UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DESIGNING DATA-AIDED DEMAND-DRIVEN USER-CENTRIC ARCHITECTURE FOR 6G AND BEYOND NETWORKS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND COMPUTER ENGINEERING

BY

Shahrukh Khan Kasi
Norman, Oklahoma
2023

DESIGNING DATA-AIDED DEMAND-DRIVEN USER-CENTRIC ARCHITECTURE FOR 6G AND BEYOND NETWORKS

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Ali Imran, Chair

Dr. James J. Sluss, Jr.

Dr. Samuel Cheng

Dr. Timothy G. Ford

Acknowledgments

I am deeply grateful to Dr. Ali Imran, my supervisor, whose exceptional guidance and mentorship have been crucial in completing this dissertation. I will forever be indebted to Dr. Ali Imran for instilling in me the qualities of a diligent researcher and providing invaluable advice on excelling in both personal and professional aspects of life. I would also like to express my deepest appreciation to Dr. Sabit Ekin for his support and encouragement throughout this journey. His guidance has been instrumental in shaping my academic path. Additionally, I am grateful to Professor Subir Biswas for giving me the opportunity to benefit from his expertise.

I would also like to thank the committee members, Dr. James J. Sluss, Dr. Timothy G. Ford, and Dr. Samuel Cheng, for their valuable contributions. I would also like to express my sincere appreciation to Denise Davis and Krista Pettersen for their exceptional support and assistance in handling all administrative tasks during my study at OU-Tulsa.

I am also immensely grateful to my family, who have been with me through all the highs and lows of this challenging journey. The contributions of my parents, Ami and Baba, cannot be overstated. Their prayers and constant encouragement have played a pivotal role in my success. I specifically extend my heartfelt thanks to my wife, Hifza Afzal, for her support, patience, and belief in me. Her presence has been a consistent source of strength and motivation. I extend my sincere gratitude to my friends and colleagues at the AI4Networks Research Center. Their steadfast support and assistance have been invaluable during my academic journey.

Table of Contents

List of Figures

List of Tables

Abstract

Despite advancements in capacity-enhancing technologies like massive MIMO (multiple input, multiple output) and intelligent reflective surfaces, network densification remains crucial for significant capacity gains in future networks such as 6G. However, network densification increases interference and power consumption. Traditional cellular architectures struggle to minimize these without compromising service quality or capacity, which necessitates a shift to a user-centric radio access network (UC-RAN).

The UC-RAN approach offers additional degrees of freedom to ease the spectral-energy efficiency interlock while improving the service quality. However, its increased degrees of freedom make its optimal design and operation more challenging. This dissertation introduces four novel approaches for UC-RAN optimal design and operation. The objectives include mitigating interference, reducing power consumption, ensuring diverse user/vertical service quality, facilitating proactive network operation, risk-aware optimization, adopting an open radio access network, and enabling universal coverage.

First, we construct an analytical framework to assess the effects of incorporating Coordinated Multipoint (CoMP) technology into UC-RAN to reduce interference and power consumption. We use stochastic geometry tools to derive expressions for network-wide coverage, spectral efficiency, and energy efficiency as a function of UC-RAN Configuration and Optimization Parameters (COPs), including data base station densities and user-centric service zone sizes.

While the analytical framework provides insightful performance analysis that can guide overall system design, it cannot fully capture the dynamics of a UC-RAN system to enable optimal operation. Next, we present a Deep Reinforcement Learning (DRL) based method to dynamically orchestrate the UC-RAN service zone size to satisfy varying application demands of various service verticals during its operation. We define a novel multi-objective optimization problem that fairly optimizes otherwise conflicting key performance indicators (KPIs).

DRL's practical adaptation by the industry remains thwarted by the risk it poses to the safe operation of a live network. To address this challenge, we propose a digital twin-enabled approach to enrich the DRL-based optimization framework, ensuring risk-aware COP optimization. We use Open Radio Access Network standards-based simulations to show that the proposed risk-aware DRL framework can maximize system-level KPIs while maintaining safe operational requirements.

Lastly, we propose a hybrid model of aerial and terrestrial UC-RAN deployment to ensure universal coverage. We assess the impact of aerial base station parameters on system-level KPIs, providing a quantitative analysis of the advantages of a hybrid over a solely terrestrial UC-RAN. We develop a robust multi-objective function solvable via our DRL-based framework to balance and optimize these KPIs in a hybrid UC-RAN.

Our extensive analytical and system-level simulation results suggest that these contributions can foster the much-needed paradigm shift towards demand-driven, elastic, and user-centric architecture in emerging and future cellular networks.

CHAPTER 1

Introduction

1.1 Motivation

In the realm of 6G and beyond networks, ultra-dense networks (UDNs) will play a vital role in offering seamless coverage, remarkably high throughput, and unparalleled low latency, surpassing the benefits of additional spectrum or advancements in physical layer technologies like massive MIMO (multiple input, multiple output) or intelligent reflective surfaces. Network operators are actively exploring the potential of UDNs to meet the escalating demands for throughput and latency envisioned for 6G and beyond users.

While both academia and industry researchers agree that network densification is key to enhancing the coverage and capacity of existing cellular networks, it is essential to acknowledge the inherent complexities associated with UDNs [1]. Densifying the network reduces the average distance between users and the interferring base stations. This leads to increased interference from neighboring base stations, which overshadows the benefits of the decreased average distance from serving base stations. Furthermore, the deployment of a large number of small cells contributes to an increase in the network's power consumption.

Traditional cellular architectures encounter difficulties in addressing these issues without compromising service quality or capacity. Therefore, there is a need to shift the cellular network design towards a user-centric radio access network (UC-RAN) to overcome these challenges [2]. The UC-RAN approach offers additional degrees of freedom to relax the spectral-energy efficiency interlock while improving the service quality [2–4]. However, its increased flexibility makes its optimal design and operation more challenging, particularly when considering