (3.30)$$

$$E_k^r = \frac{p_k D_k}{R_k} \quad (3.31)$$

Problem *P3.3* represents both power and computational resource allocation. Although the channels are non-interfering, the shared resources make the power allocation interdependent, complicating our problem. To tackle this problem, we first ignore constraints (3.29e) and (3.29f), then we find the best resource allocation scenario based on the obtained energy reduction. Furthermore, we notice that the objective function comprises a fractional term. The numerator is a linear function concerning p_k , and the denominator is already considered in (3.29b). Given the information above, we can simplify the optimization problem as follows:

$$P3.4 : \min_{\mathbf{p}} \sum_{k \in \mathcal{K}} p_k \quad (3.32a)$$

$$\text{s.t. } R_k \geq \frac{D_k}{T_k^D - T_k^c} \quad (3.32b)$$

$$\sum_{n \in \mathcal{N}} p_{kn} \leq P_k^{\max}, k \in \mathcal{K} \quad (3.32c)$$

$$0 \leq p_{kn} \quad (3.32d)$$

The constraint (3.32b) in *P3.4*, considering (3.30), makes the problem non-convex. We tackle this issue by relaxing the non-convex problem to eliminate non-convex structure. First, we make use of the following formula:

$$\log_2 x = \frac{\ln x}{\ln(2)} \quad (3.33)$$

where \ln denotes the natural logarithm. Let us define the constant $\nu = \ln(2)$; then (3.33) can be written as

$$\log_2 x = \frac{1}{\nu} \ln x \quad (3.34)$$

Furthermore, we make use of the following lower bound [73]:

$$\zeta \log(z) + \gamma \leq \log(1 + z) \quad (3.35)$$

The bound is tight at $z = z_0$ such that:

$$\zeta = \frac{z_0}{1 + z_0} \quad (3.36a)$$

$$\gamma = \log(1 + z_0) - \zeta \log(z_0) \quad (3.36b)$$

Where z_0 is the SINR and expressed as:

$$z_0 = \frac{p_k g_k}{\sum_{i=1}^{j-1} p_i g_k + \sigma_n^2}, \quad (3.37)$$

based on (3.35), and (3.36), we introduce the following inequality:

$$\begin{aligned} & \log \left(1 + \frac{p_k g_k}{\sum_{i=1}^{j-1} p_i g_k + \sigma_n^2} \right) \geq \\ & \geq \zeta \log \left(\frac{p_k g_k}{\sum_{i=1}^{j-1} p_i g_k + \sigma_n^2} \right) + \gamma \end{aligned} \quad (3.38)$$

With ζ and γ defined as above, and because ν is positive, it is dropped from both sides of (3.38). The right-hand side of (3.38) can be expressed as:

$$\begin{aligned} & \zeta \log \left(\frac{p_k g_k}{\sum_{i=1}^{j-1} p_i g_k + \sigma_n^2} \right) + \gamma = \\ & \zeta \left[\log(p_k g_k) - \log \left(\sum_{i=1}^{k-1} g_k p_i + \sigma_n^2 \right) \right] + \gamma \end{aligned} \quad (3.39)$$

Still, we have a non-convex form in (3.39). Furthermore, we notice that in [15, p. 72], the log-sum-exp is convex. So, by using the fact that $x = e^{\log(x)}$, and if we express $\widehat{p}_k := \log(p_k)$, we modify (3.39) into a convenient form as follows:

$$\begin{aligned} & \zeta \left[\log(p_k g_k) - \log \left(\sum_{i=1}^{k-1} g_k p_i + \sigma_n^2 \right) \right] + \gamma = \\ & \zeta \left[\widehat{p}_k + \log(g_k) - \log \left(\sum_{i=1}^{k-1} g_k e^{\widehat{p}_i} + \sigma_n^2 \right) \right] + \gamma \end{aligned} \quad (3.40)$$

We can see that (3.40) is structured as the sum of linear and convex functions, which forms a convex function. If we express (3.40) as modified rate \widehat{R}_k , then the problem P3.4 writes:

$$P3.5 : \min_{\widehat{p}} \sum_{k \in \mathcal{K}} e^{\widehat{p}_k} \quad (3.41a)$$

s.t.

$$\begin{aligned} & \zeta \left[\widehat{p}_k + \log(g_k) - \log \left(\sum_{i=1}^{k-1} g_k e^{\widehat{p}_i} + \sigma_n^2 \right) \right] + \gamma \\ & \geq \frac{\nu D_k}{T_k^D - T_k^c} \end{aligned} \quad (3.41b)$$

$$\sum_{n \in \mathcal{N}} e^{\widehat{p}_k} \leq \widehat{P}_k^{\max}, k \in \mathcal{K} \quad (3.41c)$$

$$0 \leq e^{\widehat{p}_k} \quad (3.41d)$$

The constraint (3.41d) is redundant as the right-hand side will always be positive. We can solve Problem P3.5 in a convex form using efficient convex optimisation algorithms. For this purpose, we apply KKT to find the optimal power allocation as shown in Appendix C. Next, we follow the steps outlined in Algorithm 2 to determine the optimal power allocation.

Algorithm 4 Power Allocation

Input:

- P_k : Power levels [W]
- D_k : Data demands [bits]
- T_k^D : Delay tolerance [s]
- T_k^c : Channel conditions [dimensionless]

Output: Power allocation for each user p_k 1: **Step 1: Initialization**

2: Set initial conditions:

- $z_0 \gg 1 \rightarrow \{\zeta_k = 1, \gamma_k = 0\}$
- Initialize Lagrange multipliers λ_k, μ_k
- Set iteration index $i = 1$
- Set convergence threshold δ

3: **Step 2: Iterative Optimization**4: **repeat**5: Compute gradients for λ_k and μ_k (See equations B-9 and B-10)

6: Update Lagrange multipliers:

$$\lambda_k^{(i+1)} = \left[\lambda_k^{(i)} - \eta_\lambda \cdot \text{Grad}_{\lambda_k} \right]^+$$
$$\mu_k^{(i+1)} = \left[\mu_k^{(i)} + \eta_\mu \cdot \text{Grad}_{\mu_k} \right]^+$$

7: Update power value:

$$\hat{p}_k = \log \left(\frac{\mu_k \zeta}{\lambda_k + 1} \right)$$

8: Check for convergence:

- Ensure primal and dual feasibility conditions
- Check complementary slackness

9: Increment iteration index $i \leftarrow i + 1$ 10: **until** convergence criteria are met11: **Step 3: Final Power Update**

12: Set actual power:

$$p_k = e^{\hat{p}_k}$$

13: Update $\{\zeta_k, \gamma_k\}$ based on equation (35)14: **End Procedure**

Where $[x]^+ = \max(0, x)$, and η is the step size. By applying **Alg. 2**, we obtain the optimum power for mobile up-link in the NOMA scenario. If we recall problem *P3.3*, our last challenge is to find the best approach to assigning computation resources. From (3.12) we get:

$$T_j^c \leq T_j^D - T_j^t \quad (3.42)$$

Combining (3.9) in (3.42) we get:

$$T_j^c \leq T_j^D - \frac{D_j}{R_j} \quad (3.43)$$

Also, writing (3.11) in (3.43), we get a final expression for q_j as follows:

$$q_j \geq \frac{W_j}{c \left(T_j^D - \frac{D_j}{R_j} \right)} \quad (3.44)$$

Where R_j from (3.6) with p_j from **Alg. 2**. In (3.44), the lower limit is shown. We can then assign resources based on availability to enhance the margin for each device.

3.6.3 Massive Multiple Input Multiple Output

When considering mMIMO network configuration, we also consider the multi-user case by wisely applying spatial multiplexing through beam forming. In this scenario, we consider that the number of antennas (M) is large enough to achieve channel hardening and favorable condition as defined in [64]. This means the channel gain is arbitrarily close to its mean, and the normalized channels of two users become asymptotically orthogonal.

Based on this scenario, we also consider a SIMO up-link, which means users are equipped with a single antenna, whereas BS serves users through multiple antennas. The channel between user k and BS is $\mathbf{h}_k \in \mathbb{C}^M$. Furthermore, we consider the antenna elements equally spaced, with an interdistance of d forming the Uniform Linear Array (ULA). Because of the nature of the network, the BS is elevated and has no near-field scatters while users suffer from scattering; thus, each multipath component results in a plane wave reaching the array from a particular angle ϕ . Then the total array response writes:

$$\mathbf{c}(\phi) = g_l [c_1(\phi), c_2(\phi), \dots, c_M(\phi)] \quad (3.45)$$

Where

$$c_m(\phi) = e^{-j2\pi(m-1)\left(\frac{d}{\lambda}\right)\sin(\phi)}, \quad m \in [1, 2, \dots, M] \quad (3.46)$$

where $g_l \in \mathbb{C}$ represents the gain of the channel from multi-path component l , λ is the wavelength, and $\phi \in [0, 2\pi]$ is the azimuth angle to the user relative to the bore-sight of the array. The channel between user k and BS is a superposition of multipath components, as follows:

$$\mathbf{h}_k = \sum_l^L \mathbf{c}_k(\phi_l) \quad (3.47)$$

as $L \rightarrow \infty$, and from the central limit theorem we find that $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_k)$. Where \mathbf{R} is the spatial correlation matrix and can be defined as:

$$\mathbf{R}_k = \mathbb{E} \left\{ \sum_l^L \mathbf{c}_k(\phi_l) \mathbf{c}_k^H(\phi_l) \right\} \quad (3.48)$$

$\mathcal{R}_k \in \mathbb{C}^{M \times M}$. Then, from (3.46), any element in \mathcal{R}_k can be found as follows:

$$\mathcal{R}_k^{m,m'} = \sum_l^L \mathbb{E}[|g_l|^2 e^{-j2\pi(m-1)(\frac{d}{\lambda})\sin(\phi_l)} e^{j2\pi(m'-1)(\frac{d}{\lambda})\sin(\phi_l)}] \quad (3.49)$$

$$= \sum_l^L \mathbb{E}[|g_l|^2] \int e^{-j2\pi(m-m')(\frac{d}{\lambda})\sin(\phi)} f(\phi) d\phi \quad (3.50)$$

Where $f(\phi)$ is the PDF. Furthermore, ϕ is considered as a combination of a nominal angle with a deviation from it. In the literature, the deviation can be considered Gaussian, Laplace, or uniformly distributed. We base the expectation on the angle of arrival. Furthermore, based on the favorable channel condition mentioned above, one can find the expression for the first part of (3.50) as follows:

$$\mathbb{E}[|g_l|^2] = \frac{1}{M} \text{tr}(\mathcal{R}_k) \quad (3.51)$$

Where $\text{tr}(\cdot)$ is the trace. $f(\phi)$ is the probability density function. For uncorrelated channels, \mathcal{R}_k would form a diagonal matrix.

The up-link received signal can be expressed as:

$$\mathbf{y} = \sum_{k=1}^K \mathbf{h}_k s_k + \mathbf{n} \quad (3.52)$$

Where $\mathbf{y} \in \mathbb{C}^M$, $s \sim \mathcal{N}_{\mathbb{C}}(0, p)$, and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \sigma^2 \mathbf{I}_M)$.

For signal reception, the received signal is correlated with a beam-forming vector to detect the signal from user k as follows:

$$\mathbf{w}_k^H \mathbf{y} = \sum_{k=1}^K \mathbf{w}_k^H \mathbf{h}_k s_k + \mathbf{n} \quad (3.53)$$

From (3.53), the rate can be expressed as:

$$\text{SINR}_k = \frac{p_k |\mathbf{w}_k^H \mathbf{h}_k|^2}{\sum_{j \neq k} P_j |\mathbf{w}_k^H \mathbf{h}_j|^2 + \mathbf{w}_k^H \mathcal{R}_n \mathbf{w}_k} \quad (3.54)$$

Where \mathcal{R}_n is the noise covariance matrix, and in the case of white noise, it reduces to $\mathbf{R}_n = \sigma^2 \mathbf{I}$. Then (3.54) reduces to:

$$\text{SINR}_k = \frac{p_k |\mathbf{w}_k^H \mathbf{h}_k|^2}{\sum_{j \neq k} p_j |\mathbf{w}_k^H \mathbf{h}_j|^2 + \sigma^2 \|\mathbf{w}_k\|^2} \quad (3.55)$$

Here we assume that the channel for each user k is known (estimated) at the BS as $\hat{\mathbf{h}}_k$, then the channel matrix estimation for all users as:

$$\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_K] \quad (3.56)$$

Finally, by applying zero forcing receive combining we find that :

$$\mathbf{W} = \hat{\mathbf{H}}(\hat{\mathbf{H}}^H \hat{\mathbf{H}})^{-1} \quad (3.57)$$

Next, for formulate the optimization problem for up-link power control.

$$P3.6 : \min \sum_{k \in \mathcal{K}} p_k \quad (3.58a)$$

$$\text{s.t. } \theta \leq \frac{R_k}{R_k^{\min}} \quad (3.58b)$$

$$0 \leq p_k \leq P^{\max} \quad (3.58c)$$

Where:

$$R_k = \log_2 \left(1 + \frac{p_k |\mathbf{w}_k^H \mathbf{h}_k|^2}{\sum_{j \neq k} p_j |\mathbf{w}_k^H \mathbf{h}_j|^2 + \sigma^2 \|\mathbf{w}_k\|^2} \right) \quad (3.59)$$

$$R_k^{\min} = \frac{D_k}{T_k^D - T_k^c} \quad (3.60)$$

In *P3.6*, we introduced θ in constraint (3.58b) as a fairness parameter. In general $\theta \geq 1$, where the case $\theta = 1$ guarantees the minimum rate. We proposed constraint (3.58c) because of the expected large number of users to enforce a balance in the received power levels across users, ensuring that no user's signal is disproportionately stronger or weaker than others. This prevents unfair resource allocation and minimizes interference, leading to a more stable and efficient network. Because of the fractional form of the SINR, the problem *P3.6* is nonconvex.

Next, we will work on solving *P3.6*; to achieve this, we will implement the Concave-Convex Procedure (CCCP). The objective function is linear so that we will work on the nonconvex part (3.58b). Let us rewrite it as follows:

$$R_k = \log_2 \left(1 + \sum_j p_j |\mathbf{w}_k^H \mathbf{h}_j|^2 + \sigma^2 \|\mathbf{w}_k\|^2 \right) - \log_2 \left(1 + \sum_{j \neq k} p_j |\mathbf{w}_k^H \mathbf{h}_j|^2 + \sigma^2 \|\mathbf{w}_k\|^2 \right) \quad (3.61)$$

$$= f_1(\mathbf{p}) - f_2(\mathbf{p}) \quad (3.62)$$

Where (3.62) represents a Difference of Convex (DC) function. Next, we implement Taylor expansion to linearize the second term of (3.62):

$$f_2(\mathbf{p}) = f_2(\mathbf{p}^{(t)}) + (\mathbf{p} - \mathbf{p}^{(t)})^T \nabla f_2(\mathbf{p}^{(t)}) \quad (3.63)$$

Where $\mathbf{p}^{(t)}$ is the power allocation vector at the t_{th} step (or iteration) of the CCCP algorithm and:

$$\nabla f_2(\mathbf{p}^{(t)}) = \frac{1}{\ln(2)} \cdot \frac{|\mathbf{w}_k^H \mathbf{h}_j|^2}{\sum_{j \neq k} p_j^{(t)} |\mathbf{w}_k^H \mathbf{h}_j|^2 + \sigma^2 \|\mathbf{w}_k\|^2} \quad (3.64)$$

By applying CCCP, we have transformed problem *P3.6* into a convex optimization problem, making it solvable with known convex optimization tools. However, the problem might present high complexity if the number of users is high. To reduce the complexity, we propose user clustering before applying CCCP.

When defining a clustering criterion, we discover that clustering based on AoA could be

deceptive, as our system presumes a localized scattering around the users, whereas the BS remains elevated. Channel gain is a good criterion, as it plays a significant role in deciding the power allocations. In algorithm 3, we apply K-means clustering.

Algorithm 5 K-means Clustering Based on Channel Gain

Input:

- K : Number of users
- \mathbf{h}_k for $k = 1, 2, \dots, K$: Channel vectors
- N_{clusters} : Number of clusters
- T_{max} : Maximum number of iterations
- ϵ : Convergence threshold

Output:

- Cluster assignments \mathcal{C}_i for each user k
- Cluster centers $\boldsymbol{\mu}_i$

- 1: **Step 1: Compute Channel Gains**
- 2: Compute channel gain for each user: $g_k = \|\mathbf{h}_k\|$ for each user k
- 3: **Step 2: Initialize Cluster Centers**
- 4: Randomly initialize N_{clusters} cluster centers $\boldsymbol{\mu}_i$ from $\{g_k\}$
- 5: **Step 3: Repeat Until Convergence or T_{max} Iterations**
- 6: **for** each iteration $t = 1$ to T_{max} **do**
- 7: **Assignment Step**
- 8: Assign each user k to the nearest cluster center:

$$\mathcal{C}_i = \left\{ k \mid \min_j |g_k - \boldsymbol{\mu}_j| \right\}$$

- 9: **Update Step**
- 10: Update the cluster centers:

$$\boldsymbol{\mu}_i = \frac{1}{|\mathcal{C}_i|} \sum_{k \in \mathcal{C}_i} g_k \quad \text{for each cluster } i$$

- 11: **Convergence Check**
 - 12: **if** $\max_i |\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_i^{(t-1)}| < \epsilon$ **then**
 - 13: Break loop
 - 14: **end if**
 - 15: **end for**
 - 16: **Step 4: Output Final Clusters**
 - 17: Return final cluster assignments \mathcal{C}_i and centers $\boldsymbol{\mu}_i$
-

To solve the problem, we first cluster users, then apply CCCP within each cluster and manage inter-cluster interference by considering each cluster as a single user and applying scaling factors to each cluster based on its interference contributions to other clusters. Algorithm 2 lists the detailed steps for solving the optimization problem.

Algorithm 6 Power Allocation in mMIMO Uplink using K-means Clustering and CCCP

Input:

- K : Number of users
- \mathbf{h}_k for $k = 1, 2, \dots, K$: Channel vectors
- P_k^{\max} : Maximum power for each user
- R_k^{\min} : Minimum rate for each user
- N_{clusters} : Number of clusters
- T_{\max} : Maximum number of CCCP iterations
- ϵ : Convergence tolerance
- β : Scaling factor parameter

Output: Optimized power allocation p_k for each user k

- 1: **Step 1: K-means Clustering**
 - 2: Perform K-means clustering to group users based on their channel vectors
 - 3: **Step 2: Intra-Cluster Power Allocation (CCCP)**
 - 4: **for** each cluster \mathcal{C}_i **do**
 - 5: Initialize power values $p_k^{(0)}$ for all $k \in \mathcal{C}_i$
 - 6: **for** iteration $t = 1$ to T_{\max} **do**
 - 7: Linearize the non-convex part of the SINR constraints
 - 8: Solve the convex subproblem to update power values $p_k^{(t+1)}$
 - 9: **if** $\|p_k^{(t+1)} - p_k^{(t)}\| < \epsilon$ **then**
 - 10: Stop the iteration
 - 11: **end if**
 - 12: **end for**
 - 13: **end for**
 - 14: **Step 3: Inter-Cluster Coordination**
 - 15: **for** each cluster \mathcal{C}_i **do**
 - 16: Calculate inter-cluster interference I_{ij} for all clusters
 - 17: Determine scaling factor:
$$\alpha_i = \frac{1}{1 + \beta \cdot I_{i,\text{total}}}$$
 - 18: Apply the scaling factor to each user $k \in \mathcal{C}_i$:
$$p_k \leftarrow \alpha_i \cdot p_k$$
 - 19: **end for**
 - 20: **Step 4: Finalization**
 - 21: Output the final optimized power allocations p_k for all users
-

Where ϵ , N_{clusters} , T_{\max} , and β are decided by system design. Finally, we apply (3.42) through (3.44) for resource allocation.

Few words about Complexity

Clustering reduces the complexity of the power allocation problem in mMIMO systems by partitioning the original large-scale optimization problem into smaller, more manageable sub-problems. Without clustering, the complexity typically grows exponentially or polynomially with the number of users K , as the system must optimize power allocation across all users simultaneously. By applying K-means clustering, the problem is divided into N clusters, each with a smaller number of users K_i , leading to a reduced overall complexity of $N \times \mathcal{O}(f(K/N))$, where $f(K)$ represents the original complexity function. If $f(K)$ is linear (1st order), the overall computational complexity does not theoretically decrease with clustering. However, in most practical power allocation problems, $f(K)$ is a higher order due to interference terms and SINR constraints, making the reduction in complexity with clustering significant. Although inter-cluster coordination adds some complexity, it is generally outweighed by the reduction achieved through clustering, resulting in a more scalable and computationally feasible solution.

Iterative Power Scaling Bisection Algorithm (IPSBA)

The power allocation algorithm shows high complexity. For this purpose, we introduce a less strict approach to reduce the complexity, as demonstrated by the following problem:

$$P3.7 : \min \alpha \tag{3.65a}$$

s.t.

$$\gamma \left(\sum_j p_j |\mathbf{w}_k^H \mathbf{h}_j|^2 + \sigma^2 \|\mathbf{w}_k\|^2 \right) \leq p_k |\mathbf{w}_k^H \mathbf{h}_k|^2 \tag{3.65b}$$

$$0 \leq p_k \leq \alpha P^{\max} \tag{3.65c}$$

In $P3.7$, we have solely considered the SINR within the log, considering an increase in this term equivalent to an increase in the log itself. The challenge in the new problem is to find the best γ and then solve the problem. For this purpose, we implement the well-known bisection algorithm to find an acceptable value for γ without compromising energy consumption. This can be controlled by carefully selecting the range for γ , then applying bisection until finding the best value. The IPSBA algorithm illustrates the steps to solve $P3.7$.

Algorithm 7 Iterative Power Scaling Bisection Algorithm (IPSBA)

Input:

- Maximum uplink power P^{\max}
- Precision level ϵ

Output:

- Optimal SINR threshold γ_{opt}
- Scaling factor α_{opt}
- Power allocations $\{p_k^{\text{opt}}\}$

1: **Step 1: Initialization**2: Set $\gamma_{\text{lower}} \leftarrow \gamma_{\text{min}}$ 3: Set $\gamma_{\text{upper}} \leftarrow \min_k \left(\frac{P^{\max} |\mathbf{w}_k^H \mathbf{h}_k|^2}{\sigma^2 \|\mathbf{w}_k\|^2} \right)$ 4: Initialize $\rho_{jk}^{\text{opt}} \leftarrow 0$ for all k 5: **Step 2: Bisection Iteration**6: **while** $\gamma_{\text{upper}} - \gamma_{\text{lower}} > \epsilon$ **do**

7: Compute midpoint:

$$\gamma_{\text{mid}} = \frac{\gamma_{\text{lower}} + \gamma_{\text{upper}}}{2}$$

8: Solve problem *P3.7* using γ_{mid} 9: **if** feasible solution found **then**10: Set $\gamma_{\text{lower}} \leftarrow \gamma_{\text{mid}}$ 11: Update p_k^{opt} and α_{opt} 12: **else**13: Set $\gamma_{\text{upper}} \leftarrow \gamma_{\text{mid}}$ 14: **end if**15: **end while**16: **Step 3: Finalization**17: **Return:** Final power allocations $\{p_k^{\text{opt}}\}$, SINR threshold γ_{opt} , and scaling factor α_{opt}

Finally, after power allocation, we apply (3.42) through (3.44) to allocate resources.

3.7 Results and Discussion

In this section, we will study the algorithms derived so far. We will divide the numerical results into two parts: NOMA and mMIMO.

NOMA algorithm

We implement the parameters in Table 1. We begin by demonstrating the efficiency of the modified Gale-Shapely algorithm for channel assignment.

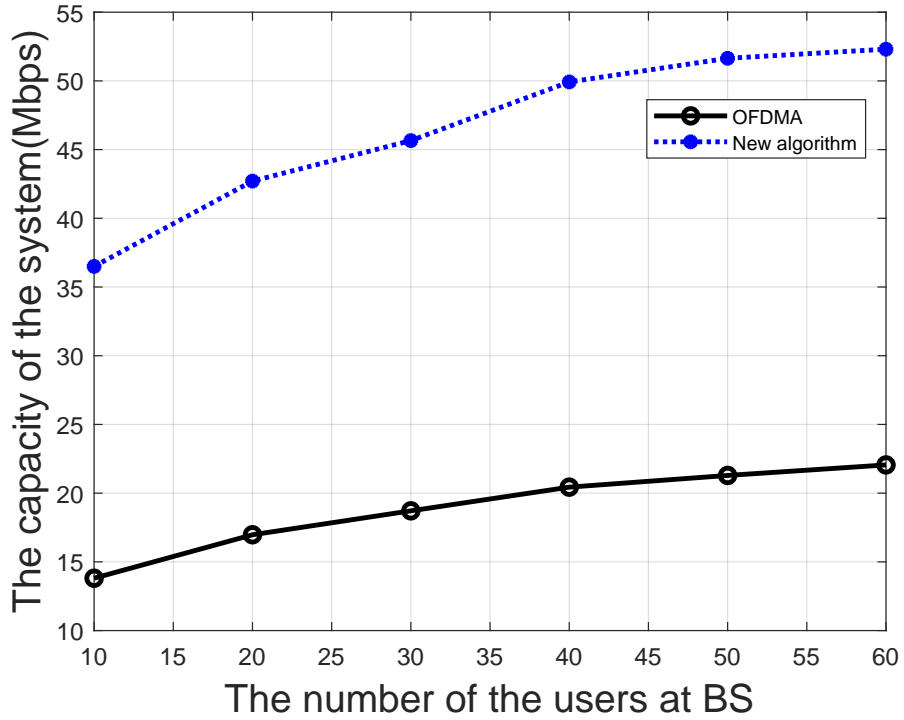
The simulation parameters for the NOMA network are summarized in Table 3.1.

As shown in Fig. 3.1, implementing the NOMA algorithm has significantly increased

Table 3.1: Simulation Parameters

Parameter	Value
Cell radius	500 m
Minimum distance between BS and UEs	50 m
Minimum distance between UEs	40 m
System bandwidth (BW)	5 MHz
Maximum number of UEs	60
Noise power spectral density	-174 dBm/Hz
Effective switched capacitance (τ)	10^{-8} Farad
Processing unit capacity	1 G cycles/s
UE data size	80–100 Kbits
UE peak power (P_k)	23 dBm

system capacity. For this purpose, we have fixed each mobile output power to 20 dBm for both OFDMA and NOMA users. Furthermore, we multiply the data rate expressed in (3.6) by each sub-channel BW to get the results shown in

**Figure 3.1:** Capacity of the system versus different numbers of users.

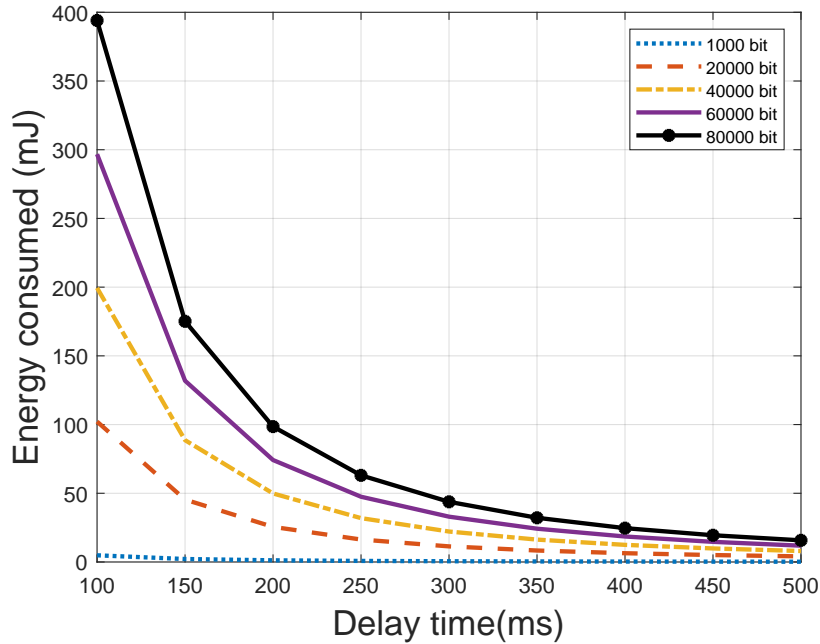


Figure 3.2: Local energy consumption for different delay times.

As shown in Fig. 3.2, the delay is inversely related to the energy consumed. We can justify this by remembering that a longer delay necessitates a slower task speed, affecting the energy consumed. Fig. 3.2 also shows the impact of data size on energy consumption, which, as expected, demonstrates a proportional relationship between the two.

We also studied the overall energy consumption of different numbers of users. While each user is solely focused on their power consumption, we have comprehensively understood the overall energy impact across multiple users.

As shown in Fig. 3.3, the energy is computed according to (3.31), we notice that it increases with different numbers of users, and the decrease in delay requirement causes a significant increase in energy required.

In Fig. 3.4, we depict a scenario where a user's local processing consumes energy, and we compare this with the user's energy consumption under both high and low SINR conditions. The scenario choice is dependent on channel conditions and delay time.

Furthermore, we see the impact of increasing processing units in Fig. 3.5.

Increasing the PCs will relax the delay requirement on the user's equipment, which requires a slower speed and reduces energy consumption.

massive MIMO

In mMIMO simulation part, we implement the parameters detailed in Table. 3.2. We continue to use a single-cell scenario. We modify BW to 20 MHz, and we will consider Imperfect Channel State Information (CSI). Moreover, we estimate the channels using the Minimum Mean Square Error (MMSE) estimation method [14]. For each mobile,

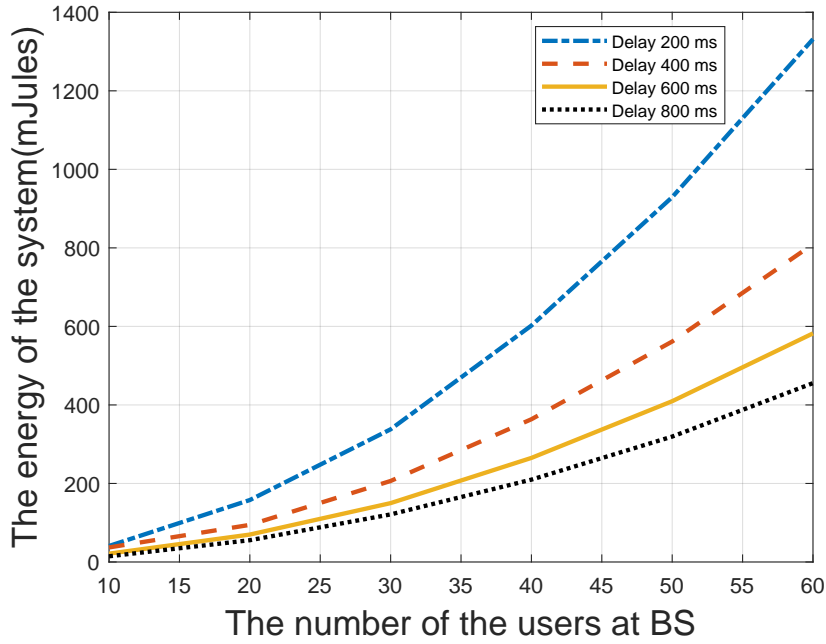


Figure 3.3: Energy consumed for remote processing vs. number of users.

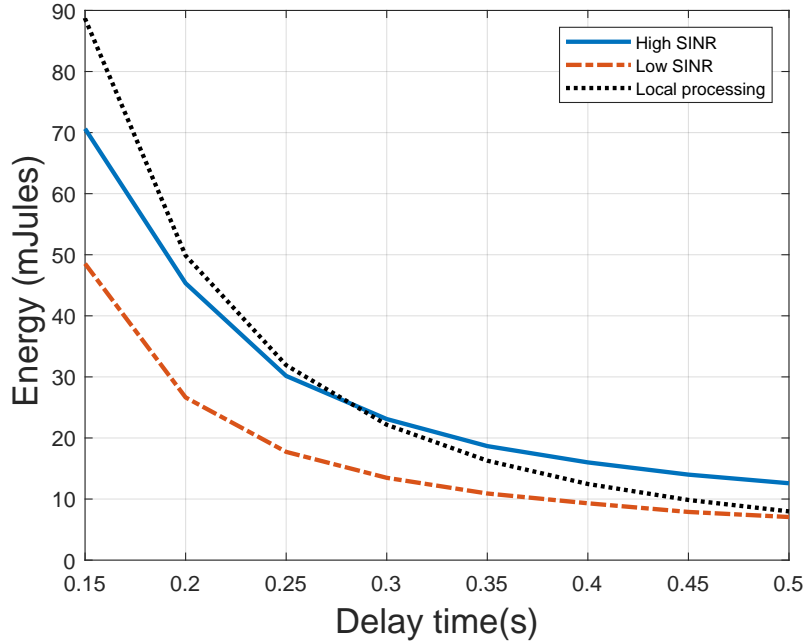


Figure 3.4: Energy consumed for local and remote processing.

we consider random AoA with respect to the BS, and the unresolved angle distributions around the AoA ($f(\phi)$) are considered Gaussian, which is justified because of the assumed large number of multipath reflected rays (the scenario assumes that the user is surrounded with many reflectors while the BS is elevated such that no reflectors exist in the nearby). The results shown in Fig. 3.6 show that the CCCP and PSSB algorithms introduce very

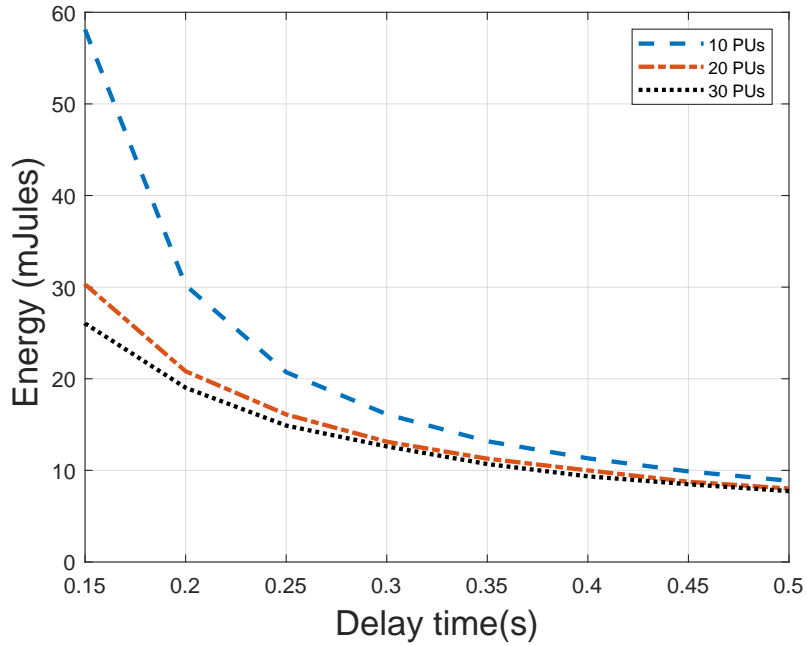


Figure 3.5: Impact of increasing PCs ($W_k = 1e9$).

close results in terms of spectral efficiency (SE), with the former giving about 10% higher SE, while the equal power distributions benefit from channel gain and give high SE when channel conditions are improving for users.

In Fig. 3.7, we show the impact of increasing antenna number on total SE in the cell. The figure shows that SE increases proportionally with antenna numbers, as evidenced by the equal power distributions. However, we observe that the rate does not significantly increase due to the application of optimisation algorithms. This is because we have set the minimum rate, which decreases energy consumption as the number of antennas increases. Finally, in Fig. 3.8, we show how the delay requirements impact energy consumption in the mMIMO scenario. As expected, increasing the delay has a positive effect on energy consumption.

Table 3.2: Simulation Parameters

Parameter	Value
Cell radius	500 m
Minimum distance between BS and UEs	50 m
System bandwidth (BW)	20 MHz
Maximum number of UEs	60
Number of Antenna	100
Path-loss exponent	3.76
Shadow fading (standard deviation)	10
Bandwidth	20 MHz
Transmit power (mW)	100

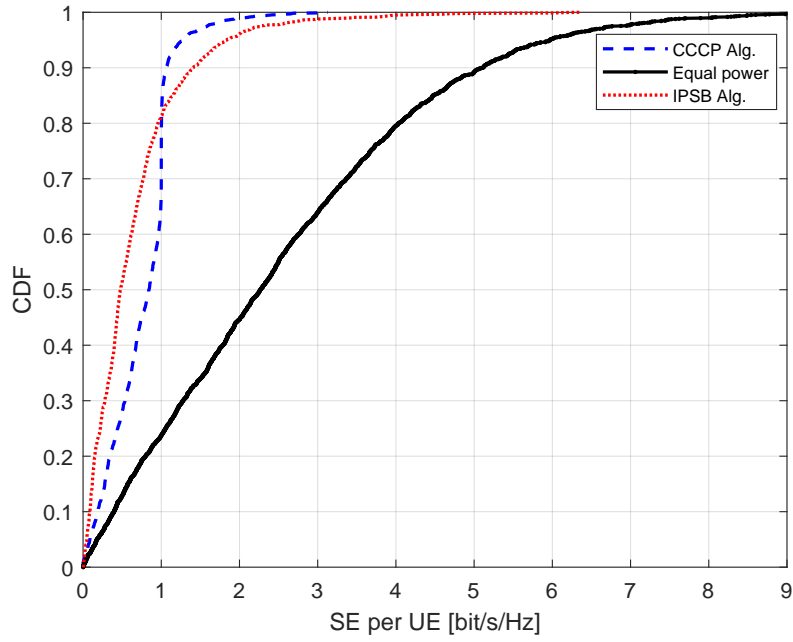


Figure 3.6: CDF of Spectral Efficiency.

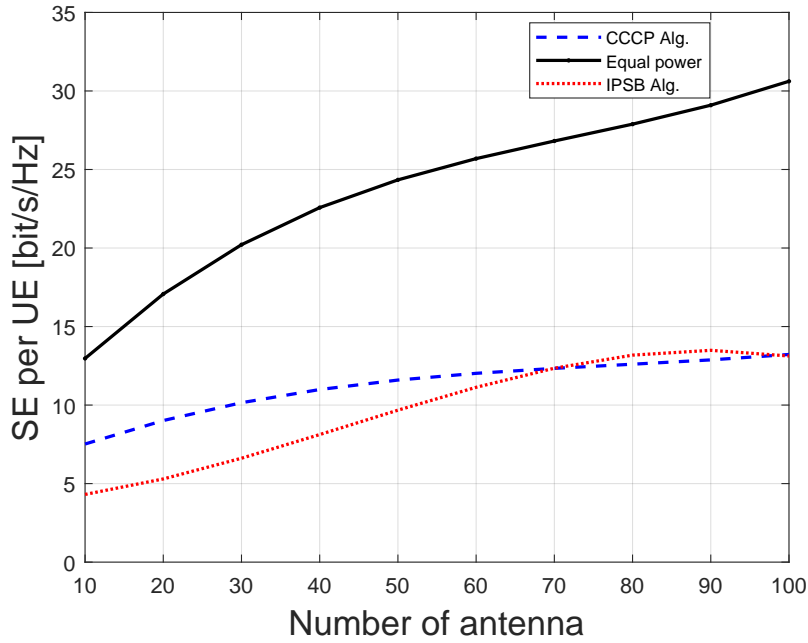


Figure 3.7: Spectral Efficiency Vs Number of Antenna.

Overall, the numerical results confirm the effectiveness of the proposed methods in both NOMA and mMIMO scenarios. In the NOMA case, the modified Gale-Shapely-based channel assignment improves system capacity compared with OFDMA, while the proposed local and remote processing models show that energy consumption strongly depends on delay constraints, user number, and available processing units. In the mMIMO case, the

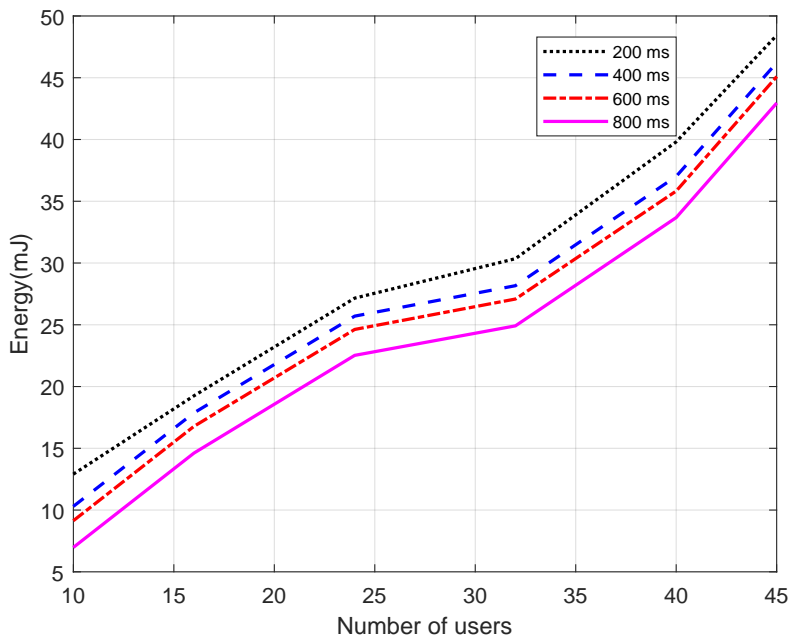


Figure 3.8: Energy consumed when applying remote processing for different delay times.

CCCP and IPSB-based power allocation methods provide better spectral-efficiency and energy-consumption trade-offs than equal power allocation, with CCCP achieving the best overall performance. These results show that the proposed framework improves resource utilization, enhances capacity, and reduces energy consumption under different wireless system settings.

3.8 Conclusion

This work developed a comprehensive energy-efficient resource allocation framework for mobile edge computing by jointly addressing local task execution and remote data offloading. For local processing, both deterministic and probabilistic energy-consumption models were derived, enabling efficient task execution under known and uncertain workload conditions. This provided a more realistic characterization of mobile-device energy usage and showed how delay tolerance can be exploited to reduce local processing cost.

For remote processing, two important wireless access scenarios were investigated. In the NOMA case, a stable matching-based sub-channel assignment scheme and a convexified uplink power allocation method were proposed to efficiently support offloading under cloudlet and delay constraints. In the massive MIMO case, I formulated an uplink power-control problem and developed efficient power allocation solutions, together with clustering support to improve scalability in larger systems.

The numerical results demonstrated that the proposed methods improve energy efficiency, enhance resource utilization, and maintain the required service constraints in both NOMA and massive MIMO settings. Overall, the work shows that integrating optimized local execution with communication-aware offloading decisions provides an effective and practical approach for energy-efficient MEC systems.

Chapter 4

Graph Coloring and User Clustering-Based Resource Allocation for Device-to-Device Communication in 5G Networks

4.1 Introduction

Device-to-Device (D2D) communication is a key enabler in 5G and emerging 6G networks, addressing challenges such as efficient spectrum utilization, low latency, and energy optimization. By enabling direct communication between devices, D2D reduces dependence on centralized infrastructure and supports applications like smart cities, industrial IoT, autonomous vehicles, and remote healthcare systems [3, 83, 108].

Incorporating artificial intelligence (AI) and particularly machine learning (ML) has expanded the potential of D2D communication. AI-based frameworks have demonstrated significant improvements in optimizing Quality of Service (QoS) in dense and dynamic network environments [108]. Future 6G systems must tackle additional challenges, such as ultra-reliable low-latency communication (URLLC) and connecting billions of devices globally, as outlined in [62, 83].

Resource allocation and interference management remain significant challenges in spectrum-scarce environments where D2D users coexist with cellular users (CUs). Researchers have explored numerous solutions to address these issues. For example, Al-Imari et al. [3] introduced uplink non-orthogonal multiple access (NOMA) to enhance resource sharing, while Zhou et al. [128] proposed graph-coloring techniques for efficient interference management. Additionally, Zhao and Wang [127] developed optimization methods for resource allocation in D2D underlying cellular networks, and Lee et al. [49] applied graph-coloring algorithms to tackle interference problems in resource allocation. Power allocation for NOMA-based D2D systems has also been investigated, demonstrating that it is possible to cut down on interference while maintaining stability [5, 75].

Congestion control has also emerged as a crucial aspect of D2D communications. Malini and Karthigaikumar [62] introduced a weighted AIMD congestion control algorithm that optimizes resource utilization and mitigates congestion in 5G networks. Furthermore, multi-hop D2D resource allocation strategies, such as those outlined by Chauhan et al. [19], have addressed the growing complexity of network management.

This chapter introduces a novel two-phase heuristic framework to address resource allocation challenges in mixed CU and D2D networks. First, K-means clustering groups the users into clusters, simplifying interference management. Second, a graph-coloring algo-

rithm assigns subchannels, ensuring minimal interference. Additionally, a power allocation optimization framework ensures fairness while meeting stringent QoS requirements.

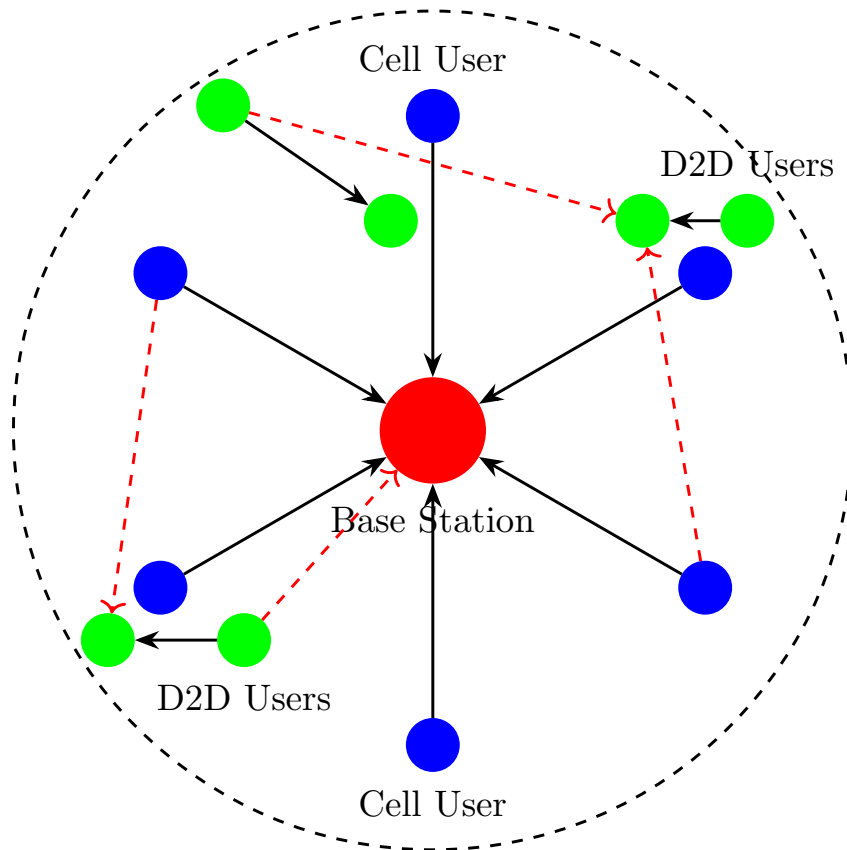


Figure 4.1: A cellular network consists of a single BS, 6 CUs and 6 D2D users. Solid lines are communication channels, and dashed lines are interference channels.

4.2 System Model

This model considers – for simplicity, without loss of generality – a single BS that manages the network within the cell. It is responsible for channel assignment to cellular users and management of D2D communications. Furthermore, consider the cellular users as $u_j \in \mathcal{U}_{\text{CU}} \subseteq \mathcal{K}$, where $\mathcal{U}_{\text{CU}} = \{1, 2, \dots, N_{\text{CU}}\}$. Also, D2D users are the set of users that communicate directly with each other using the uplink channels of the cellular network and are denoted by $u_j \in \mathcal{U}_{\text{D2D}} \subseteq \mathcal{K}$, where $\mathcal{U}_{\text{D2D}} = \{1, 2, \dots, N_{\text{D2D}}\}$. The BS manages D2D communications under its protocol framework, without passing through it. The channels provided by the BS are denoted by $c_n \in \mathcal{N}$ where $\mathcal{N} = \{1, 2, \dots, N\}$ and N is the total number of available channels. The reason for choosing the uplink channels for D2D is that power transmitted by CUs is considered much less than that transmitted from BS and thus causes less interference. Because of the scarcity of the spectrum, D2D users are supposed to reuse channels assigned to CUs. The channels assigned to CUs are usually based on the BS’s convenience, depending on several factors that are out of the scope of this work. Based on the above information, there are three main ways that interference in this network arises:

- Interference from D2D users to the BS (D2D-BS) and vice versa: This occurs when the D2D users' signals on the uplink channels create interference that affects the signals from CUs to the BS. The interference can deteriorate the quality of the uplink communications that the BS receives.
- Interference from Cellular Users to D2D Devices (CU-D2D) and vice versa: If CUs use the same or adjacent channels while transmitting on their assigned channels, D2D transmissions may encounter interference. This kind of interference impacts the reception quality at the D2D receivers.
- Interference among D2D Devices (D2D-D2D): Since multiple D2D users may share the same uplink channels or use closely spaced frequencies, their transmissions can interfere. This scenario requires careful management to mitigate the impact on communication quality.

Formulation of Channel Assignment Problem

Before defining the main objective function and constraints, we need to define the following variables:

- $x_{j,n}$: Binary variable where $x_{j,n} = 1$ if CU $u_j \in \mathcal{U}_{\text{CU}}$ is assigned channel $c_n \in \mathcal{N}$, otherwise 0.
- $y_{j,n}$: Binary variable where $y_{j,n} = 1$ if D2D pair $u_j \in \mathcal{U}_{\text{D2D}}$ uses channel $c_n \in \mathcal{N}$, otherwise 0.
- z_j : Binary variable that is 1 if D2D pair $u_j \in \mathcal{U}_{\text{D2D}}$ is active (transmitting).

Then the optimization problem is formulated as:

$$P4.1 : \min_{x,y,z} \sum_{j \in \mathcal{U}_{\text{CU}}} \sum_{k \in \mathcal{U}_{\text{D2D}}} \sum_{n \in \mathcal{N}} \alpha_{j,k,n} x_{j,n} y_{k,n} \quad (4.1a)$$

$$+ \sum_{j \in \mathcal{U}_{\text{D2D}}} \gamma_j z_j + \sum_{\substack{j \neq l \\ n \in \mathcal{N}}} \beta_{j,l,n} y_{j,n} y_{l,n} \quad (4.1b)$$

$$\text{s.t.} \quad \sum_{n \in \mathcal{N}} x_{j,n} = 1, \quad \forall u_j \in \mathcal{U}_{\text{CU}}, \quad (4.1c)$$

$$\sum_{n \in \mathcal{N}} y_{j,n} = 1, \quad \forall u_j \in \mathcal{U}_{\text{D2D}}, \quad (4.1d)$$

$$z_j \geq y_{j,n}, \quad \forall u_j \in \mathcal{U}_{\text{D2D}}, \forall c_n \in \mathcal{N}, \quad (4.1e)$$

$$\sum_{j \in \mathcal{U}_{\text{CU}}} x_{j,n} + \sum_{j \in \mathcal{U}_{\text{D2D}}} y_{j,n} \leq 1 + M, \quad \forall c_n \in \mathcal{N}, \quad (4.1f)$$

$$x_{j,n}, y_{j,n}, z_j \in \{0, 1\}. \quad (4.1g)$$

where:

- $\alpha_{j,k,n}$ is the interference coefficient between a CU $u_j \in \mathcal{U}_{\text{CU}}$ and a D2D pair $u_k \in \mathcal{U}_{\text{D2D}}$ on channel c_n .
- γ_j represents the interference impact of D2D pair $u_j \in \mathcal{U}_{\text{D2D}}$ on the BS.

- $\beta_{j,l,n}$ is the interference coefficient between different D2D pairs $u_j, u_l \in \mathcal{U}_{\text{D2D}}$ on the same channel c_n .
- σ^2 denotes the noise power (AWGN), applied consistently across all channels.

4.3 User Clustering

Problem $P4.1$ is a quadratic binary optimization problem, since its objective function contains products of binary decision variables, such as $x_{j,n}y_{k,n}$ and $y_{j,n}y_{l,n}$. Due to its combinatorial complexity, finding the exact optimal solution in large-scale networks with many CUs and D2D pairs is challenging. Therefore, we adopt a heuristic two-phase solution: first, users are clustered according to their proximity to the BS, and then graph coloring is applied for channel assignment.

We first apply K-means clustering. By applying this, we divide the users $u_j \in \mathcal{K}$ (where \mathcal{K} includes both CUs and D2Ds) into k groups, or clusters, using the k-means algorithm. Each cluster is represented by a centroid c_i , and each user is assigned to the closest centroid, ensuring that the users within each group have a closer relationship than those in other groups.

The k-means clustering aims to minimize the variance within each cluster, expressed as the following objective function:

$$J(C) = \sum_{i=1}^k \sum_{u_j \in \mathcal{L}_i} \|u_j - c_i\|^2$$

where:

- \mathcal{L}_i represents the set of all users assigned to the i -th cluster.
- $\|u_j - c_i\|^2$ is the squared Euclidean distance between user u_j and the centroid c_i of its cluster.

Process of K-Means Clustering

To achieve the objective, k-means follows these steps iteratively:

1. **Initialization:** Select k initial centroids randomly.
2. **Assignment:** Assign each data point u_j to the nearest centroid c_i , such that:

$$\mathcal{L}_i = \{u_j : \|u_j - c_i\| \leq \|u_j - c_l\| \text{ for all } l \neq i\}$$

3. **Update:** Recalculate each centroid c_i as the mean of all data points assigned to \mathcal{L}_i :

$$c_i = \frac{1}{|\mathcal{L}_i|} \sum_{u_j \in \mathcal{L}_i} u_j$$

4. **Repeat:** Continue the assignment and update steps until the centroids no longer change significantly, indicating convergence.

We find this algorithm efficient in our scenario and believe it will provide fast convergence for two reasons: Firstly, while the number of users is significant, it is still significantly smaller than the extensive data sets that the k-means algorithm applies to. Second, we will only need a few iterations to achieve a satisfactory clustering; after a few iterations, we will rely on graph coloring to determine the optimal channel assignments.

4.4 Graph Coloring

Graph colouring, a fundamental concept in graph theory, assigns colours to graph elements under specific constraints. Our work uses graph colouring to minimize interference between cellular and D2D users. We assign each colour to a unique subchannel $c_n \in \mathcal{N}$ to prevent interference between adjacent users.

After clustering, we use a $G(V, E)$ graph to illustrate a cellular network structure that permits CU and D2D communications. V and E are the groups of users $u_j \in \mathcal{K}$ (including both D2D and CUs) and interference between them, respectively. This graph serves as the basis for developing the auxiliary conflict graph $\hat{G}(\hat{V}, \hat{E})$, defined as a weighted undirected graph where each vertex $\hat{v}_j \in \hat{V}$ represents a user. An edge $\hat{e} \in \hat{E}$ connects vertices \hat{v}_j and \hat{v}_m if and only if they are assigned the same subchannel $c_n \in \mathcal{N}$, i.e., $x_{j,n} = x_{m,n} = 1$.

Definition 1: The conflict graph $\hat{G}(\hat{V}, \hat{E})$ is a weighted undirected graph whose vertex set \hat{V} represents the set of active users. Any two vertices $\hat{v}_j, \hat{v}_m \in \hat{V}$ (which represent users u_j and u_m) are connected by an edge $\hat{e} \in \hat{E}$ if their chosen subchannels overlap. Furthermore, we can define a proper spectrum allocation for the conflict graph: let w_j denote the weight (priority) of user u_j . If \hat{v}_j and \hat{v}_m are adjacent in \hat{G} , then the distance between w_j and w_m should ensure minimal interference. This allocation forms a proper subchannel assignment for the conflict graph.

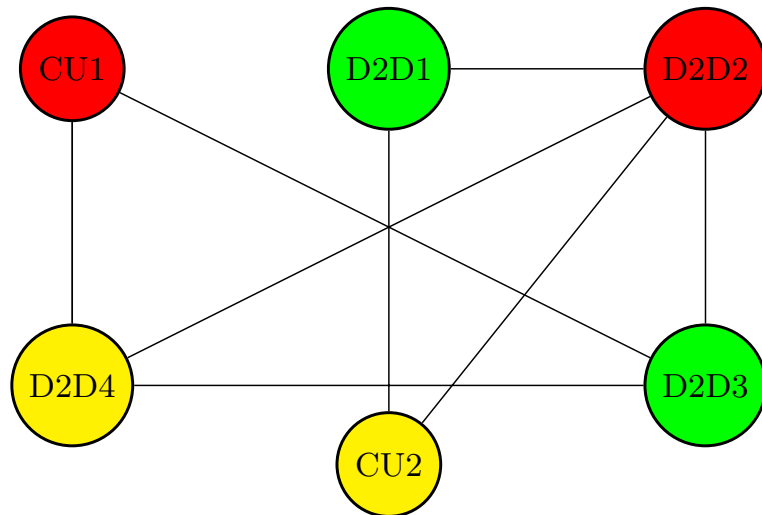


Figure 4.2: Conflict graph with nodes colored to avoid interference

After applying the graph coloring algorithm, Fig. 4.2 illustrates a six-node conflict graph, each corresponding to different users assigned to their respective subchannels. According to the established definition, a proper subchannel assignment in this conflict graph ensures minimal interference among these users, mirroring a proper subchannel configuration for the respective subchannels. This means that the best subchannel assignment is directly

related to the chromatic number $\chi(\hat{G})$, which gives the fewest colours required to avoid interference.

Figure 4.2 presents an illustrative example of the conflict graph used to explain the graph coloring process. The selected nodes (CU1, CU2, D2D1–D2D4) are representative users chosen for visualization purposes only and do not correspond to a specific simulation snapshot. The figure demonstrates how interference relationships are modeled and how graph coloring ensures that users sharing strong interference links are assigned different subchannels.

4.5 Power Allocation

After channel assignment, each user $u_j \in \mathcal{K}$ will suffer interference, which depends on the number of users in the geographic area and the number of assigned channels. We can express the data rate as follows:

$$R_j = B \cdot \log_2 \left(1 + \frac{P_j \cdot g_{jj}}{\sum_{m \neq j} P_m \cdot g_{mj} + \sigma^2} \right) \quad (4.2)$$

Where R_j is the data rate for user u_j , B is the channel bandwidth, P_j is the transmit power of user u_j , g_{jj} is the channel gain (or path loss) between user u_j and its receiver, P_m is the transmit power of interfering user u_m , g_{mj} is the channel gain from interfering user u_m to user u_j , σ^2 is the noise power.

$$P4.2 : \max_{\mathbf{P}} \sum_{j=1}^{|\mathcal{K}|} R_j \quad (4.3)$$

subject to:

$$R_j = B \cdot \log_2 \left(1 + \frac{P_j \cdot g_{jj}}{\sum_{m \neq j} P_m \cdot g_{mj} + \sigma^2} \right), \quad (4.4)$$

$$= B \cdot \log_2 (1 + \text{SINR}) \quad (4.5)$$

$$0 \leq P_j \leq P_{\max}, \quad \forall u_j \in \mathcal{K} \quad (4.6)$$

Where $\mathbf{P} = [P_1, P_2, \dots, P_{|\mathcal{K}|}]$ is the power vector for all users. Problem *P4.2* is nonconvex due to the interference term. Our goal is to modify the problem to a more convenient form. We focused our attention on the term present within the log function. We can achieve our ultimate goal by increasing the SINR term, which also increases the long term. To increase the SINR and to guarantee fairness, we convert the problem into a multiplying product of SINR. We then formulate the new problem as follows:

$$\max_{\mathbf{P}} \prod_{j=1}^{|\mathcal{K}|} \text{SINR}_j \quad (4.7)$$

subject to:

$$\text{SINR}_j = \frac{P_j \cdot g_{jj}}{\sum_{m \neq j} P_m \cdot g_{mj} + \sigma^2}, \quad \forall u_j \in \mathcal{K} \quad (4.8)$$

$$0 \leq P_j \leq P_{\max}, \quad \forall u_j \in \mathcal{K} \quad (4.9)$$

Next, if we impose a minimum rate (QoS) for each user, say Γ_j , then the problem writes:

$$P4.3 : \max_{\mathbf{P}} \prod_{j=1}^{|\mathcal{K}|} \Gamma_j \quad (4.10)$$

subject to:

$$\sum_{m \neq j} \frac{\Gamma_j (P_m \cdot g_{mj} + \sigma^2)}{P_j \cdot g_{jj}} \leq 1, \quad \forall u_j \in \mathcal{K} \quad (4.11)$$

$$0 \leq P_j \leq P_{\max}, \quad \forall u_j \in \mathcal{K} \quad (4.12)$$

We notice that $P4.3$ is posynomial and thus a geometric program, i.e., an optimization problem with posynomial functions that can be transformed into a convex form and solved efficiently.

4.6 Simulation

This section assumes a single-cell scenario with 40 D2D users and 10 cellular users (CUs). The cell has a coverage diameter of 1 km. We utilize both K-means clustering and graph coloring algorithms for user grouping and channel assignment. Fig. 4.3 illustrates the

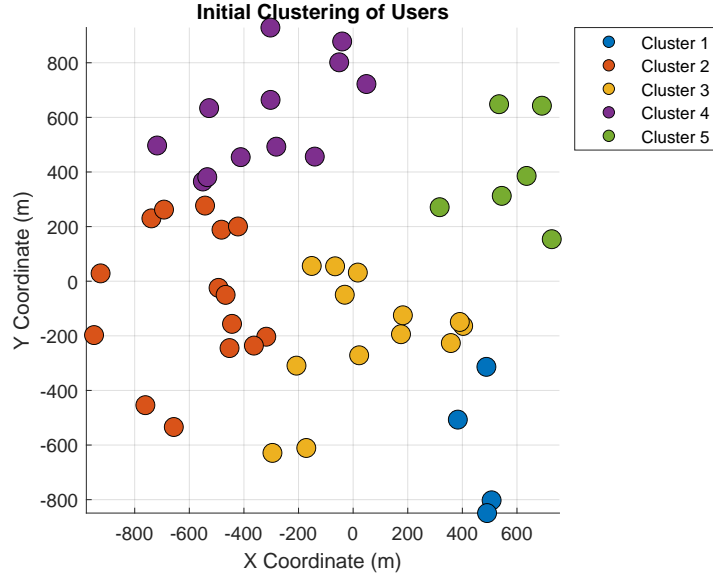


Figure 4.3: Initial user clustering based on proximity to the BS using K-means.

clustering of users based on their spatial proximity to the base station (BS), which serves as the reference point. Using the K-means algorithm, users are grouped into clusters such that those located closer to each other are assigned to the same cluster, minimizing the intra-cluster distance. This grouping simplifies interference management by limiting strong interference interactions within smaller regions, and prepares the network for efficient channel assignment in the subsequent graph coloring phase.

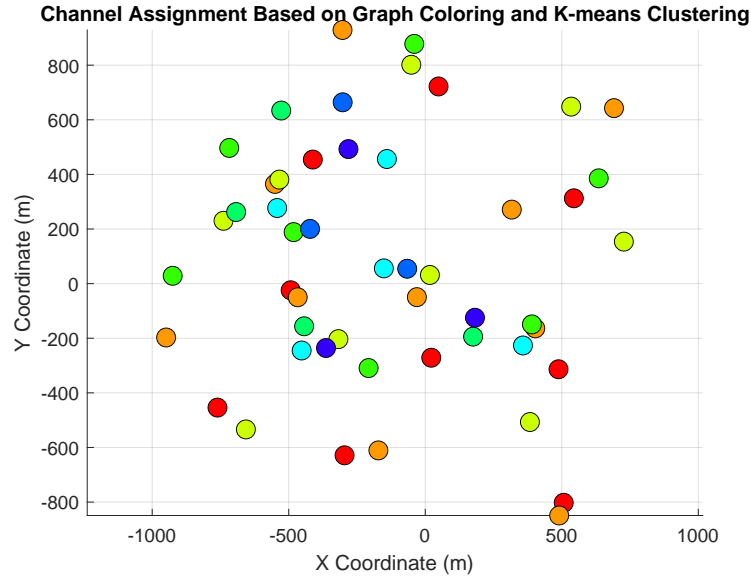


Figure 4.4: Channel distribution after applying graph coloring.

After clustering, graph coloring is applied to assign channels, as illustrated in Fig. 4.4. In this figure, each color represents a distinct channel, and users connected through strong interference relationships are assigned different colors to reduce channel conflicts. This allocation strategy helps separate interfering users while allowing efficient channel reuse among users with low mutual interference, thereby improving overall spectrum utilization.

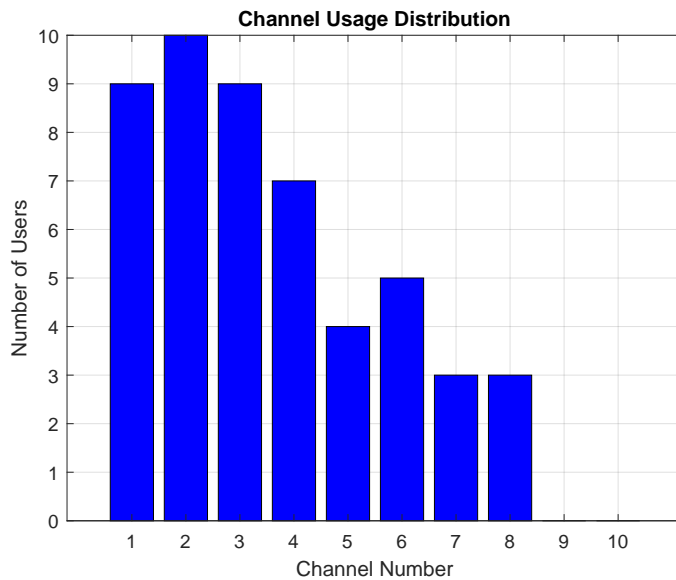


Figure 4.5: Users per channel after allocation.

Fig. 4.5 presents the number of users assigned to each channel after the allocation process. The bar chart provides insight into the channel utilization pattern and shows how the proposed scheme distributes users across the available channels. A relatively balanced distribution indicates that the allocation is fair and avoids overloading specific channels, which helps improve spectrum efficiency and reduce excessive interference.

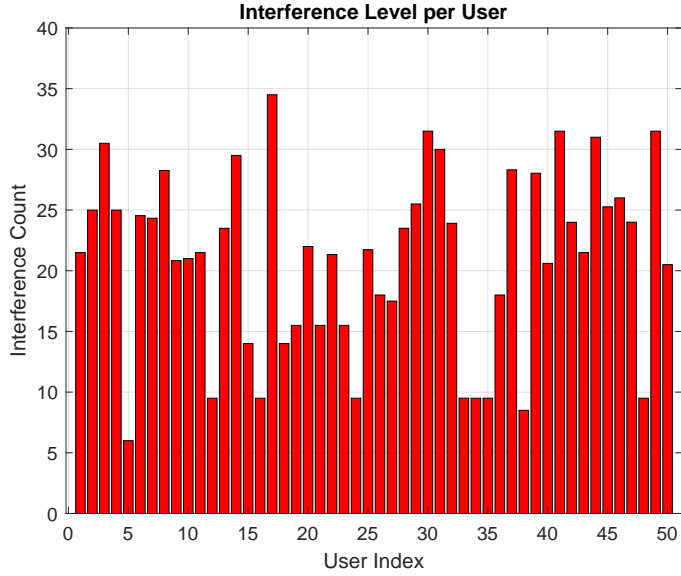


Figure 4.6: Interference level experienced by each user.

The effectiveness of the channel assignment is evaluated in Fig. 4.6, where the interference level for each user is quantified by the number of neighboring users reusing the same channel. This figure reflects how well the graph-coloring-based allocation mitigates harmful channel conflicts. Lower interference levels for most users indicate that the proposed assignment successfully separates strongly interfering users, thereby improving communication reliability and overall spectrum reuse efficiency.

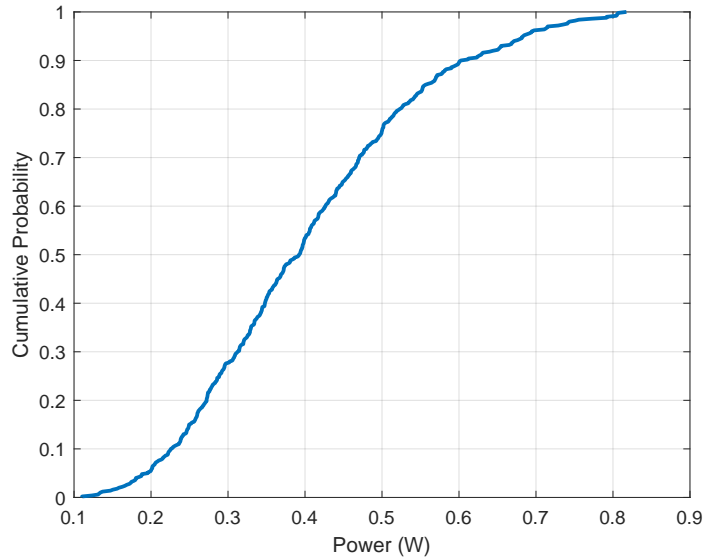


Figure 4.7: CDF of power allocation with fairness consideration.

Fig. 4.7 presents the cumulative distribution function (CDF) of the allocated transmit power among D2D users after applying the proposed power control scheme. It can be observed that approximately 90% of users are assigned less than 0.6 W, which indicates that the required quality-of-service constraints are satisfied without excessively increasing the transmit power. This behavior reflects a fair and efficient power allocation strategy, where

users are provided with only the power needed to meet their minimum rate requirements while avoiding unnecessary interference.

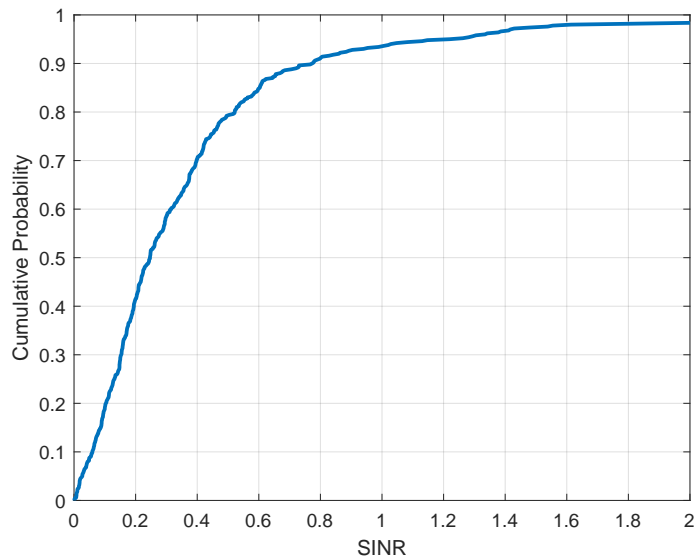


Figure 4.8: CDF of SINR achieved per user.

Finally, Fig. 4.8 shows the distribution of the achieved SINR values among users. The steepness of the CDF indicates that the SINR values are concentrated within a relatively narrow range, which reflects the fairness of the proposed power allocation approach. This result confirms that the algorithm effectively controls interference and allocates power in a way that allows most users to achieve the target communication quality with balanced performance across the network.

4.7 Conclusion

This chapter addressed the problem of resource allocation in D2D-enabled cellular networks under channel reuse conditions. The channel assignment problem was formulated as a binary quadratic optimization problem, which is difficult to solve directly in large-scale scenarios. To provide a practical solution, a two-phase heuristic framework based on K-means clustering and graph coloring was developed for user grouping and interference-aware channel assignment. In addition, a fairness-oriented power allocation scheme was introduced to improve user performance while satisfying QoS requirements. The simulation results confirmed that the proposed framework achieves efficient channel reuse, mitigates interference, and provides balanced power and SINR performance among users.

Chapter 5

Mobility-Aware Resource Allocation in D2D Communications Using Genetic Algorithms

5.1 Introduction

With the rapid evolution of wireless technology, it has become increasingly clear that we need more spectrum, faster data speeds, and less latency. The arrival of 5G and all the planning behind what is coming next have thrown these issues into sharp focus. The growing number of connected devices has made the spectrum increasingly congested. One approach that is gaining more attention is device-to-device, or D2D, communication, as it allows devices to talk directly without the base station involvement. According to [23], this method helps free up bandwidth, improves overall efficiency, and keeps the network from getting overwhelmed.

Although implementing device-to-device (D2D) communication in cellular networks is promising, it presents several challenges, particularly in managing power levels under highly dynamic conditions when users are in motion. To address these issues, researchers are increasingly considering (NOMA) as a practical solution to improve spectral efficiency while handling interference more intelligently through coordinated power and channel allocation [1, 5]. Nevertheless, maintaining stable communication in fast-changing real-world scenarios—where user mobility plays a major role—remains a significant challenge.

User mobility significantly influences resource allocation strategies, affecting link reliability, interference levels, and the overall quality of service (QoS). As users move throughout the network, variations in channel conditions and frequent handovers complicate traditional static allocation methods, which typically assume fixed user positions. This mismatch can result in inefficient spectrum usage and reduced system performance [85]. Consequently, mobility-aware approaches have become essential for sustaining connectivity and improving overall network fairness and efficiency [109].

Researchers have explored different strategies to tackle these issues. Centralized optimization methods might deliver good results, but in practice, they are often too computationally heavy to be used in real-time [35]. Game theory offers more flexibility, allowing for decentralized decision-making, but it does not always perform well in environments that change quickly [43]. Lately, machine learning — especially Reinforcement Learning (RL) has been getting attention for its ability to adapt to complex wireless scenarios [44]. However, RL can be data intensive and may take too long to settle into a practical solution when networks evolve rapidly [2].

To address these challenges, our work introduces a resource allocation framework that takes user mobility into account by integrating supervised learning with evolutionary optimization techniques. The approach anticipates user movement patterns and uses this insight to adapt allocation strategies in real time. In our work, we introduce an enhanced version of the Non-Dominated Sorting Genetic Algorithm II (NSGA-II), which manages power control and channel assignment. This joint optimization is aimed at achieving a practical trade-off between system efficiency and fairness in dynamic mobile scenarios.

The main contributions of this work are as follows:

- **Mobility-Aware Learning Framework:** A machine learning-based system is developed to predict mobility patterns and inform resource allocation decisions in D2D-enabled cellular networks.
- **Multi-Objective Optimization:** A tailored NSGA-II algorithm is used to address conflicting objectives such as throughput maximization, interference mitigation, and fairness in resource distribution.
- **Adaptive Spectrum Reuse:** The proposed method dynamically adjusts spectrum reuse policies in response to mobility patterns, enhancing spectral efficiency and minimizing interference.
- **Extensive Simulation Analysis:** Comprehensive simulations under varying mobility and network load conditions demonstrate the proposed method's superiority over conventional static and heuristic techniques.

5.2 Related Work

Device-to-Device (D2D) communication has emerged as a significant research focus in next-generation wireless systems, owing to its potential to improve spectral efficiency and reduce communication latency. A considerable body of literature has studied the development of secure and efficient mechanisms to support D2D operations within cellular frameworks [6]. For example, [38] proposed a distributed ML-based design with modularity, specifically for D2D implementation. Furthermore, Logeshwaran *et al.* [58] focused on the enhancement of energy efficiency in resource allocation for D2D in 5G wireless networks. Other contributions, such as [88], mitigated interference management challenges that affect the implementation of D2D communications in cellular networks.

Rajab *et al.* [76] investigated power allocation in NOMA-based D2D communication by implementing the Greedy Asynchronous Distributed Interference Avoidance Algorithm (GADIA). Resource allocation frameworks combining NOMA and massive multiple-input multiple-output (mMIMO) in mobile edge computing environments have also been proposed to enhance energy efficiency and scalability [4].

Despite significant progress, a large portion of existing research continues to depend on static resource allocation strategies, which are typically inadequate in environments characterized by high user mobility and dense network deployments. In order to address these limitations, more recent studies have addressed the use of machine learning techniques to enable mobility-aware resource management. For instance, Hou *et al.* [37] employed radial basis function neural networks to predict vehicle velocity in real-time, highlighting the practical advantages of mobility forecasting. Lin *et al.* [52] implemented Markov chain models with driving behavior recognition to develop adaptive energy management approaches. Similarly, Jia *et al.* [45] examined learning-based predictive optimization for

dynamic environments, and Gandhi *et al.* [31] provided a thorough survey of machine learning applications in mobility-aware wireless resource management. Xu *et al.* [104] developed deep learning-based prediction of time-varying channels in dynamic environments, providing useful insights into channel state information CSI acquisition for systems requiring adaptive power and resource control under mobility.

Federated Learning (FL) has earned notable attention as a privacy-preserving paradigm for decentralized model training across large-scale Internet of Things (IoT) and edge computing environments. However, conventional FL methods often face performance degradation due to data heterogeneity and variable computational resources among participating devices. To overcome these limitations, Ma *et al.* [61] and Zhang *et al.* [119] proposed Clustered Federated Learning (CFL) frameworks. Furthermore, Xu *et al.* [101] introduced optimized resource usage in non-uniform conditions. On the other hand, Jiang *et al.* [46] introduced a CFL scheme designed for personalized adaptation models. In parallel, privacy-preserving clustering mechanisms have been studied by [33, 81] to balance model performance with communication efficiency.

RL, mainly in its deep learning variants, has been implemented to improve the performance of FL systems operating under resource-constrained conditions. For instance, Zhao *et al.* [126] proposed a deep RL-based framework for joint resource allocation and scheduling in hierarchical FL systems deployed within NOMA-enabled industrial IoT networks. Pan *et al.* [72] introduced a dynamic RL approach that effectively addresses latency and energy limitations. Xu *et al.* [103] and Zhang *et al.* [122] applied deep RL to enhance energy efficiency and optimize user selection in FL. The implementation of deep RL with clustered FL, as illustrated by Zhang *et al.* [120], demonstrates improved performance in artificial intelligence (AI) systems.

Previous works have treated mobility, optimization, and fairness as separate challenges. Few studies offer an integrated approach that simultaneously addresses these aspects in D2D-enabled cellular networks. In contrast, the framework proposed in this work combines machine learning-based mobility prediction with a hybrid NSGA-II algorithm to jointly optimize power control, channel assignment, and dynamic spectrum reuse. This comprehensive design offers an adaptive and fairness-aware resource management strategy suitable for highly dynamic network environments, setting our work apart from prior efforts and contributing to the evolution of intelligent, scalable resource allocation in future wireless systems.

5.3 System Model

The modeling begins by defining the network setup, user types, and their communication modes. A zone-based spatial division is introduced to capture user location and mobility. This is followed by the wireless channel model and SINR formulations for both D2D and cellular users. Power constraints are then outlined, and finally, a Markov-based mobility model is presented to reflect user movement and its impact on interference and resource allocation.

We consider a single-cell wireless network with a radius of R_{cell} , consisting of N_{CU} cellular users and N_{D2D} device-to-device (D2D) pairs. The sets of cellular users and D2D pairs are denoted as:

$$\mathcal{U}_{\text{CU}} = \{1, 2, \dots, N_{\text{CU}}\}, \quad \mathcal{U}_{\text{D2D}} = \{1, 2, \dots, N_{\text{D2D}}\}. \quad (5.1)$$

Cellular users communicate directly with the Base Station (BS) using orthogonal channels, while D2D pairs opportunistically reuse cellular resources for direct communication. The network incorporates dynamic user mobility, channel variations, and interference constraints to ensure efficient resource allocation and mitigate spectrum congestion.

5.3.1 Zone-Based Spatial Division

To efficiently analyze mobility and channel conditions, the cell is divided into N_{zones} concentric zones of equal width, as shown in Fig. 5.1, enabling location-aware resource allocation strategies:

$$\Delta R = \frac{R_{\text{cell}}}{N_{\text{zones}}}, \quad R_z = z \cdot \Delta R, \quad z = 1, \dots, N_{\text{zones}}. \quad (5.2)$$

Each user at a distance r from the BS belongs to zone z if:

$$R_{z-1} \leq r < R_z, \quad \forall z \in \{1, \dots, N_{\text{zones}}\}, \quad \text{where } R_0 = 0. \quad (5.3)$$

Users are assumed to transition between zones based on probabilistic mobility patterns. The adoption of a zone-based spatial segmentation facilitates an efficient and scalable means of incorporating user location into both mobility modeling and resource allocation. By categorizing users according to their radial distance from the base station, the system simplifies the prediction of mobility dynamics using zone-level transitions, enables more accurate estimation of interference patterns, and allows for proactive adjustments to power control and channel assignment. This approach offers a practical trade-off between spatial modeling granularity and computational complexity, making it particularly effective for real-time optimization in dynamic wireless environments.

5.3.2 Channel Model

The wireless channel is modeled as a combination of large-scale path loss and small-scale Rayleigh fading. The composite channel gain between a transmitter i and receiver k on channel j at time t is given by:

$$g_{i,k}^{(j)}(t) = h_{i,k}^{(j)}(t) \cdot \frac{\beta}{d_{i,k}(t)^\alpha}, \quad h_{i,k}^{(j)}(t) \sim \mathcal{CN}(0, 1), \quad (5.4)$$

where:

- β : Path loss constant.
- α : Path loss exponent.
- $d_{i,k}(t)$: Euclidean distance.
- $h_{i,k}^{(j)}(t)$: Rayleigh fading coefficient.

The set of available orthogonal channels is:

$$\mathcal{N} = \{1, 2, \dots, N_{\text{ch}}\}, \quad (5.5)$$

where N_{ch} is the number of available channels.

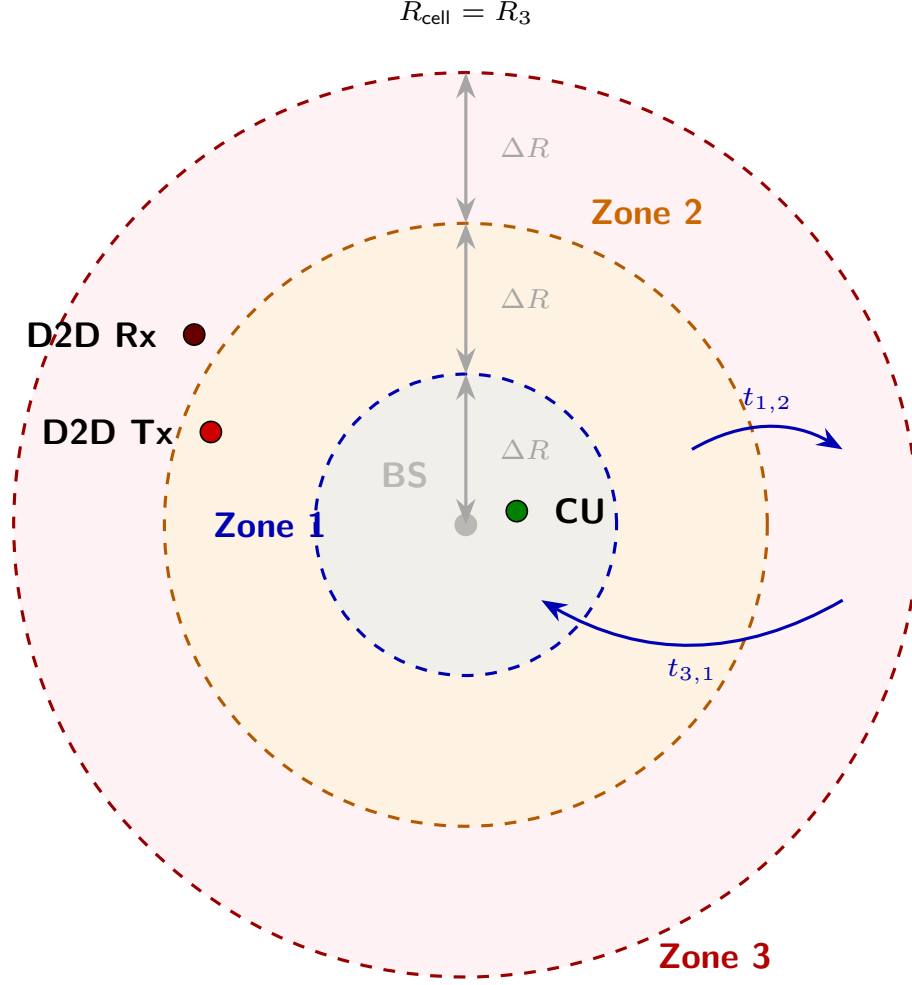


Figure 5.1: Enhanced concentric zone architecture in a cellular network. The base station (BS) is located at the center of the cell, surrounded by three radial zones. Examples of users include a cellular user (CU), a device-to-device (D2D) transmitter (Tx), and a receiver (Rx). Transition probabilities between zones are shown with arrows.

We analyze SINR separately for D2D and cellular uplink communications. The SINR expressions for the considered communication modes are summarized as follows:

(a) **D2D Communication**

D2D receivers experience both intra-tier and cross-tier interference. The SINR is:

$$\gamma_{i,k}^{(j)}(t) = \frac{P_i^{(j)}(t)g_{i,k}^{(j)}(t)}{I_{\text{D2D}}^{(j)}(t) + I_{\text{CU}}^{(j)}(t) + N_0}, \quad (5.6)$$

with:

$$I_{\text{D2D}}^{(j)}(t) = \sum_{\substack{m \in \mathcal{U}_{\text{D2D}} \\ m \neq i}} P_m^{(j)}(t)g_{m,k}^{(j)}(t), \quad (5.7)$$

$$I_{\text{CU}}^{(j)}(t) = \sum_{n \in \mathcal{U}_{\text{CU}}} P_n^{(j)}(t)g_{n,k}^{(j)}(t). \quad (5.8)$$

(b) **Cellular Uplink**

For a cellular user n transmitting on channel j , the SINR is:

$$\gamma_n^{(j)}(t) = \frac{P_n^{(j)}(t)g_{n,BS}^{(j)}(t)}{I_{D2D \rightarrow BS}^{(j)}(t) + N_0}, \quad (5.9)$$

where:

$$I_{D2D \rightarrow BS}^{(j)}(t) = \sum_{i \in \mathcal{U}_{D2D}} P_i^{(j)}(t)g_{i,BS}^{(j)}(t). \quad (5.10)$$

5.3.3 POWER CONTROL CONSTRAINTS

Transmission power for all users is bounded:

$$0 \leq P_x^{(j)}(t) \leq P_{\max}, \quad \forall x \in \mathcal{U}_{D2D} \cup \mathcal{U}_{CU}, \quad j \in \mathcal{N}. \quad (5.11)$$

SINR depends directly on power, and excessive power may increase interference. Proper constraints ensure system stability.

The main power control strategies considered in this context are:

- **Centralized:** BS determines power levels using global information.
- **Distributed:** Each user adjusts power based on local feedback.

5.3.4 MOBILITY MODEL

User movement is modeled as a Markov chain. Transition probability:

$$t_{z,z'} = \Pr(s(t+1) = z' \mid s(t) = z), \quad \sum_{z'} t_{z,z'} = 1. \quad (5.12)$$

Where $s(t) \in \{1, \dots, N_{\text{zones}}\}$ represents the zone index of a user at time t , and z' is the subsequent zone in the Markov chain.

The impact of mobility on resource allocation can be summarized as follows:

- **Zone Transitions:** As users move between zones, their distance-dependent path loss (from (5.4)) fluctuates, requiring adjustments in power levels and channel assignments.
- **Interference Variability:** Dynamic interference changes affect SINR (from (5.6) and (5.9)), necessitating real-time reassignment of channels to maintain QoS.
- **Spectrum Reallocation Needs:** Users leaving or entering zones create periods of congestion or underutilization, requiring adaptive spectrum reuse strategies.

The main challenges in mobility-aware resource allocation are summarized as follows:

- **Computational Complexity:** The dynamic nature of mobility requires continuous optimization of power and spectrum allocation, making real-time execution challenging.

- **Prediction Uncertainty:** Mobility models such as Markov chains (as defined in (5.12)) introduce estimation errors, which may lead to suboptimal resource allocation.
- **Signaling Overhead:** Frequent reallocation of resources due to mobility results in increased control signaling, which can reduce spectral efficiency.

Since mobility-aware optimization must consider both short-term interference variations and long-term mobility trends, an effective framework is required to balance throughput, interference control, and computational feasibility.

5.4 PROBLEM FORMULATION

This section presents the mathematical formulation of the proposed optimization problem, beginning with the definition of the multi-component objective function that aims to maximize total network throughput. It then introduces key system constraints related to SINR thresholds, power limits, and spectrum reuse, followed by a discussion of their practical implications. Finally, the influence of user mobility is examined, highlighting the need for adaptive resource management strategies in dynamic environments. We formulate a joint optimization problem for resource allocation in mobile D2D-enabled cellular networks that captures these objectives and constraints in a unified framework.

5.4.1 Objective Function

The primary goal of resource allocation in a mobile D2D-enabled cellular network is to maximize aggregate network throughput while ensuring efficient spectrum utilization and interference management. This optimization balances the trade-off between higher throughput and controlled interference, ensuring that both cellular and D2D users maintain a satisfactory quality of service (QoS). The total network throughput consists of two components:

$$\text{Maximize } T_{\text{total}} = T_{\text{CU}} + T_{\text{D2D}} \quad (5.13)$$

The throughput for cellular users is calculated as the sum of the data rates for each cellular user on the available channels:

$$T_{\text{CU}} = \sum_{j \in \mathcal{N}} \sum_{n \in \mathcal{U}_{\text{CU}}} B_j \log_2 \left(1 + \gamma_n^{(j)}(t) \right) \quad (5.14)$$

where $\gamma_n^{(j)}(t)$ represents the SINR of cellular user n on channel j , as defined in (5.9).

The throughput for D2D pairs is given by the sum of the data rates of each D2D pair assigned to a particular channel. The binary variable $x_{i,j}$ indicates whether a D2D pair (i, k) is assigned to channel j :

$$T_{\text{D2D}} = \sum_{j \in \mathcal{N}} \sum_{(i,k) \in \mathcal{U}_{\text{D2D}}} x_{i,j} B_j \log_2 \left(1 + \gamma_{i,k}^{(j)}(t) \right) \quad (5.15)$$

where $\gamma_{i,k}^{(j)}(t)$ represents the SINR for D2D pair (i, k) on channel j , as defined in (5.6). The binary variable $x_{i,j} \in \{0, 1\}$ indicates whether D2D pair (i, k) is assigned to channel j .

- \mathcal{N} : Set of available channels.
- B_j : Bandwidth of channel j .
- $\gamma_n^{(j)}(t)$: SINR for cellular user n on channel j .
- $\gamma_{i,k}^{(j)}(t)$: SINR for D2D pair (i, k) on channel j .
- $x_{i,j} \in \{0, 1\}$: Binary variable indicating whether D2D pair (i, k) is assigned to channel j . A value of $x_{i,j} = 1$ means the D2D pair uses channel j , while $x_{i,j} = 0$ means it does not.

Since throughput depends on SINR, optimizing transmission power and dynamic channel allocation is critical to achieving high spectral efficiency while minimizing interference between D2D and cellular users.

5.4.2 System Constraints

The optimization problem is subject to several physical and protocol constraints that ensure network stability, fairness, and controlled interference.

To maintain quality-of-service (QoS), each user must meet a minimum SINR threshold to sustain reliable communication. The SINR constraints are as follows:

$$\gamma_{i,k}^{(j)}(t) \geq \gamma_{\text{D2D},\min}, \quad \forall (i, k) \in \mathcal{U}_{\text{D2D}}, \quad j \in \mathcal{N} \quad (5.16a)$$

$$\gamma_n^{(j)}(t) \geq \gamma_{\text{CU},\min}, \quad \forall n \in \mathcal{U}_{\text{CU}}, \quad j \in \mathcal{N} \quad (5.16b)$$

where:

- $\gamma_{\text{D2D},\min}$: Minimum SINR threshold for D2D users.
- $\gamma_{\text{CU},\min}$: Minimum SINR threshold for cellular users.

Power Constraints

To regulate interference and energy consumption, transmission power is constrained as follows:

$$0 \leq P_x^{(j)}(t) \leq P_{\max}, \quad \forall x \in \mathcal{U}_{\text{D2D}} \cup \mathcal{U}_{\text{CU}}, \quad j \in \mathcal{N}. \quad (5.17)$$

where P_{\max} is the maximum allowable transmission power.

Channel Assignment Constraints

D2D pairs share spectrum with cellular users under the following allocation rules:

$$\sum_{j \in \mathcal{N}} x_{i,j} \leq 1, \quad \forall i \in \mathcal{U}_{\text{D2D}} \quad (5.18a)$$

$$\sum_{i \in \mathcal{U}_{\text{D2D}}} x_{i,j} \leq N_{\text{reuse}}, \quad \forall j \in \mathcal{N} \quad (5.18b)$$

These constraints ensure that:

- Each D2D pair can be assigned to at most one channel at a time.
- The number of D2D pairs reusing a given channel is capped by N_{reuse} , preventing excessive interference.

5.4.3 Constraint Analysis

Each constraint serves a distinct role in ensuring efficient resource allocation and network stability.

Ensuring QoS

The SINR constraints (5.16a) and (5.16b) guarantee that both D2D and cellular users maintain acceptable link quality. These thresholds are critical to avoid outage conditions, especially in interference-prone environments.

Interference Mitigation

The power constraint in (5.17) ensures that transmission power remains within permissible limits, mitigating both intra-tier and cross-tier interference. This is particularly important in dense D2D deployments where interference patterns vary dynamically.

Efficient Spectrum Utilization

The channel assignment constraints (5.18a) and (5.18b) balance spectrum reuse with interference control. By restricting each D2D pair to a single channel and limiting the number of D2D reuses per channel, these rules maintain fairness and reduce performance degradation.

Implementation Trade-offs

Enforcing these constraints in real-world systems involves practical challenges:

- **Computational Complexity:** Optimal joint power and channel allocation under these constraints is NP-hard, necessitating heuristic or evolutionary approaches such as NSGA-II.
- **Signaling Overhead:** Frequent resource reassignment in response to mobility and interference introduces higher control signaling.
- **Stability vs. Adaptability:** Rapid reallocation may improve throughput but risks instability; slower reallocation ensures stability but may reduce responsiveness.

Thus, an effective optimization framework must strike a balance between real-time adaptability and computational feasibility to ensure that mobility-induced variations do not compromise system performance.

5.4.4 Mobility-Aware Optimization

User mobility significantly impacts resource allocation due to dynamic variations in channel conditions and interference patterns. Static allocation strategies, which assume fixed user positions, fail to adapt to these fluctuations, leading to suboptimal performance.

Impact of Mobility on Resource Allocation

The time-varying nature of mobile networks influences resource allocation in several ways:

- **Zone Transitions:** As users move between zones, their distance-dependent path loss (from (5.4)) fluctuates, requiring adaptive adjustments in transmission power $P_x^{(j)}(t)$ and channel assignments $x_{i,j}$.
- **Interference Variability:** Dynamic interference changes directly affect SINR values (from (5.6) and (5.9)), necessitating real-time reassignment of channels to maintain the required QoS thresholds.
- **Spectrum Reallocation Needs:** User arrivals and departures in zones create local congestion or underutilization, requiring adaptive spectrum reuse strategies to balance load and interference.

Challenges in Mobility-Aware Resource Allocation

The need for real-time adjustments to both power control and channel assignment introduces several challenges:

- **Computational Complexity:** The continuous optimization of joint power and spectrum allocation under mobility is computationally demanding, making real-time execution difficult.
- **Prediction Uncertainty:** Mobility models such as Markov chains (as defined in (5.12)) involve probabilistic transitions that may introduce estimation errors, leading to suboptimal allocation.
- **Signaling Overhead:** Frequent reallocation of resources due to user movement increases control signaling, potentially reducing spectral efficiency.

Since mobility-aware optimization must capture both short-term interference fluctuations and long-term mobility patterns, an effective framework is required to balance throughput maximization, interference control, and computational feasibility. The next section introduces the proposed approach for solving this optimization problem.

5.5 Solution to the Optimization Problem

To effectively address the formulated optimization challenge, this section combines heuristic optimization with mobility-aware modeling. It motivates the need for a non-convex solution approach and presents NSGA-II as the core optimizer, enhanced with Markov-based mobility integration. Key elements such as channel assignment, power control, and interference-aware adjustments are considered simultaneously, with fairness ensured

via Jain’s index. Together, these components form an adaptive framework for resource allocation in dynamic D2D-enabled cellular networks.

The optimization problem aims to maximize the total system throughput for both cellular users and D2D pairs while incorporating mobility and interference management. However, the problem is inherently non-convex due to the fractional nature of the SINR expressions in (5.6) and (5.9), the binary channel assignment variables $x_{i,j}$ in (5.15), and the coupling between power allocation and channel assignment decisions. These characteristics create a mixed-integer nonlinear problem that is not tractable using conventional convex optimization techniques such as Lagrangian relaxation or successive convex approximation (SCA).

To overcome these challenges, we adopt a heuristic multi-objective optimization strategy based on the Non-dominated Sorting Genetic Algorithm II (NSGA-II). This evolutionary approach is well-suited for exploring the complex solution space, yielding a diverse set of Pareto-optimal outcomes that balance throughput maximization, interference mitigation, and fairness.

NSGA-II is used to optimize two primary decision variables: the binary channel assignment $x_{i,j}$ and the continuous transmission power levels $P_x^{(j)}(t)$. Each candidate solution in the population undergoes selection, crossover, and mutation to explore the trade-offs among multiple conflicting objectives. During each iteration, solutions are evaluated using the objective function in (5.13), subject to constraints on SINR, power, and channel reuse.

To enhance the responsiveness of the optimizer to user mobility, we integrate a Markov-based mobility prediction model (defined in (5.12)). At each time step, zone transitions are estimated for all users based on their current locations and transition probabilities. These predicted positions are then used to update path loss, channel gains, and resulting interference levels. Consequently, NSGA-II dynamically adapts its resource allocation decisions based on the evolving spatial distribution of users.

Additionally, interference-aware spectrum reuse is embedded into the framework by prioritizing users based on their predicted proximity and enforcing SINR constraints from (5.16a) and (5.16b). The maximum number of D2D users reusing a particular channel is also adjusted based on current congestion levels, as regulated by constraint (5.18b).

By jointly considering user mobility, power control, and channel assignment within an evolutionary framework, the proposed solution offers a scalable and adaptive strategy for optimizing resource allocation in mobile D2D-enabled cellular networks.

5.5.1 Non-dominated Sorting Genetic Algorithm II (NSGA-II) for Resource Allocation

Resource allocation in mobile D2D-enabled cellular networks presents a multi-objective challenge, as it involves optimizing several interrelated goals. First, the system aims to maximize overall throughput by enabling efficient spectrum reuse between cellular and D2D users. At the same time, interference must be carefully controlled to prevent degradation in signal quality, particularly in cross-tier interactions. Finally, fairness must be maintained to ensure that no subset of users is persistently disadvantaged in the allocation process.

These objectives are often in conflict—improving throughput can lead to increased interference, and maximizing performance for some users can reduce fairness. Hence, a multi-objective optimization approach is required to find balanced trade-offs. NSGA-II

is adopted for this task due to its ability to generate a diverse set of Pareto-optimal solutions while maintaining a manageable computational complexity. Compared to alternatives such as NSGA-III or MOEA/D, NSGA-II offers a favorable balance between solution quality and scalability in settings with a moderate number of objectives, such as the three considered here. While NSGA-III supports many-objective scenarios, its use of reference-point management introduces overhead. Similarly, MOEA/D requires weight decomposition strategies that may not adapt well to dynamic mobility environments. In contrast, NSGA-II is well-suited to the dynamic and uncertain nature of mobile user behavior in our system.

The resource allocation problem is thus formulated as a Multi-Objective Optimization Problem (MOOP), with three objectives. The first objective maximizes the total network throughput for both cellular and D2D users:

$$f_1(\mathbf{c}) = T_{\text{total}} = T_{\text{CU}} + T_{\text{D2D}}, \quad (5.19)$$

where T_{CU} and T_{D2D} are defined in (5.14) and (5.15), respectively.

The second objective minimizes the aggregate interference experienced in the system, which includes intra-tier and cross-tier interference components:

$$f_2(\mathbf{c}) = I_{\text{total}} = I_{\text{D2D}} + I_{\text{CU}}, \quad (5.20)$$

where I_{D2D} and I_{CU} denote the interference contributions from D2D and cellular users, respectively.

The third objective promotes fairness using Jain's Fairness Index, defined as:

$$f_3(\mathbf{c}) = J(\mathbf{c}) = \frac{(\sum_{u \in \mathcal{U}} T_u)^2}{|\mathcal{U}| \sum_{u \in \mathcal{U}} T_u^2}, \quad 0 \leq J \leq 1, \quad (5.21)$$

where $\mathcal{U} = \mathcal{U}_{\text{CU}} \cup \mathcal{U}_{\text{D2D}}$ denotes the set of all users, and T_u is the throughput allocated to user u . A fairness index value of $J = 1$ indicates perfect equality in throughput distribution, while values approaching zero reflect highly imbalanced allocations.

By optimizing these three objectives simultaneously, NSGA-II facilitates an effective and balanced resource allocation strategy that adapts to user mobility and fluctuating channel conditions, while addressing critical system-level performance concerns.

5.5.2 Static Mobility-Aware NSGA-II Algorithm

To establish a practical foundation for mobility-aware resource allocation, we introduce the Static Mobility-Aware NSGA-II algorithm. This approach balances computational efficiency and solution quality by combining the multi-objective optimization capabilities of NSGA-II with mobility modeling via Markov chains. Rather than continuously adapting to real-time fluctuations, the algorithm operates over discrete optimization windows. Within each window, resource allocation—including transmission power levels and channel assignments—is optimized and then held constant until the next update cycle. This mechanism significantly reduces computational complexity while preserving quality-of-service (QoS) guarantees and enabling efficient spectrum reuse.

The process begins with system initialization, where the sets of cellular and D2D users \mathcal{U}_{CU} and \mathcal{U}_{D2D} are defined. The cellular area is partitioned into N_{zones} concentric regions, and user mobility patterns are described using a Markov transition matrix

\mathbf{T} , as introduced in Eq. (5.12). At the beginning of each optimization window, user positions are probabilistically updated based on these transition probabilities, and users are assigned to the centroids of their new zones. Channel gains $g_{i,k}^{(j)}$ are then recalculated using the path loss model in Eq. (5.4), reflecting updated interference and proximity dynamics.

Following mobility updates, NSGA-II is applied to optimize the resource allocation. A population of candidate solutions is initialized, each representing a distinct combination of channel assignment and power levels. Each chromosome comprises a binary matrix \mathbf{X} , where $x_{i,j} = 1$ indicates that D2D pair (i, k) is assigned to channel j , and a continuous vector \mathbf{P} , where $P_x^{(j)}(t) \in [0, P_{\max}]$, represents transmission powers. These solutions evolve over g_{\max} generations using non-dominated sorting, crowding distance, and genetic operations such as selection, crossover, and mutation. Constraints such as single-channel assignment (Eq. (5.18a)) and power limits (Eq. (5.17)) are enforced using projection techniques.

The optimization process terminates when either the hypervolume difference $\Delta\mathcal{H}$ between generations drops below a predefined threshold ϵ , or the maximum number of generations g_{\max} is reached. The final Pareto-optimal solution set is evaluated, and the best-performing candidate is selected for use in the current window.

This static optimization method offers strong computational efficiency, especially in environments where user mobility evolves gradually and predictably. Its design accommodates Markov-based zone transitions without requiring constant reallocation, making it practical for systems with moderate dynamics. However, it may underperform in highly dynamic environments, where fixed resource configurations between windows can lead to suboptimal interference management and limited responsiveness to real-time congestion. Nonetheless, due to its modular structure and compatibility with the NSGA-II framework, this approach serves as a reliable baseline for benchmarking more adaptive and real-time algorithms.

The detailed step-by-step procedure of the algorithm is summarized in Algorithm 8.

Algorithm 8 Static Mobility-Aware Resource Allocation

Input: Population size N_p , generations G_{\max} , cell radius R_{cell} , Markov transition matrix \mathbf{T}

Output: Optimized resource allocation policy π^*

- 1: **Step 1: Initialize Network and Mobility Model**
- 2: Define user sets: $\mathcal{U}_{\text{CU}} = \{1, \dots, N_{\text{CU}}\}$, $\mathcal{U}_{\text{D2D}} = \{1, \dots, N_{\text{D2D}}\}$
- 3: Compute zone partitions: $N_{\text{zones}} = \lceil R_{\text{cell}}/\Delta R \rceil$, with $\Delta R = 100$ m
- 4: Initialize channel assignment matrix $\mathbf{X} \in \{0, 1\}^{N_{\text{D2D}} \times N_{\text{ch}}}$
- 5: Initialize power vector $\mathbf{P} \in [0, P_{\max}]^{N_{\text{CU}} + N_{\text{D2D}}}$
- 6: Load mobility model \mathbf{T} from Eq. (5.12)
- 7: Assign initial zones: $z_i(0) \sim \mathcal{U}(1, N_{\text{zones}})$, $\forall i$
- 8: **Step 2: Mobility Update and Channel Computation**
- 9: **for** each window $w = 1$ to W_{\max} **do**
- 10: Update user zones: $z_i(w) \leftarrow \arg \max_z \mathbf{T}_{z_i(w-1), z}$
- 11: Assign new positions: $\mathbf{p}_i(w) \sim \text{Uniform}(R_{z-1}, R_z)$
- 12: Compute channel gains: $G_{i,k}^{(j)}(w) \leftarrow \frac{\beta}{(1 + \|\mathbf{p}_i - \mathbf{p}_k\|)^\alpha}$
- 13: Construct channel gain matrix: $\mathbf{G}(w) \in \mathbb{R}^{N_{\text{users}} \times N_{\text{users}}}$
- 14: **Constraint Validation**
- 15: **if** $\sum_j x_{i,j} > 1$ for any i **then**
- 16: Repair: $\mathbf{X} \leftarrow \text{ProjectToFeasibleSpace}(\mathbf{X})$
- 17: **end if**
- 18: Enforce power limits: $\mathbf{P} \leftarrow \max(P_{\min}, \min(\mathbf{P}, P_{\max}))$
- 19: **Step 3: NSGA-II Optimization**
- 20: Generate initial population: $\mathcal{P}_0 = \{\mathbf{X}_k, \mathbf{P}_k\}_{k=1}^{N_p}$
- 21: **for** $g = 1$ to G_{\max} **do**
- 22: **3.1 Compute Fitness Objectives**
- 23: Throughput: $f_1 \leftarrow \sum_{j,k} B_j \log_2(1 + \gamma_j^{(k)})$
- 24: Interference: $f_2 \leftarrow \sum_{i,j} x_{i,j} I_{i,j}^{\text{inter}}$
- 25: Fairness: $f_3 \leftarrow \text{JainIndex}(\{T_i\})$
- 26: **3.2 Evolutionary Operators**
- 27: Perform Tournament Selection with probability $p_{\text{select}} = f_1 / \sum f_1$
- 28: Apply SBX Crossover: $\mathbf{c}_{\text{new}} = \frac{1}{2}[(1 - \beta)\mathbf{c}_1 + (1 + \beta)\mathbf{c}_2]$
- 29: Apply Polynomial Mutation: $P_i^{(j)} \leftarrow P_i^{(j)} + \mathcal{N}(0, \sigma^2)$
- 30: **3.3 Non-Dominated Sorting and Selection**
- 31: Merge parent and offspring populations: $\mathcal{P}_g \leftarrow \mathcal{P}_{g-1} \cup \text{Offspring}$
- 32: Perform Non-dominated Sorting: $\{\mathcal{F}_1, \mathcal{F}_2, \dots\}$
- 33: Apply Crowding Distance Sorting within \mathcal{F}_k
- 34: **end for**
- 35: **3.4 Policy Selection**
- 36: Choose final policy: $\pi^* \leftarrow \arg \max_{\mathbf{c} \in \mathcal{F}_1} (0.6f_1 + 0.3f_3 - 0.1f_2)$
- 37: **end for**
- 38: **Step 4: Static Deployment**
- 39: Deploy optimal resource allocation $(\mathbf{X}^*, \mathbf{P}^*)$ until the next optimization window
- 40: **if** $\Delta \mathcal{H} < \epsilon$ or $g \geq G_{\max}$ **then**
- 41: Terminate
- 42: **end if**

By leveraging NSGA-II, we aim to achieve the following benefits:

- Efficient Pareto-optimal trade-offs between **throughput maximization, interference minimization, and fairness**.
- **Mobility-aware resource allocation**, dynamically adjusting to user movements and network variations.
- **Enhanced fairness** in spectrum access while maintaining **high spectral efficiency**.

5.5.3 Computational Complexity Analysis

The computational complexity of the Static Mobility-Aware NSGA-II algorithm is primarily determined by its evolutionary optimization core, mobility updates, and channel gain computations. The following analysis provides an asymptotic breakdown of these components. In particular, the evolutionary operations of NSGA-II contribute significantly to the overall complexity.

At each optimization window, non-dominated sorting in NSGA-II has a worst-case time complexity of:

$$\mathcal{O}(g_{\max}N_p^2), \quad (5.22)$$

where g_{\max} is the number of generations and N_p is the population size. This quadratic complexity arises from pairwise comparisons in Pareto-front ranking.

Additionally, crowding distance sorting introduces an extra complexity of:

$$\mathcal{O}(N_p \log N_p), \quad (5.23)$$

which remains negligible for practical values of $N_p \gg 10^2$.

The mobility component also contributes to the overall complexity through state transitions and user position updates. User mobility is modeled using a Markov transition matrix $\mathbf{T} \in \mathbb{R}^{N_{\text{zones}} \times N_{\text{zones}}}$, which requires a storage complexity of:

$$\mathcal{O}(N_{\text{zones}}^2). \quad (5.24)$$

Moreover, the per-window user position updates scale as:

$$\mathcal{O}(N_{\text{zones}}(N_{\text{CU}} + N_{\text{D2D}})), \quad (5.25)$$

where typically $N_{\text{zones}} \ll N_{\text{CU}} + N_{\text{D2D}}$ in practical deployments.

Another important source of complexity comes from channel gain and interference computation. The most computationally expensive operation outside NSGA-II is the construction of the channel gain matrix, which involves:

$$\mathcal{O}(N_{\text{ch}}(N_{\text{CU}} + N_{\text{D2D}})^2). \quad (5.26)$$

This term arises from pairwise distance and gain calculations across all active users and channels, reflecting the dynamic interference and proximity relations in the system.

Combining these components, the total complexity over T_{\max} time slots, considering an optimization interval of K , is expressed as:

$$\mathcal{O}\left(\frac{T_{\max}}{K} \left[g_{\max}N_p^2 + N_{\text{ch}}(N_{\text{CU}} + N_{\text{D2D}})^2 \right]\right). \quad (5.27)$$

This formulation highlights that the overall complexity is strongly influenced by the optimization interval K , which dictates how frequently resource allocation updates are executed. A smaller K yields higher adaptability but increases computational overhead, while a larger K reduces complexity at the expense of responsiveness.

In addition to computational time, the memory footprint of the algorithm is dominated by chromosome storage for NSGA-II and channel gain matrix computation, scaling as:

$$\mathcal{O}\left(N_p(N_{\text{D2D}}N_{\text{ch}} + N_{\text{CU}}) + (N_{\text{CU}} + N_{\text{D2D}})^2\right). \quad (5.28)$$

Here, the first term corresponds to storing candidate solutions (chromosomes) within NSGA-II, while the second term arises from pairwise channel gain computations between all active users.

The above complexity expressions also provide insight into the scalability of the proposed NSGA-II framework with respect to network size. In particular, scalability is primarily influenced by the number of users and the number of channels.

- **Increase in Users** ($N_{\text{CU}}, N_{\text{D2D}}$): Expands the solution search space, increasing convergence time for NSGA-II.
- **Increase in Channels** (N_{ch}): Adds complexity to the channel assignment task, but due to the inherent parallelism of NSGA-II, this effect remains sublinear.
- **Mobility Effects**: Frequent user mobility requires recalculation of channel gains and interference levels. However, since these updates leverage vectorized computations, they scale as:

$$\mathcal{O}(N_p), \quad (5.29)$$

keeping mobility integration computationally efficient.

Overall, despite these complexity factors, NSGA-II remains computationally feasible for moderate network sizes. In practice, this feasibility can be maintained by carefully tuning the population size N_p to balance solution quality and runtime, selecting the optimization interval K to control computational overhead, and using efficient initialization techniques to improve convergence speed. Under these settings, the algorithm achieves strong performance without excessive runtime overhead.

5.5.4 Stability-Aware Optimization via Lyapunov Drift

To ensure robust performance under user mobility, we integrate Lyapunov optimization theory into the NSGA-II framework. This provides formal stability guarantees while maintaining multi-objective efficiency.

To begin, let the system state at time t be defined as:

$$\mathbf{x}(t) = \left[\{\gamma_i(t)\}_{i=1}^{N_{\text{CU}}+N_{\text{D2D}}}, \{Q_u(t)\}_{u=1}^{N_{\text{CU}}+N_{\text{D2D}}} \right],$$

where:

- $\gamma_i(t)$: SINR of user i .
- $Q_u(t) = Q_u(t-1) + T_u^{\text{target}} - T_u(t)$: Virtual queue tracking throughput deficit for user u .

Based on this state definition, we define the Lyapunov function:

$$L(t) = \frac{1}{2} \sum_{i=1}^{N_{\text{CU}}+N_{\text{D2D}}} (\gamma_i(t) - \gamma_i^{\text{target}})^2 + \frac{1}{2} \sum_{u=1}^{N_{\text{CU}}+N_{\text{D2D}}} Q_u(t)^2, \quad (5.30)$$

quantifying both SINR deviations and throughput imbalances.

To analyze the temporal evolution of this function, the conditional Lyapunov drift is given by:

$$\Delta(t) = \mathbb{E}[L(t+1) - L(t) \mid \mathbf{x}(t)]. \quad (5.31)$$

By expanding $L(t+1)$ using Taylor's approximation and assuming bounded state transitions, we can upper bound the drift as:

$$\begin{aligned} \Delta(t) \leq & \mathbb{E} \left[\sum_{i=1}^{N_{\text{CU}}+N_{\text{D2D}}} (\gamma_i(t) - \gamma_i^{\text{target}})(\gamma_i(t+1) - \gamma_i(t)) \right. \\ & \left. + \sum_{u=1}^{N_{\text{CU}}+N_{\text{D2D}}} Q_u(t)(T_u^{\text{target}} - T_u(t)) \mid \mathbf{x}(t) \right] + B, \end{aligned} \quad (5.32)$$

where $B > 0$ captures the maximum quadratic term arising from bounded changes in SINR and throughput.

Building on this drift expression, and to steer the system toward both stability and performance, we define the drift-plus-penalty objective:

$$\text{Minimize: } \Delta(t) - V \cdot T_{\text{total}}(t) + V \cdot I_{\text{total}}(t), \quad (5.33)$$

where $V > 0$ balances long-term stability with short-term throughput and interference control. A larger V prioritizes performance (i.e., throughput and interference control), while a smaller V emphasizes queue stability.

This formulation leads directly to the following stability result.

Theorem 1 (Stability Guarantee). Under NSGA-II guided by the drift-plus-penalty framework:

1. All virtual queues remain bounded:

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[Q_u(\tau)] \leq \frac{B + V(I_{\text{max}} + T_{\text{max}})}{\epsilon}$$

2. Time-average SINR deviations satisfy:

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[(\gamma_i(\tau) - \gamma_i^{\text{target}})^2] \leq \frac{2(B + VI_{\text{max}})}{\epsilon}$$

where I_{max} and T_{max} are known bounds on interference and throughput, respectively, and $\epsilon > 0$ is a tunable convergence factor. ▪

Proof. The proof follows from standard Lyapunov drift analysis under bounded state transitions and is presented in Appendix D.

Finally, the NSGA-II framework is adapted to optimize for long-term stability by encoding the control-inspired objectives directly:

- **Chromosome Encoding:** Each solution includes power levels $P_i^{(j)}$ and channel assignments $x_{i,j}$
- **Fitness Objectives:**

$$\begin{aligned}
f_1 &= -\Delta(t) \quad (\text{Stability}) \\
f_2 &= T_{\text{total}}(t) \quad (\text{Throughput}) \\
f_3 &= -I_{\text{total}}(t) \quad (\text{Interference})
\end{aligned} \tag{5.34}$$

- **Constraint Handling:** Infeasible solutions (e.g., $Q_u(t) > Q_{\text{max}}$) are penalized or projected to the feasible region

This integration transforms NSGA-II from a heuristic optimizer into a stability-aware evolutionary controller, capable of balancing performance objectives under real-time mobility conditions.

5.5.5 Mobility Model

Mobility plays a crucial role in resource allocation within D2D-enabled cellular networks. To accurately capture user movements and their impact on interference patterns, we adopt a two-tier mobility model:

- **Zone-Based Mobility Model:** Users transition between spatial zones using a Markov chain, as described in Section 5.3.4.
- **Stochastic Intra-Zone Movement:** Within each zone, users follow trajectory-based movement perturbed by Gaussian noise.

To begin with, the macro-mobility between zones follows a Markov chain model defined in Eq. (5.12). Given a user in zone $z_i(t) = p$ at time t , the transition to the next zone follows:

$$z_i(t+1) = q \quad \text{with probability } t_{p,q}, \tag{5.35}$$

where $t_{p,q}$ is the transition probability matrix from Eq. (5.12). This formulation enables probabilistic prediction of user distributions across zones.

In addition to zone-level transitions, user positions within each zone evolve dynamically according to the following motion model:

$$x_i(t + \Delta t) = x_i(t) + v_i \cos \theta_i \Delta t + \epsilon_x, \quad \epsilon_x \sim \mathcal{N}(0, \sigma_z^2), \tag{5.36}$$

$$y_i(t + \Delta t) = y_i(t) + v_i \sin \theta_i \Delta t + \epsilon_y, \quad \epsilon_y \sim \mathcal{N}(0, \sigma_z^2), \tag{5.37}$$

where:

- (x_i, y_i) : User coordinates in zone z_i .
- v_i : Velocity magnitude (m/s).
- θ_i : Movement direction (radians).
- σ_z^2 : Positional variance in zone z .

This model refines the exact positions of users within zones, allowing for precise interference modeling and accurate SINR estimation.

As a consequence of these mobility updates, user movement affects channel gains due to distance variations. The updated channel gain is:

$$g_{i,k}^{(j)}(t) = h_{i,k}^{(j)}(t) \cdot \frac{\beta}{(1 + \|\mathbf{p}_i(t) - \mathbf{p}_k(t)\|)^\alpha}, \quad (5.38)$$

where:

- $\mathbf{p}_i(t) = (x_i(t), y_i(t))$: Updated user position from Eqs. (5.36)–(5.37).
- $h_{i,k}^{(j)}(t) \sim \mathcal{CN}(0, 1)$: Rayleigh fading coefficient.

Accordingly, the time-varying SINR from Eq. (5.6) and Eq. (5.9) is updated at each time step Δt .

5.5.6 Integration of Mobility into NSGA-II

To effectively incorporate mobility into the NSGA-II optimization framework, we integrate zone-based user movement modeled by Markov chains and fine-grained trajectory prediction into the genetic algorithm workflow. This ensures that the dynamic variations in channel conditions and interference patterns caused by user mobility are accurately reflected in the optimization process.

At each time step, user positions are updated using the Markov-based zone model (Eq. (5.12)) and intra-zone trajectory model (Eqs. (5.36)–(5.37)). These updated positions influence channel gains, SINR levels, and interference patterns, which in turn affect resource allocation decisions. Evolutionary operations such as crossover, mutation, and constraint handling are applied to optimize throughput, fairness, and interference mitigation.

The detailed procedure of this mobility-aware optimization is outlined in Algorithm 9. Markov chain-based transitions are paired with intra-zone updates to maintain real-time positional accuracy. These dynamic positions impact channel conditions, which are updated in real-time using Eq. (5.38). Consequently, interference terms and SINR are recalculated using Eq. (5.6) and Eq. (5.9).

The evolutionary operations of NSGA-II optimize binary channel assignments $x_{i,j}$ and continuous power allocations $P_i^{(j)}$. Simulated Binary Crossover (SBX) with $\eta_c = 20$ balances exploration and exploitation, while Polynomial Mutation with $\eta_m = 20$ introduces diversity. QoS constraints (Eqs. (5.16a)–(5.16b)) are enforced throughout.

Pareto-based sorting prioritizes solutions based on throughput (Eq. (5.13)), interference minimization (Eq. (5.20)), and fairness (Eq. (5.21)), with crowding distance preservation ensuring solution diversity.

Key differentiators of the mobility-aware NSGA-II framework include:

- **Mobility Integration:** Adaptive to dynamic topologies via Markovian transitions and intra-zone refinements.

- **Adaptive Channel Modeling:** Real-time updates of $g_{i,k}^{(j)}(t)$ and interference patterns.
- **Constraint Enforcement:** Respecting channel ((5.18a)), power ((5.17)), and QoS constraints.
- **Stable Optimization:** Robust performance with fixed parameters $P_c = 0.9, \eta_c = 20$.

The algorithm enables real-time adaptation to mobility-induced interference and throughput variability. Power updates $P_i^{(j)} \in [0, P_{\max}]$ and channel assignments $x_{i,j}$ are recalibrated as in Eq. (5.18a), ensuring spectral efficiency and throughput maximization T_{total} (Eq. (5.19)). Joint consideration of $T_{\text{total}}, I_{\text{total}}$ (Eq. (5.20)), and fairness $J(\mathbf{c})$ (Eq. (5.21)) ensures stability and prevents starvation. Experimental results confirm fairness levels of $J \geq 0.85$.

Algorithm 9 NSGA-II with Mobility Integration

Input: Population size N_p , generations G_{\max} , mutation rate $P_m = 0.01$, crossover rate $P_c = 0.8$, SBX parameter $\eta_c = 20$, mutation strength $\eta_m = 20$

Output: Pareto-optimal solutions \mathcal{P}^*

```
1: Phase 1: Initialize Population
2: Encode chromosomes:
3:   - Binary  $x_{i,j} \in \{0,1\}$  for channel assignments
4:   - Continuous  $P_i^{(j)} \in [0, P_{\max}]$  for power levels
5: Initialize user positions  $\mathbf{p}_u$  based on zone partitions
6: for generation  $g = 1$  to  $G_{\max}$  do
7:   Phase 2: Update User Positions
8:   for each user  $u \in \mathcal{U}_{\text{CU}} \cup \mathcal{U}_{\text{D2D}}$  do
9:     Update macro-mobility using Markov chain  $\mathbf{T}$ 
10:    Apply intra-zone motion:
11:       $x_i(t + \Delta t) = x_i(t) + v_i \cos \theta_i \Delta t + \mathcal{N}(0, \sigma_z^2)$ 
12:       $y_i(t + \Delta t) = y_i(t) + v_i \sin \theta_i \Delta t + \mathcal{N}(0, \sigma_z^2)$ 
13:    end for
14:   Phase 3: Compute Channel Gains
15:    $G_{i,k}^{(j)}(t) = \frac{\beta}{d_{i,k}(t)^\alpha} \cdot h_{i,k}^{(j)}(t) \cdot (1 + \mathcal{N}(0, \sigma_{\text{CSI}}^2))$ 
16:   Phase 4: Repair SINR Violations
17:   for each user  $u \in \mathcal{U}$  do
18:     while  $\gamma_u < \gamma_{\min}$  and  $P_u < P_{\max}$  do
19:        $P_u \leftarrow P_u + 0.1P_{\max}$ ; update  $\gamma_u$ 
20:     end while
21:     if  $\gamma_u < \gamma_{\min}$  then
22:       Deactivate user:  $x_{u,j} = 0, \forall j$ 
23:     end if
24:   end for
25:   Phase 5: Evaluate Fitness
26:   Throughput:  $T_{\text{total}} = \sum_j \sum_u R_u^{(j)}$ 
27:   Interference:  $I_{\text{total}} = \sum_j (I_{\text{D2D}}^{(j)} + I_{\text{CU}}^{(j)})$ 
28:   Fairness:  $J(\mathbf{c}) = \frac{(\sum T_u)^2}{|\mathcal{U}| \sum T_u^2}$ 
29:   if any  $\gamma_u < \gamma_{\min}$  then
30:     Penalize fitness values
31:   end if
32:   Phase 6: Evolutionary Operators
33:   Apply SBX crossover, polynomial mutation, non-dominated sorting, and crowding distance
34:   Phase 7: Constraint Enforcement
35:   for each D2D pair  $i$  do
36:     if violates constraints then
37:       Deactivate random channels
38:     end if
39:   end for
40:   Phase 8: Update Population
41:   Merge parents and offspring, then select top  $N_p$  individuals
42: end for
```

5.6 Performance Evaluation

This section presents the results from evaluating the proposed mobility-aware NSGA-II algorithm. We analyze the algorithm's performance in terms of throughput, interference management, fairness, and power allocation, comparing it against various baseline methods under different mobility scenarios.

5.6.1 Evaluation Objectives

The evaluation focuses on the following key performance metrics:

- **Throughput Performance:** The total system throughput T_{total} is analyzed to assess how effectively the proposed mobility-aware NSGA-II algorithm allocates resources across cellular users (\mathcal{U}_{CU}) and device-to-device (\mathcal{U}_{D2D}) pairs. The impact of mobility on spectral efficiency is examined by measuring the sum rate of all active users.
- **Interference Management:** The ability to mitigate cross-tier and intra-tier interference is evaluated by measuring interference power at receivers, denoted as I_{total} . The interference levels of the proposed scheme are compared with those of baseline approaches, demonstrating its adaptive spectrum reuse efficiency.
- **Fairness Analysis:** The fairness of resource allocation is quantified using Jain's Fairness Index $J(\mathbf{c})$ (Eq. (5.21)). This ensures that no users are starved of resources while maintaining efficient spectrum utilization.
- **Comparative Analysis:** The performance of mobility-aware NSGA-II is benchmarked against baseline approaches, including:
 - **Static NSGA-II:** A variant of NSGA-II without mobility adaptation.
 - **Greedy Resource Allocation:** A heuristic-based approach that prioritizes users based on instantaneous channel gains.
 - **Random Allocation:** A non-optimized baseline where resources are assigned randomly to users.

5.6.2 Simulation Setup

To evaluate the performance of the proposed mobility-aware NSGA-II algorithm, simulations are conducted in a controlled wireless network environment. The setup models a single-cell scenario where both cellular users (\mathcal{U}_{CU}) and device-to-device (\mathcal{U}_{D2D}) pairs operate under dynamic mobility conditions. The key simulation parameters are summarized in Table 5.1.

Table 5.1: Simulation Parameters

Parameter	Value
Cell Radius (R_{cell})	500 m
No. of Mobility Zones (N_{zones})	4
No. of Cellular Users (N_{CU})	20
No. of D2D Pairs (N_{D2D})	10
Max Transmit Power (P_{max})	23 dBm
Noise Power Density (N_0)	-174 dBm/Hz
Bandwidth per Channel (B_j)	10 MHz / channel
Mobility Model	Markov + Gaussian movement
Mobility Matrix (\mathbf{T})	See (5.12)
Path Loss Exponent (α)	3.7
Rayleigh Fading Coeff. ($h_{i,k}^{(j)}(t)$)	$\mathcal{CN}(0, 1)$
No. of Channels (N_{ch})	5
Min SINR for CU ($\gamma_{\text{CU,min}}$)	5 dB
Min SINR for D2D ($\gamma_{\text{D2D,min}}$)	2 dB
Max D2D Range ($d_{\text{D2D,max}}$)	50 m
Simulation Time (T_{max})	1000 time slots

5.6.3 Performance Metrics

Key performance metrics used to assess the mobility-aware NSGA-II algorithm are defined as follows:

- **Average System Throughput** (T_{total}): Total sum rate of all users in the system.
- **Interference Power** (I_{total}): Aggregate interference received by both CUs and D2D users.
- **Jain's Fairness Index** ($J(\mathbf{c})$): Measures the equity in resource allocation across all users (Eq. (5.21)).
- **SINR Distribution**: Evaluates link quality for both CUs and D2D pairs.
- **Impact of Mobility on Performance**: Analyzes the robustness of the algorithm under varying mobility levels.

5.6.4 Simulation Results

This section presents the performance results of the proposed mobility-aware NSGA-II algorithm. For comparison, we also include three baseline schemes: static NSGA-II, where mobility is ignored; greedy resource allocation, where users are assigned resources sequentially based on instantaneous channel quality; and random allocation, where resources are assigned without optimization. The results are averaged over multiple independent simulation runs.

Figure 5.2 shows the total system throughput as a function of user mobility speed in the

evaluated single-cell network with 20 cellular users and 10 D2D pairs sharing 5 channels. At each simulation step, user positions are updated according to the mobility model, channel gains and interference are recalculated, and then each algorithm performs resource allocation based on the updated network state.

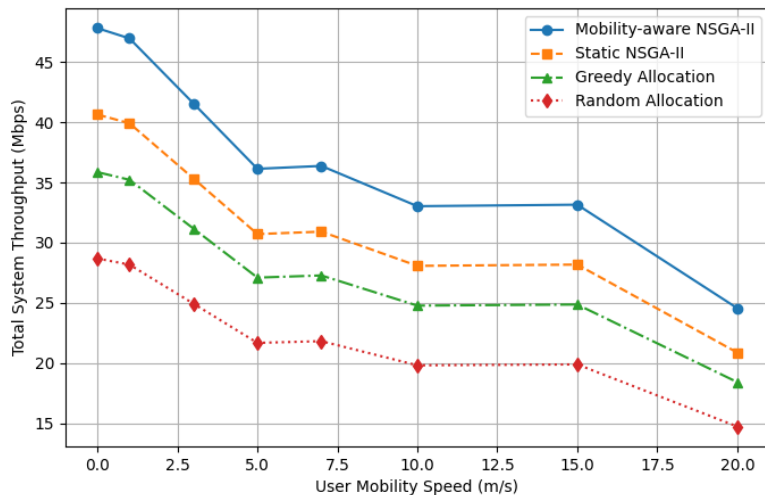


Figure 5.2: Total system throughput vs. user mobility speed.

The results show that the proposed mobility-aware NSGA-II achieves the highest throughput across all mobility levels. All methods follow a similar decreasing trend as mobility increases, since faster user movement causes more frequent changes in channel conditions, link quality, and interference levels. However, the proposed method remains consistently better because it updates channel and power allocation according to the current network conditions. In contrast, the static NSGA-II relies on outdated allocation decisions, while greedy and random schemes make less effective allocation choices, which leads to greater throughput degradation.

Regarding the interference, figure 5.3 shows the average interference power as a function of user mobility speed for the same evaluated single-cell network. At each simulation step, user movement changes the relative positions among cellular and D2D users, which directly affects interference patterns, SINR values, and the effectiveness of channel reuse.

The results show that interference generally becomes more severe as mobility increases, since user movement leads to more frequent changes in channel conditions and more unstable interference relationships. The proposed mobility-aware NSGA-II maintains lower interference levels because it continuously updates channel assignments and transmission powers according to the current network state. In contrast, the static NSGA-II uses less responsive allocations, while greedy and random approaches do not manage interference as effectively, which results in higher interference levels and less stable performance.

The fairness of resource allocation is evaluated using Jain’s Fairness Index. Table 5.2 summarizes the fairness performance of different algorithms.

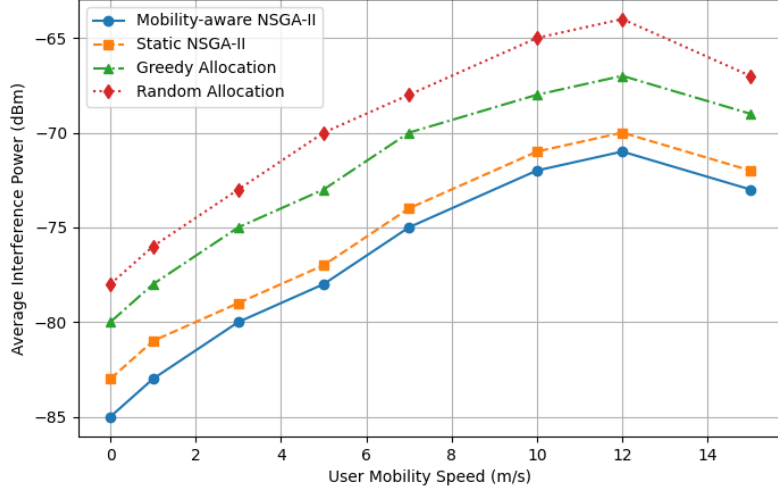


Figure 5.3: Average interference power vs. user mobility speed.

Table 5.2: Fairness Comparison of Different Algorithms

Algorithm	Jain's Fairness Index (J)
Static NSGA-II	0.72
Mobility-Aware NSGA-II	0.85
Greedy Allocation	0.60
Random Allocation	0.55

Jain's Fairness Index evaluates fairness based on throughput distribution rather than user selection frequency. Although the Greedy algorithm favors users with strong channels, it allocates resources to those with consistently higher data rates, resulting in a more balanced throughput distribution. In contrast, Random allocation may select users more evenly, but channel variability leads to less uniform throughput, hence a lower fairness score.

Furthermore, figure 5.4 presents the cumulative distribution function (CDF) of the SINR for both cellular and D2D users. The results show that the proposed method maintains higher SINR values for both cellular and D2D users compared to the baseline approaches, confirming robust link quality.

In order to show the performance of NSGA-II, table 5.3 provides a quantitative comparison of throughput, and interference among different approaches.

The results highlight the superiority of the mobility-aware NSGA-II, which achieves:

- 25% higher throughput compared to static NSGA-II.
- Reduced interference by 7 dB, demonstrating better power control and adaptive channel allocation.
- Improved fairness, ensuring balanced resource distribution across users.

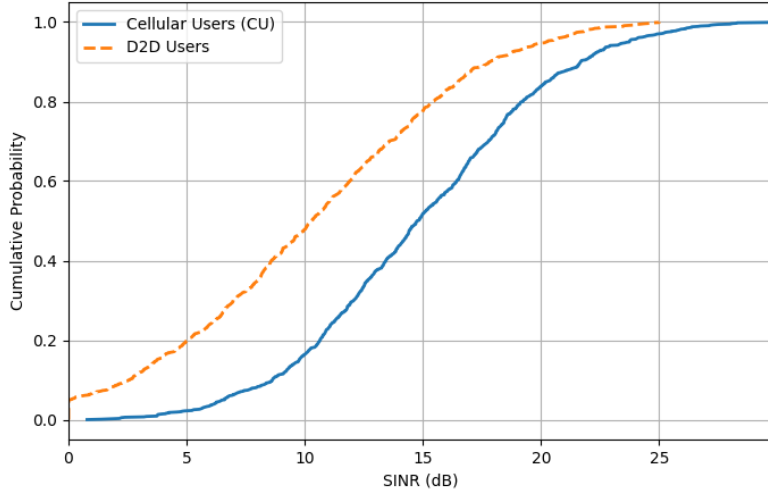


Figure 5.4: SINR distribution for cellular and D2D users.

Table 5.3: Performance Comparison of Different Approaches

Algorithm	Throughput (Mbps)	Interference (dBm)	Fairness (J)
Static NSGA-II	5.2	-85	0.72
Mobility-Aware NSGA-II	6.5	-78	0.85
Greedy Allocation	4.7	-80	0.60
Random Allocation	3.9	-82	0.55

Finally, to further analyze the impact of mobility, we evaluate system performance at different user speeds. Figure 5.5 illustrates the variation in fairness (Jain’s Index) with increasing user velocity. The results confirm that the mobility-aware NSGA-II dynamically adapts to user movement, maintaining higher fairness levels compared to static and greedy allocation methods. As mobility increases, fairness decreases across all strategies due to frequent handovers and dynamic spectral conflicts. However, the mobility-aware NSGA-II effectively mitigates fairness degradation, ensuring a more balanced resource allocation even in highly mobile scenarios.

5.6.5 Power Allocation vs. Mobility Speed

To analyze the impact of mobility on transmission power, we evaluate the average power allocation for different mobility speeds. Figure 5.6 illustrates how power allocation varies across different resource allocation strategies.

The results indicate that the mobility-aware NSGA-II optimizes power allocation dynamically, maintaining lower average transmission power compared to static and greedy approaches. At low mobility speeds, power allocation remains stable across all methods. However, as mobility increases, static and greedy strategies allocate excessive power due to their inability to adapt to changing network conditions, leading to inefficient energy usage and higher interference levels. In contrast, the mobility-aware NSGA-II dynamically

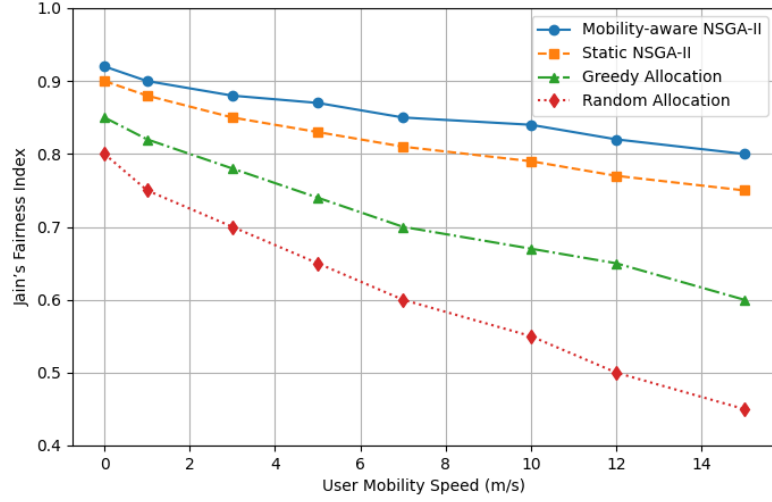


Figure 5.5: Impact of user mobility on fairness (Jain's Index).

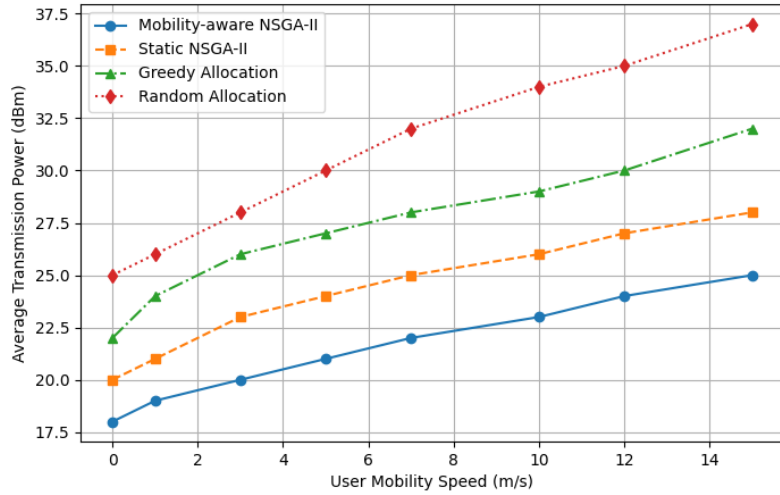


Figure 5.6: Impact of user mobility on power allocation.

adjusts power levels, minimizing unnecessary power expenditure while ensuring QoS for both cellular and D2D users.

Figure 5.7 presents the Pareto front between system throughput and interference, where each point represents a Pareto-optimal operating solution. The mobility-aware NSGA-II exhibits a smoother trend compared to static, greedy, and random methods because it continuously adapts resource allocation based on updated user positions and channel conditions. In contrast, baseline methods rely on fixed or instantaneous decisions, which leads to abrupt changes in allocation when network conditions vary. As a result, their curves appear less smooth, while the proposed method maintains more stable and consistent transitions across different operating points. Overall, it achieves a better throughput–interference trade-off than the baseline approaches. Figure 5.8 further illustrates the

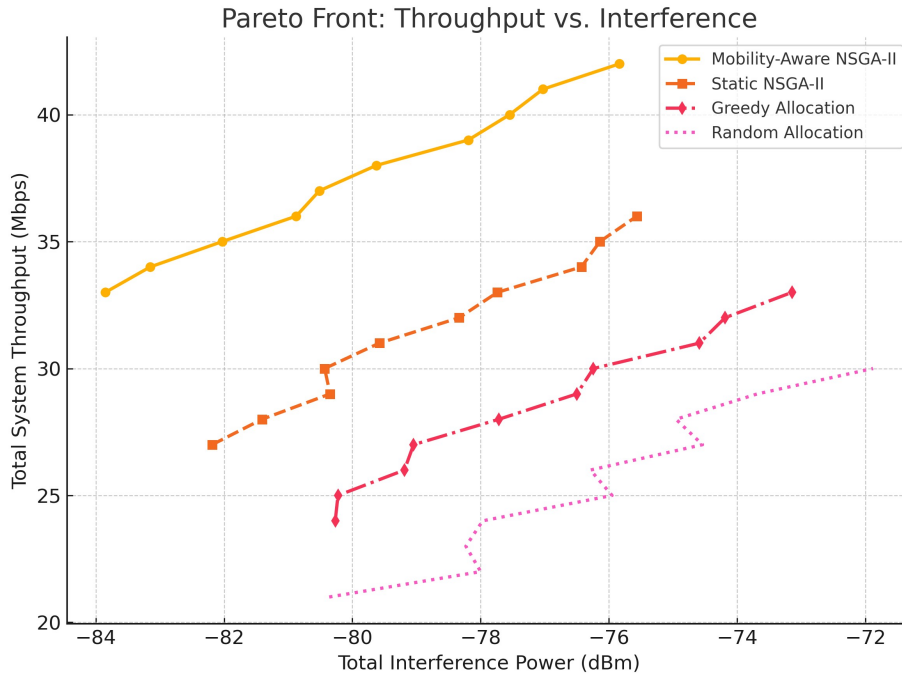


Figure 5.7: Pareto Front: Throughput vs. Interference for different allocation strategies.

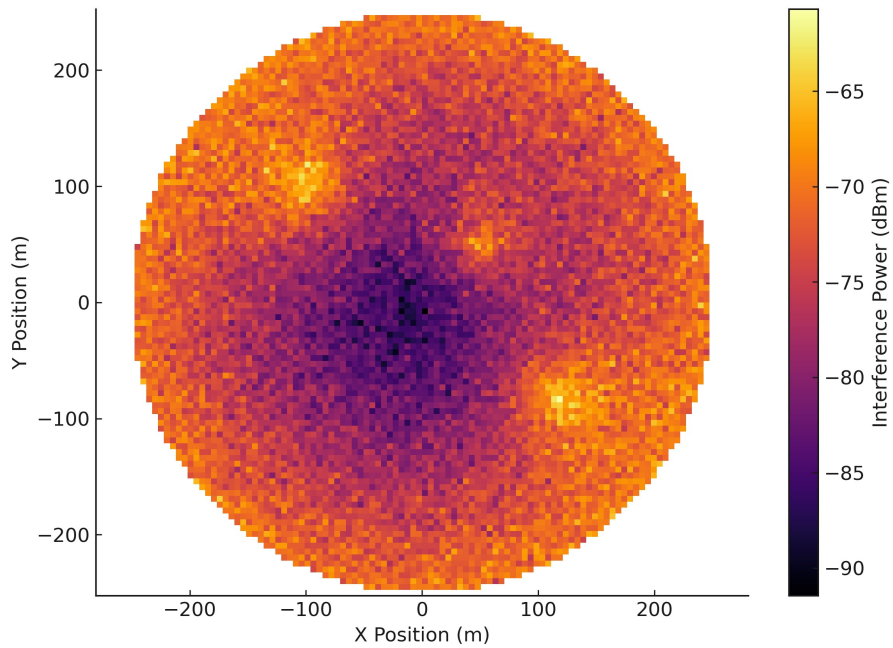


Figure 5.8: Spatial distribution of interference power across the cellular coverage area. Interference is elevated near the cell edge and around dense user clusters (hotspots), highlighting the importance of spatially-aware resource allocation.

spatial distribution of interference, highlighting elevated levels near cell edges and user

clusters. The proposed approach effectively suppresses such hotspots through adaptive, location-aware resource allocation.

5.7 Conclusion

This chapter presented a mobility-aware resource allocation framework for D2D-enabled cellular networks based on NSGA-II and Markovian mobility modeling. The proposed approach jointly optimized channel assignment and power allocation while accounting for dynamic user movement, interference variation, and fairness requirements. The obtained results showed that incorporating mobility awareness significantly improves throughput, fairness, and interference management compared with static, greedy, and random allocation schemes. Therefore, the proposed framework provides an effective and scalable solution for adaptive resource allocation in dynamic wireless environments and offers a strong basis for future extensions toward more intelligent real-time optimization methods.

Chapter 6

Summary of Results and Future Work

6.1 Summary of Results

This dissertation addressed key challenges in resource allocation for 5G and beyond wireless systems, focusing on power control, channel assignment, energy efficiency, and mobility-aware optimization. Each chapter introduced a novel solution tailored to a distinct scenario in next-generation communication networks.

Chapter 2 introduced a hybrid optimization framework for Non-Orthogonal Multiple Access (NOMA) in OFDMA systems. To tackle the non-convex and NP-hard nature of the joint channel and power allocation problem, the work proposed a two-stage approach. A novel Channel User Sorting and Filling (CUSF) algorithm was developed to handle channel allocation, followed by a power control mechanism integrating water-filling and Fractional Transmit Power Control (FTPC). This decomposition method significantly improved spectral efficiency and reduced computational burden, setting the foundation for advanced NOMA-based access schemes.

Building on this, Chapter 3 turned attention to energy-efficient computation offloading in Mobile Edge Computing (MEC). A comprehensive optimization framework was introduced to minimize energy consumption for both local and remote processing. The study incorporated two distinct access technologies—NOMA and massive MIMO (mMIMO)—under dynamic user and network conditions. For the NOMA case, a sub-channel and power allocation mechanism was developed, while in the mMIMO scenario, an uplink power control problem was formulated and solved using convex optimization techniques. The results demonstrated substantial energy savings and provided insights into trade-offs involving delay, throughput, and energy usage.

Chapter 4 tackled the issue of interference management in D2D-enabled cellular networks. A graph-theoretic approach was adopted to model channel reuse as an Integer Linear Programming (ILP) problem. Due to the complexity of the optimization, the work introduced a heuristic strategy involving K-means user clustering followed by graph coloring to assign channels efficiently within each cluster. A fairness-aware power allocation scheme complemented the channel assignment, ensuring a balanced SINR distribution. Simulation results confirmed the effectiveness of the method in minimizing cross-tier interference and enhancing network performance.

Finally, Chapter 5 introduced a machine learning-based optimization framework for dynamic D2D communication under mobility. Combining mobility prediction with a multi-objective Non-Dominated Sorting Genetic Algorithm II (NSGA-II), the proposed system jointly optimized channel allocation and power control with fairness constraints. Com-

pared to static and greedy baselines, the mobility-aware solution consistently achieved higher throughput, lower interference, and improved fairness. The analysis also revealed the sensitivity of SINR and power efficiency to user mobility, underscoring the importance of predictive, adaptive algorithms in modern wireless systems.

Collectively, the contributions of this dissertation offer a cohesive and forward-looking set of strategies for intelligent resource allocation in future wireless networks. They combine mathematical rigor, algorithmic innovation, and practical relevance, forming a solid basis for scalable and energy-aware network design.

6.2 Future Work

While each chapter in this dissertation presented targeted advancements in resource allocation, several promising directions for future research emerge from the findings in Chapters 2–5.

In Chapter 2, the CUSF algorithm and FTPC-based power control demonstrated effective performance in NOMA-OFDMA settings. Future work could extend this solution to alternative multi-user transmission paradigms such as Multi-User Linear Precoding (MU-LP) or Rate-Splitting Multiple Access (RSMA), thereby increasing the generality of the proposed approach. Additionally, incorporating user mobility or imperfect CSI into the model could provide more robust resource allocation under practical deployment conditions.

Chapter 3 focused on uplink energy optimization in MEC scenarios. A natural extension is to develop a unified framework that includes both uplink and downlink energy models, capturing the full offloading cycle and return transmission of processed data. Another avenue involves incorporating stochastic user demands and real-time service arrivals to further enhance the realism of the model. Benchmarking the proposed algorithms against reinforcement learning or other AI-based solutions will also help assess their relative competitiveness.

Chapter 4 introduced heuristic methods for solving an ILP-based channel reuse problem in D2D networks. While graph coloring and clustering proved effective, more adaptive schemes such as evolutionary algorithms or reinforcement learning-based graph embeddings could enhance channel selection in real-time. Moreover, extending the framework to heterogeneous networks (HetNets) with varying transmission ranges and QoS demands would broaden its applicability.

In Chapter 5, a mobility-aware optimization framework was developed using NSGA-II. The integration of predictive modeling and evolutionary optimization offers a powerful basis for further work. One promising direction is to incorporate deep reinforcement learning (DRL) to enable real-time decision-making under dynamic conditions. Federated learning may also be employed to maintain decentralized intelligence across edge devices while preserving data privacy. Finally, deploying the proposed framework in large-scale testbeds or simulated 5G environments will be key to evaluating its scalability, latency tolerance, and system-level trade-offs.

Overall, this dissertation contributes significantly to the field of intelligent and energy-efficient wireless communication. By combining optimization theory, graph algorithms, and machine learning, it lays the groundwork for future work at the intersection of network intelligence, edge computing, and mobility-aware connectivity.

List of Publications

1. Journals

- J1. **Qusay Alghazali**, Husam Al-Amaireh, Tibor Cinkler, "Joint Power and Channel Allocation for Non-Orthogonal Multiple Access in 5G Networks and Beyond", *Sensors*, Vol. 23, No. 19, 2023, Article 8040. DOI: 10.3390/s23198040. (WoS, Impact Factor: 3.4, Scopus, CiteScore: 7.3, **Q1**).
- J2. **Qusay Alghazali**, Husam Al-Amaireh, Tibor Cinkler, "Energy-Efficient Resource Allocation in Mobile Edge Computing Using NOMA and Massive MIMO", *IEEE Access*, Vol. 13, 2025, pp. 21456–21470. DOI: 10.1109/ACCESS.2025.3535233. (WoS, Impact Factor: 3.4, Scopus, CiteScore: 9.8, **Q1**).
- J3. **Qusay Alghazali**, Husam Al-Amaireh, Tibor Cinkler, "Mobility-Aware Resource Allocation in D2D Communications Using Genetic Algorithms," *IEEE Access*, Vol. 13, 2025, pp. 144591–144606. DOI: 10.1109/ACCESS.2025.3599051. (WoS, Impact Factor: 3.4; Scopus, CiteScore: 9.8, **Q1**).

Additional Journals

- J1. **Qusay Alghazali**, Abdulbasit M. A. Sabaawi, Husam Al-Amaireh, Mohammed R. Almasaoodi, and Tibor Cinkler, "Quantum Genetic Algorithm-Based Joint User Grouping and Power Control for Energy Efficiency Optimization in Multi-Cell Massive MIMO Uplink Systems," *To be submitted to Springer Journal of Quantum Information Processing*.
- J2. **Qusay Alghazali**, Husam Al-Amaireh, and Tibor Cinkler, "Beamforming-Aware Joint Resource and Power Allocation for Energy-Efficient Task Offloading in Massive MIMO MEC Networks," *To be submitted to IEEE open journal of the communications society*.

2. Conferences

- C1. **Qusay Alghazali**, Husam Al-Amaireh, Tibor Cinkler, "Graph Coloring and User Clustering-Based Resource Allocation for Device-to-Device Communication in 5G Networks", *2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Paris, France, 3–6 July 2025. Accepted and presented, waiting for publication to IEEE Xplore.
- C2. Viktória Nemkin, **Qusay Alghazali**, Tibor Cinkler, Katalin Friedl, László Kabódi, "Experiments with QUBO on D-Wave", Technical Report, Budapest University of Technology and Economics (BME), 2023. Available at: <https://www.bme.hu>

Bibliography

- [1] 3GPP. 3gpp tr 36.843: Study on lte device-to-device proximity services. Technical Report TR 36.843, 3rd Generation Partnership Project, 2014. URL <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2655>.
- [2] Pouya Akhoundzadeh and Ghasem Mirjalily. Joint antenna selection and resource allocation for mm-wave directional d2d communications using distributed deep reinforcement learning. *Electronics Letters*, 60(20), October 2024. DOI: [10.1049/e112.70066](https://doi.org/10.1049/e112.70066). URL <https://doi.org/10.1049/e112.70066>.
- [3] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli. Uplink non-orthogonal multiple access for 5g wireless networks. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 781–785, 2014.
- [4] Q. Alghazali and et al. Energy-efficient resource allocation in mobile edge computing using noma and massive mimo. *IEEE Access*, 13:21456–21470, 2025. URL <https://doi.org/10.1109/ACCESS.2025.3535233>.
- [5] Qusay Alghazali, Husam Al-Amaireh, and Tibor Cinkler. Joint power and channel allocation for non-orthogonal multiple access in 5g networks and beyond. *Sensors*, 23(19), 2023. ISSN 1424-8220. DOI: [10.3390/s23198040](https://doi.org/10.3390/s23198040). URL <https://www.mdpi.com/1424-8220/23/19/8040>.
- [6] A. M. H. Alibraheemi, M. N. Hindia, and K. Dimiyati. A survey of resource management in d2d communication for b5g networks. *IEEE Access*, 11:7892–7923, 2023. DOI: [10.1109/ACCESS.2023.3240655](https://doi.org/10.1109/ACCESS.2023.3240655).
- [7] Next G Alliance. Next g alliance: 6g research and roadmap, 2024. URL <https://nextgalliance.org>. Accessed: July 29, 2025.
- [8] A. Asadi and Q. Wang. A survey on device-to-device communication in cellular networks. *IEEE Communications Surveys & Tutorials*, 16:1801–1819, 2014.
- [9] M. W. Baidas. Resource allocation for offloading-efficiency maximization in clustered noma-enabled mobile edge computing networks. *Computer Networks*, 189:107919, 2021.
- [10] M. Bashar, K. Cumanan, A.G. Burr, H.Q. Ngo, L. Hanzo, and P. Xiao. Noma/oma mode selection-based cell-free massive mimo. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2019.
- [11] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura. Concept and practical considerations of non-orthogonal multiple access (noma) for future radio access. In *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 770–774. IEEE, 2013.

- [12] G. Bi and Y. Zeng. *Transforms and fast algorithms for signal analysis and representations*. Springer Science and Business Media, 2003.
- [13] E. Bjornson, J. Hoydis, and L. Sanguinetti. Massive mimo: Ten myths and one critical question. *IEEE Communications Magazine*, 54(2):114–123, 2017.
- [14] E. Björnson, J. Hoydis, and L. Sanguinetti. Massive mimo networks: Spectral, energy, and hardware efficiency. *Foundations and Trends® in Signal Processing*, 11(3-4):154–655, 2017.
- [15] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [16] A. Celik, R.M. Radaydeh, F.S. Al-Qahtani, A.H.A. El-Malek, and M.S. Alouini. Resource allocation and cluster formation for imperfect noma in dl/ul decoupled hetnets. In *2017 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2017.
- [17] A. Celik, M.C. Tsai, R.M. Radaydeh, F.S. Al-Qahtani, and M.S. Alouini. Distributed cluster formation and power-bandwidth allocation for imperfect noma in dl-hetnets. *IEEE Transactions on Communications*, 67(2):1677–1692, 2018.
- [18] A. Celik, M.C. Tsai, R.M. Radaydeh, F.S. Al-Qahtani, and M.S. Alouini. Distributed user clustering and resource allocation for imperfect noma in heterogeneous networks. *IEEE Transactions on Communications*, 67(10):7211–7227, 2019.
- [19] S. Chauhan, A. Gupta, and N. S. Nandakumar. Efficient resource allocation algorithms for multihop d2d approach in 5g network. *International Journal of Communication Systems*, 36(4):e5520, 2023.
- [20] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis. Performance optimization in mobile-edge computing via deep reinforcement learning. In *Proc. IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–6, 2018.
- [21] M. Chiang and T. Zhang. Fog and iot: An overview of research opportunities. *IEEE Internet of Things Journal*, 3(6):854–864, 2016.
- [22] D. Chizhik, J. Ling, P.W. Wolniansky, R.A. Valenzuela, N. Costa, and K. Huber. Multiple-input-multiple-output measurements and modeling in manhattan. *IEEE Journal on Selected Areas in Communications*, 21(3):321–331, 2003.
- [23] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, and G. Caire. Is noma efficient in multi-antenna networks? a critical look at next generation multiple access techniques. *IEEE Open Journal of the Communications Society*, 2:1310–1343, 2021. DOI: [10.1109/OJCOMS.2021.3092537](https://doi.org/10.1109/OJCOMS.2021.3092537).
- [24] L. Dai, B. Wang, Y. Yuan, and S. Han. A survey of non-orthogonal multiple access for 5g. *IEEE Communications Surveys & Tutorials*, 20(3):2294–2323, 2018.
- [25] C. de Alwis, Q.-V. Pham, and M. Liyanage. *Evolution of mobile networks*, 2023.
- [26] Z. Ding, H.V. Poor, and R. Schober. A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends. *IEEE Journal on Selected Areas in Communications*, 35(10):2181–2195, 2017.
- [27] F. Fang, J. Cheng, and Z. Ding. Joint energy efficient subchannel and power optimization for a downlink noma heterogeneous network. *IEEE Transactions on Vehicular Technology*, 68(2):1351–1364, 2018.

- [28] D. Feng, J. Cheng, and Q. Wang. Device-to-device communications in 6g: Emerging techniques and applications. *IEEE Wireless Communications*, 27(4):133–141, 2020.
- [29] 6G Flagship. 6g research visions and white papers, 2024. URL <https://www.6gflagship.com>. Accessed: July 29, 2025.
- [30] G. Fodor, E. Dahlman, G. Mildh, and S. Parkvall. Design aspects of network assisted device-to-device communications. *IEEE Communications Magazine*, 50(3):170–177, 2012.
- [31] J. Gandhi and Z. Narmawala. A comprehensive survey on machine learning techniques in opportunistic networks: Advances, challenges and future directions. *Pervasive and Mobile Computing*, 100, 2024. URL <https://doi.org/10.1016/j.pmcj.2024.101917>.
- [32] D. Gao, H. Cheng, Z. Han, and S. Yang. Resource optimization for the multi-user mimo systems assisted edge cloud computing. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pages 948–953. IEEE, October 2021.
- [33] A. Ghosh and et al. An efficient framework for clustered federated learning. *IEEE Trans. Inf. Theory*, 68(12):8076–8091, 2022. URL <https://doi.org/10.1109/TIT.2022.3192506>.
- [34] F. Giust and T. Taleb. Green 5g networks: Challenges and future research directions. *Elsevier Computer Networks*, 255:108076, 2023.
- [35] Qi Guo, Fengxiao Tang, and Nei Kato. Federated reinforcement learning-based resource allocation in d2d-enabled 6g. *IEEE Network*, 37(5):89–95, 2023. DOI: [10.1109/MNET.122.2200102](https://doi.org/10.1109/MNET.122.2200102).
- [36] M. Hoshino, T. Yoshida, and D. Imamura. Further advancements for e-utra physical layer aspects (release 9), 2010. *IEICE Transactions on Communications*, 94:3346–3353, 2011.
- [37] J. Hou, D. Yao, F. Wu, J. Shen, and X. Chao. Online vehicle velocity prediction using an adaptive radial basis function neural network. *IEEE Transactions on Vehicular Technology*, 70(4):3113–3122, 2021. URL <https://doi.org/10.1109/TVT.2021.3063483>.
- [38] I. Ioannou, C. Christophorou, V. Vassiliou, and A. Pitsillides. A novel distributed ai framework with ml for d2d communication in 5g/6g networks. *Computer Networks*, 211, 2022. DOI: [10.1016/j.comnet.2022.108987](https://doi.org/10.1016/j.comnet.2022.108987).
- [39] SMR Islam, M. Zeng, OA Dobre, and KS Kwak. Non-orthogonal multiple access (noma): How it meets 5g and beyond. *IEEE Communications Surveys & Tutorials*, 19(2):721–742, 2017.
- [40] S.R. Islam, N. Avazov, O.A. Dobre, and K.S. Kwak. Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges. *IEEE Communications Surveys and Tutorials*, 19(2):721–742, 2016.
- [41] ITU Radiocommunication Sector. Imt-2030 (6g) vision, 2024. URL <https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx>. Accessed: July 29, 2025.

- [42] T. Jabeen, Z. Ali, W. U. Khan, F. Jameel, I. Khan, G. A. S. Sidhu, and B. J. Choi. Joint power allocation and link selection for multi-carrier buffer aided relay network. *Electronics*, 8(6):686, 2019.
- [43] S. Jayakumar and S. Nandakumar. Reinforcement learning based distributed resource allocation technique in device-to-device (d2d) communication. *Wireless Networks*, 29:503–519, 2023. DOI: [10.1007/s11276-023-03230-x](https://doi.org/10.1007/s11276-023-03230-x).
- [44] Steffi Jayakumar and S. Nandakumar. Distributed resource optimisation using the q-learning algorithm, in device-to-device communication: A reinforcement learning paradigm. *Results in Engineering*, 23:102462, 2024. ISSN 2590-1230. DOI: <https://doi.org/10.1016/j.rineng.2024.102462>. URL <https://www.sciencedirect.com/science/article/pii/S2590123024007175>.
- [45] Chunchun Jia, Hongwen He, Jiaming Zhou, Jianwei Li, Zhongbao Wei, and Kunang Li. Learning-based model predictive energy management for fuel cell hybrid electric bus with health-aware control. *Applied Energy*, 355:122228, 2024. ISSN 0306-2619. DOI: <https://doi.org/10.1016/j.apenergy.2023.122228>. URL <https://www.sciencedirect.com/science/article/pii/S0306261923015921>.
- [46] S. Jiang and et al. Low-overhead clustered federated learning for personalized stress monitoring. *IEEE Internet Things J.*, 11(3):4335–4347, 2024. URL <https://doi.org/10.1109/JIOT.2023.3299736>.
- [47] P. Kannan. Artificial neural networks for voltage-frequency prediction using on-die measurements, 2023. Unpublished or in progress.
- [48] H.H. Kha, H. D. Tuan, and Ha H. Nguyen. Fast global optimal power allocation in wireless networks by local d.c. programming. *IEEE Transactions on Wireless Communications*, 11(2):510–515, 2012. DOI: [10.1109/TWC.2011.120911.110139](https://doi.org/10.1109/TWC.2011.120911.110139).
- [49] C. Lee, S.-M. Oh, and A.-S. Park. Interference avoidance resource allocation for d2d communication based on graph-coloring. In *2014 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 895–896, 2014.
- [50] C. Li, Q. Zhang, Q. Li, and J. Qin. Price-based power allocation for non-orthogonal multiple access systems. *IEEE Wireless Communications Letters*, 5(6):664–667, 2016.
- [51] H. Lim and T. Hwang. Energy-efficient beamforming and resource allocation for multi-antenna mec systems. *IEEE Access*, 10:18008–18022, 2022.
- [52] Xinyou Lin, Guangji Zhang, and Shenshen Wei. Velocity prediction using markov chain combined with driving pattern recognition and applied to dual-motor electric vehicle energy consumption evaluation. *Applied Soft Computing*, 101:106998, 2021. ISSN 1568-4946. DOI: <https://doi.org/10.1016/j.asoc.2020.106998>. URL <https://www.sciencedirect.com/science/article/pii/S1568494620309376>.
- [53] J. Liu, S. Guo, Q. Wang, C. Pan, and L. Yang. Optimal multi-user offloading with resources allocation in mobile edge cloud computing. *Computer Networks*, 221:109522, 2023.
- [54] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi. Multiobjective optimization for computation offloading in fog computing. *IEEE Internet of Things Journal*, 5(1):283–294, 2017.

- [55] Y. Liu, H. Zhang, K. Long, A. Nallanathan, and V.C. Leung. Energy-efficient subchannel matching and power allocation in noma autonomous driving vehicular networks. *IEEE Wireless Communications*, 26(4):88–93, 2019.
- [56] Y.F. Liu and Y.H. Dai. On the complexity of joint subcarrier and power allocation for multi-user ofdma systems. *IEEE Transactions on Signal Processing*, 62(3):583–596, 2013.
- [57] Yuanwei Liu, Zhijin Qin, Maged ElKashlan, Zhiguo Ding, Arumugam Nallanathan, and Lajos Hanzo. Non-orthogonal multiple access for 5g and beyond, 2018. URL <https://arxiv.org/abs/1808.00277>.
- [58] J. Logeshwaran, N. Shanmugasundaram, and J. Lloret. Energy-efficient resource allocation model for device-to-device communication in 5g wireless personal area networks. *Int. J. Commun. Syst.*, 36(4), 2023. DOI: [10.1002/dac.5524](https://doi.org/10.1002/dac.5524).
- [59] J.R. Lorch and A.J. Smith. Improving dynamic voltage scaling algorithms with pace. *ACM SIGMETRICS Performance Evaluation Review*, 29(1):50–61, 2001.
- [60] L. Ma, S. Zhou, G. Qiao, S. Liu, and F. Zhou. Superposition coding for downlink underwater acoustic ofdm. *IEEE Journal of Oceanic Engineering*, 42(1):175–187, 2016.
- [61] Q. Ma and et al. Feduc: A unified clustering approach for hierarchical federated learning. *IEEE Trans. Mobile Comput.*, 2024. URL <https://doi.org/10.1109/TMC.2024.3366947>.
- [62] S. Malini and P. Karthigaikumar. A weighted aimd congestion control algorithm for device-to-device communication in 5g network. *IETE Journal of Research*, 2024. URL <https://www.tandfonline.com/doi/abs/10.1080/03772063.2024.2400257>.
- [63] Y. Mao and J. Zhang. Mobile edge computing: Survey and research outlook. *ACM Computing Surveys*, 49:1–36, 2017.
- [64] Thomas L Marzetta, Erik G Larsson, Hong Yang, and Hien Quoc Ngo. *Fundamentals of massive MIMO*. Cambridge University Press, 2016.
- [65] TL Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):3590–3600, 2010.
- [66] TL Marzetta, H. Yang, HH Ngo, and EG Larsson. *Fundamentals of massive MIMO*. Cambridge University Press, 2016.
- [67] M.R. Mili, K.A. Hamdi, F. Marvasti, and M. Bennis. Joint optimization for optimal power allocation in ofdma femtocell networks. *IEEE Communications Letters*, 20(1):133–136, 2015.
- [68] D.W.K. Ng, E.S. Lo, and R. Schober. Energy-efficient resource allocation in ofdma systems with large numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 11(9):3292–3304, 2012.
- [69] HH Ngo, EG Larsson, and TL Marzetta. Energy and spectral efficiency of very large multiuser mimo systems. *IEEE Transactions on Communications*, 61(4):1436–1449, 2013.

- [70] H.V. Nguyen, V.D. Nguyen, O.A. Dobre, D.N. Nguyen, E. Dutkiewicz, and O.S. Shin. Joint power control and user association for noma-based full-duplex systems. *IEEE Transactions on Communications*, 67(11):8037–8055, 2019.
- [71] Z. Ning, P. Dong, X. Kong, and F. Xia. A cooperative partial computation offloading scheme for mobile edge computing enabled internet of things. *IEEE Internet of Things Journal*, 6(3):5439–5451, Jun. 2019.
- [72] W. Pan and et al. Time-sensitive federated learning with heterogeneous training intensity: A deep reinforcement learning approach. *IEEE Trans. Emerg. Topics Comput. Intell.*, 8(2):1402–1415, 2024. URL <https://doi.org/10.1109/TETCI.2023.3345366>.
- [73] J. Papandriopoulos and J. S. Evans. Scale: A low-complexity distributed protocol for spectrum balancing in multiuser dsl networks. *IEEE Transactions on Information Theory*, 55(8):3711–3724, Aug. 2009.
- [74] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Transactions on Vehicular Technology*, 64(11):5275–5287, 2014.
- [75] H. Rajab, F. Benkhelifa, and T. Cinkler. Analysis of power allocation for noma-based d2d communications using gadia. *Information*, 12(12):1–15, 2021.
- [76] H. Rajab, F. Benkhelifa, and T. Cinkler. Analysis of power allocation for noma-based d2d communications using gadia. *Information*, 12(12), 2021. URL <https://doi.org/10.3390/info12120510>.
- [77] S. Razavizadeh and M. Hamdi. Resource allocation in d2d networks underlying lte-advanced: Challenges and solutions. *IEEE Wireless Communications*, 23:112–118, 2016.
- [78] A.E. Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy*, 92(6):991–1016, 1984.
- [79] Won Jae Ryu and Soo Young Shin. Power allocation for urllc using finite blocklength regime in downlink noma systems. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 770–773. IEEE, 2019.
- [80] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura. System-level performance evaluation of downlink non-orthogonal multiple access (noma). In *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 611–615. IEEE, 2013.
- [81] F. Sattler and et al. Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(8):3710–3722, 2021. URL <https://doi.org/10.1109/TNNLS.2020.3015958>.
- [82] M. Shafi and A.F. Molisch. 5g: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, 35: 1201–1221, 2017.
- [83] L. Shen, K.-T. Feng, and L. Hanzo. Five facets of 6g: Research challenges and opportunities. *ACM Computing Surveys*, 55(11):1–39, 2023.

- [84] P. Sindhu, K.S. Deepak, and A.H. KM. A novel low complexity power allocation algorithm for downlink noma networks. In *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 36–40. IEEE, 2018.
- [85] Durgesh Singh and Sasthi C. Ghosh. Mobility-aware relay selection in 5g d2d communication using stochastic model. *IEEE Transactions on Vehicular Technology*, 68(3):2837–2849, 2019. DOI: [10.1109/TVT.2019.2893995](https://doi.org/10.1109/TVT.2019.2893995).
- [86] H. Su and X. Zhang. Cross-layer based opportunistic mac protocols for qos provisionings over cognitive radio wireless networks. *IEEE Journal on Selected Areas in Communications*, 26(1):118–129, 2008.
- [87] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan. On the ergodic capacity of mimo noma systems. *IEEE Wireless Communications Letters*, 4(4):405–408, 2015.
- [88] Mohsen Nader Tehrani, Murat Uysal, and Halim Yanikomeroglu. Device-to-device communication in 5g cellular networks: challenges, solutions, and future directions. *IEEE Communications Magazine*, 52(5):86–92, 2014. DOI: [10.1109/MCOM.2014.6815897](https://doi.org/10.1109/MCOM.2014.6815897).
- [89] S. Timotheou and I. Krikidis. Fairness for non-orthogonal multiple access in 5g systems. *IEEE Signal Processing Letters*, 22(10):1647–1651, 2015.
- [90] T.X. Tran and D. Pompili. Joint task offloading and resource allocation for multi-server mobile-edge computing networks. *IEEE Transactions on Vehicular Technology*, 68(1):856–868, Jan. 2018.
- [91] Amara Umar, Syed Ali Hassan, Haejoon Jung, Sahil Garg, M. Shamim Hossain, and Mohsen Guizani. Computation offloading in noma-mec-enabled aerial-vehicular networks exploiting mmwave capabilities. *Computer Networks*, 246, June 2024. ISSN 1389-1286. DOI: [10.1016/j.comnet.2024.110335](https://doi.org/10.1016/j.comnet.2024.110335). Publisher Copyright: © 2024 Elsevier B.V.
- [92] Z. Wan, D. Xu, and I. Ahmad. Joint computation offloading and resource allocation for noma-based multi-access mobile edge computing systems. *Computer Networks*, 196:108256, 2021.
- [93] C.L. Wang, J.Y. Chen, and Y.J. Chen. Power allocation for a downlink non-orthogonal multiple access system. *IEEE Wireless Communications Letters*, 5(5):532–535, 2016.
- [94] F. Wang, J. Xu, and Z. Ding. Multi-antenna noma for computation offloading in multiuser mobile edge computing systems. *IEEE Transactions on Communications*, 67(3):2450–2463, 2018.
- [95] L. Wang, X. Chen, and H. Zhang. Mobility-aware task offloading and resource allocation in mec systems. *IEEE Transactions on Vehicular Technology*, 69(8):8844–8858, 2020.
- [96] N. Wang, E. Hossain, and V.K. Bhargava. Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier hetnets with large-scale antenna arrays. *IEEE Transactions on Wireless Communications*, 15(5):3251–3268, 2016.
- [97] Y. Wang, J. Liu, and X. Chen. Integrated resource allocation for noma-based mec systems: A survey. *IEEE Network*, 35(3):18–25, 2021.

- [98] Z. Wei, D.W.K. Ng, J. Yuan, and H.M. Wang. Optimal resource allocation for power-efficient mc-noma with imperfect channel state information. *IEEE Transactions on Communications*, 65(9):3944–3961, 2017.
- [99] X. Wu, X. Huang, S. Yuan, Z. Tang, and Y. Wang. A resource allocation method based on weighted drf algorithm in mobile edge computing. In *Proc. IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 5, pages 86–90, 2021.
- [100] R. Xie, F.R. Yu, H. Ji, and Y. Li. Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells. *IEEE Transactions on Wireless Communications*, 11(11):3910–3920, 2012.
- [101] B. Xu and et al. Clustered federated learning in internet of things: Convergence analysis and resource optimization. *IEEE Internet Things J.*, 11(2):3217–3232, 2024. URL <https://doi.org/10.1109/JIOT.2023.3298366>.
- [102] C. Xu, L. Li, and J. Yang. Energy-efficient resource allocation for device-to-device communication underlying cellular networks. *IEEE Transactions on Vehicular Technology*, 64(4):1561–1572, 2014.
- [103] Tongyang Xu, Yuan Liu, Zhaotai Ma, Yiqiang Huang, and Peng Liu. A dqn-based multi-objective participant selection for efficient federated learning. *Future Internet*, 15(6), 2023. ISSN 1999-5903. DOI: [10.3390/fi15060209](https://doi.org/10.3390/fi15060209). URL <https://www.mdpi.com/1999-5903/15/6/209>.
- [104] Wangyang Xu, Jiancheng An, Yongjun Xu, Chongwen Huang, Lu Gan, and Chau Yuen. Time-varying channel prediction for RIS-assisted MU-MISO networks via deep learning. *IEEE Transactions on Cognitive Communications and Networking*, 8(4):1802–1815, 2022. DOI: [10.1109/TCCN.2022.3188153](https://doi.org/10.1109/TCCN.2022.3188153).
- [105] B. Yang, H. Zhang, and L. Li. Joint beamforming design for non-orthogonal multiple access with massive mimo. *IEEE Transactions on Wireless Communications*, 19(5): 3167–3181, 2020.
- [106] C. You, K. Huang, H. Chae, and BH Kim. Energy-efficient resource allocation for mobile-edge computation offloading. *IEEE Transactions on Wireless Communications*, 16(3):1397–1411, 2016.
- [107] M. Yu and M. Zhang. An energy-saving joint resource allocation approach for mobile edge computing based on noma. *Physical Communication*, 63:102224, 2024.
- [108] X. Yuan, H. Yao, J. Wang, T. Mai, and M. Guizani. Artificial intelligence empowered qos-oriented network association for next-generation mobile networks. *IEEE Transactions on Cognitive Communications and Networking*, 7(3):856–870, 2021.
- [109] Xiaou Yuan, Hui Tian, and Bo Fan. Mobility-aware joint resource allocation and power allocation for d2d communication. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2019. DOI: [10.1109/WCNC.2019.8885989](https://doi.org/10.1109/WCNC.2019.8885989).
- [110] S.S. Yilmaz and B. Özbek. Massive mimo-noma based mec in task offloading for delay minimization. *IEEE Access*, 11:162–170, 2022.

- [111] M. Zeng, A. Yadav, O.A. Dobre, and H.V. Poor. Energy-efficient power allocation for hybrid multiple access systems. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–5. IEEE, 2018.
- [112] M. Zeng, W. Hao, O.A. Dobre, and H.V. Poor. Delay minimization for massive mimo assisted mobile edge computing. *IEEE Transactions on Vehicular Technology*, 69(6):6788–6792, 2020.
- [113] D. Zhai and R. Zhang. Joint admission control and resource allocation for multi-carrier uplink noma networks. *IEEE Wireless Communications Letters*, 7(6):922–925, 2018.
- [114] H. Zhang, C. Jiang, N.C. Beaulieu, X. Chu, X. Wen, and M. Tao. Resource allocation in spectrum-sharing ofdma femtocells with heterogeneous services. *IEEE Transactions on Communications*, 62(7):2366–2377, 2014.
- [115] H. Zhang, D.K. Zhang, W.X. Meng, and C. Li. User pairing algorithm with sic in non-orthogonal multiple access system. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.
- [116] H. Zhang, J. Liu, and X. Chen. Integration of noma with mec: Challenges and opportunities. *IEEE Transactions on Wireless Communications*, 20(6):3451–3463, 2021.
- [117] J. Zhang, L. Wang, J. Ma, and N. Ge. Hybrid precoding for millimeter wave massive mimo systems with partial channel knowledge. *IEEE Access*, 4:1095–1107, 2016.
- [118] J. Zhang, X. Hu, Z. Ning, E. C.-H. Ngai, L. Zhou, J. Wei, J. Cheng, and B. Hu. Energy-latency tradeoff for energy-aware offloading in mobile edge computing networks. *IEEE Internet of Things Journal*, 5(4):2633–2645, 2017.
- [119] Ran Zhang, Fangqi Liu, Jiang Liu, Mingzhe Chen, Qinqin Tang, Tao Huang, and F. Richard Yu. Cpper-fl: Clustered parallel training for efficient personalized federated learning. *IEEE Transactions on Mobile Computing*, 23(10):9424–9436, 2024. DOI: [10.1109/TMC.2024.3365951](https://doi.org/10.1109/TMC.2024.3365951).
- [120] Tinghao Zhang, Kwok-Yan Lam, and Jun Zhao. Device scheduling and assignment in hierarchical federated learning for internet of things. *IEEE Internet of Things Journal*, 11(10):18449–18462, 2024. DOI: [10.1109/JIOT.2024.3362972](https://doi.org/10.1109/JIOT.2024.3362972).
- [121] X. Zhang, Z. Huang, L. Huang, and H. Yang. Resource allocation strategy based on service function chaining in multi-access edge computing network. *Journal of King Saud University-Computer and Information Sciences*, page 102001, 2024.
- [122] Xinran Zhang, Hui Tian, Wanli Ni, Zhaohui Yang, and Mengying Sun. Deep reinforcement learning for energy efficiency maximization in swipt-based over-the-air federated learning. *IEEE Transactions on Green Communications and Networking*, 8(1):525–541, 2024. DOI: [10.1109/TGCN.2023.3307428](https://doi.org/10.1109/TGCN.2023.3307428).
- [123] Y. Zhang and C. Liu. Mobile edge computing in next-generation networks: Resource allocation and energy efficiency trade-offs. *Computer Networks*, 236:112145, 2023.
- [124] Y. Zhang, H.M. Wang, T.X. Zheng, and Q. Yang. Energy-efficient transmission design in non-orthogonal multiple access. *IEEE Transactions on Vehicular Technology*, 66(3):2852–2857, 2016.

- [125] Y. Zhang, C. Liu, and J. Song. Task offloading and resource allocation in d2d-assisted mec systems: A survey. *IEEE Communications Surveys & Tutorials*, 22(4): 2195–2219, 2020.
- [126] T. Zhao and et al. Drl-based joint resource allocation and device orchestration for hierarchical federated learning in noma-enabled industrial iot. *IEEE Trans. Ind. Informat.*, 19(6):7468–7479, 2023. URL <https://doi.org/10.1109/TII.2022.3170900>.
- [127] W. Zhao and S. Wang. Resource allocation for device-to-device communication underlying cellular networks: An alternating optimization method. *IEEE Communications Letters*, 19(8):1398–1401, 2015.
- [128] X. Zhou, G. Chen, Y. Hu, and X. Li. D2d interference management and resource allocation scheme based on improved graph coloring. In *IEEE Conference on Computer and Communications*, pages 184–187, 2022.
- [129] Josip Zidar, Tomislav Matić, Ivan Aleksi, and Željko Hocenski. Dynamic voltage and frequency scaling as a method for reducing energy consumption in ultra-low-power embedded systems. *Electronics*, 13(5), 2024. ISSN 2079-9292. DOI: [10.3390/electronics13050826](https://doi.org/10.3390/electronics13050826). URL <https://www.mdpi.com/2079-9292/13/5/826>.

Appendices A–D

Appendix A

Water-Filling Power Allocation Derivation

The power allocation problem using the water-filling algorithm can be solved using the Karush–Kuhn–Tucker (KKT) conditions, which are necessary for optimality in the power control problem P_2 .

The Lagrangian function is:

$$L(p, \lambda) = \sum_{n \in \mathcal{N}} \log_2(1 + p_n H_n) - \lambda \left(\sum_{n \in \mathcal{N}} p_n - P_T \right) \quad (\text{A.1})$$

where λ is the Lagrange multiplier and $H_n = \frac{g_n}{\sigma^2}$ denotes the effective channel gain-to-noise ratio for sub-channel $c_n \in \mathcal{N}$.

The KKT conditions are:

1. Stationarity:

$$\frac{\partial L}{\partial p_n} = \frac{H_n}{\ln(2)(1 + H_n p_n)} - \lambda = 0, \quad n = 1, 2, \dots, N \quad (\text{A.2})$$

2. Primal feasibility:

$$\sum_{n=1}^N p_n \leq P_T, \quad p_n \geq 0 \quad (\text{A.3})$$

3. Dual feasibility:

$$\lambda \geq 0 \quad (\text{A.4})$$

4. Complementary slackness:

$$\lambda \left(\sum_{n=1}^N p_n - P_T \right) = 0 \quad (\text{A.5})$$

From (A.1), solving the stationarity condition gives:

$$\frac{H_n}{\ln(2)(1 + H_n p_n)} = \lambda \quad (\text{A.6})$$

Solving for p_n :

$$p_n = \frac{1}{\lambda \ln 2} - \frac{1}{H_n} \quad (\text{A.7})$$

This completes the derivation used in Chapter 2 (see Equation (2.14)).

Appendix B

KKT Derivation for Problem P_2

The Lagrangian for the optimization problem P_2 is:

$$L(f, \lambda) = \sum_{w=1}^W F^c(w)[f(w)]^2 + \lambda \left(\sum_{w=1}^W \frac{1}{f(w)} - T^D \right) \quad (\text{B-1})$$

The KKT conditions are:

1. **Stationarity:**

$$2F^c(w)f(w) - \lambda \frac{1}{[f(w)]^2} = 0 \quad (\text{B-2})$$

2. **Primal feasibility:**

$$\sum_{w=1}^W \frac{1}{f(w)} - T^D \leq 0 \quad (\text{B-3})$$

3. **Dual feasibility:**

$$\lambda \geq 0 \quad (\text{B-4})$$

4. **Complementary slackness:**

$$\lambda \left(\sum_{w=1}^W \frac{1}{f(w)} - T^D \right) = 0 \quad (\text{B-5})$$

From (B-2):

$$f(w) = \left(\frac{\lambda}{2F^c(w)} \right)^{1/3} \quad (\text{B-6})$$

Substitute into (B-5):

$$\sum_{w=1}^W (F^c(w))^{1/3} = T^D \left(\frac{\lambda}{2} \right)^{1/3} \quad (\text{B-7})$$

Final expression for $f(w)$:

$$f(w) = \frac{\sum_{w=1}^W (F^c(w))^{1/3}}{T^D (F^c(w))^{1/3}} \quad (\text{B-8})$$

Appendix C

KKT-Based Convex Optimization for Problem P_5

The Lagrangian for Problem P_5 is:

$$\begin{aligned} \mathcal{L}(\hat{p}, \mu, \lambda) = & \sum_{k \in \mathcal{K}} e^{\hat{p}_k} - \sum_{k \in \mathcal{K}} \mu_k \left(\zeta \left[\hat{p}_k + \log_2(g_k) - \log_2 \left(\sum_{i=1}^{k-1} g_k e^{\hat{p}_i} + \sigma_n^2 \right) \right] \right. \\ & \left. + \gamma - \frac{D_k}{T_k^D - T_k^C} \right) + \sum_{k \in \mathcal{K}} \lambda_k \left(\sum_{n \in \mathcal{N}} e^{\hat{p}_k} - \hat{P}_k^{\max} \right) \end{aligned} \quad (\text{C-1})$$

Stationarity:

$$\frac{\partial \mathcal{L}}{\partial \hat{p}_k} = e^{\hat{p}_k} (\lambda_k + 1) - \mu_k \zeta = 0 \quad (\text{C-2})$$

$$\Rightarrow \hat{p}_k = \log \left(\frac{\mu_k \zeta}{\lambda_k + 1} \right) \quad (\text{C-3})$$

Other KKT Conditions:

- **Primal feasibility:** all original constraints must hold.
- **Dual feasibility:**

$$\mu_k \geq 0, \quad \lambda_k \geq 0 \quad \forall k \quad (\text{C-4})$$

- **Complementary slackness:**

$$\mu_k \left(\zeta \left[\hat{p}_k + \log_2(g_k) - \log_2 \left(\sum_{i=1}^{k-1} g_k e^{\hat{p}_i} + \sigma_n^2 \right) \right] + \gamma - \frac{D_k}{T_k^D - T_k^C} \right) = 0, \quad (\text{C-5})$$

$$\lambda_k \left(\sum_{n \in \mathcal{N}} e^{\hat{p}_k} - \hat{P}_k^{\max} \right) = 0, \quad \forall k \quad (\text{C-6})$$

Gradient-Based Updates:

$$\text{Grad}_{\lambda_k} = \sum_{n \in \mathcal{N}} e^{\hat{p}_k} - \hat{P}_k^{\max} \quad (\text{C-7})$$

$$\text{Grad}_{\mu_k} = \zeta \left[\hat{p}_k + \log_2(g_k) - \log_2 \left(\sum_{i=1}^{k-1} g_k e^{\hat{p}_i} + \sigma_n^2 \right) \right] + \gamma - \frac{D_k}{T_k^D - T_k^C} \quad (\text{C-8})$$

Update rules:

$$\lambda_k^{\text{new}} = \lambda_k^{\text{old}} - \eta_\lambda \cdot \text{Grad}_{\lambda_k} \quad (\text{C-9})$$

$$\mu_k^{\text{new}} = \mu_k^{\text{old}} + \eta_\mu \cdot \text{Grad}_{\mu_k} \quad (\text{C-10})$$

Repeat until convergence.

Appendix D

Proof of Theorem 1: Lyapunov Stability

Theorem 2 (Stability Guarantee). Under NSGA-II guided by the drift-plus-penalty framework in (5.33), and given:

- Bounded throughput: $T_{\max} = B_j \log_2\left(1 + \frac{P_{\max}G_{\max}}{N_0}\right)$
- Bounded interference: $I_{\max} = N_{\text{reuse}}P_{\max}G_{\max}$
- Minimum Lyapunov decay rate: $\epsilon = \min\left(\Delta\gamma_{\max}, \frac{1}{N_{\text{users}}}\right)$

the following hold:

1. Virtual queues remain bounded:

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[Q_u(\tau)] \leq \sqrt{\frac{2(B + V(I_{\max} + T_{\max}))}{\epsilon}}$$

2. SINR deviations are controlled:

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left[(\gamma_i(\tau) - \gamma_i^{\text{target}})^2\right] \leq \frac{2(B + VI_{\max})}{\epsilon}$$

where $B = \frac{1}{2}(N\Delta\gamma_{\max}^2 + N_{\text{users}}T_{\max}^2)$. •

Step 1: Lyapunov Drift Expansion

From the Lyapunov function $L(t)$ in (5.30):

$$\begin{aligned} \Delta(t) &= \mathbb{E}[L(t+1) - L(t) \mid \mathbf{x}(t)] \\ &\leq \mathbb{E}\left[\sum_{i=1}^N \Delta\gamma_{\max}(\gamma_i(t) - \gamma_i^{\text{target}}) \right. \\ &\quad \left. + \sum_{u=1}^{N_{\text{users}}} Q_u(t)(T_u^{\text{target}} - T_u(t)) \mid \mathbf{x}(t)\right] + B \end{aligned} \tag{D-1}$$

Step 2: Drift-Plus-Penalty Bound

Substituting the NSGA-II objective into (5.33):

$$\Delta(t) - VT_{\text{total}}(t) + VI_{\text{total}}(t) \leq -\epsilon\mathbb{E}[L(t)] + B + V(I_{\text{max}} + T_{\text{max}}) \quad (\text{D-2})$$

Step 3: Telescoping Sum

Summing over $t \in \{0, \dots, T-1\}$ and taking $T \rightarrow \infty$:

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[L(t)] \leq \frac{B + V(I_{\text{max}} + T_{\text{max}})}{\epsilon} \quad (\text{D-3})$$

Step 4: Final Bounds

From $L(t) \geq \frac{1}{2} \sum Q_u(t)^2$ and Jensen's inequality:

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[Q_u(t)] \leq \sqrt{\frac{2(B + V(I_{\text{max}} + T_{\text{max}}))}{\epsilon}} \quad (\text{D-4})$$

Similarly, the SINR bound follows from $L(t) \geq \frac{1}{2} \sum (\gamma_i(t) - \gamma_i^{\text{target}})^2$.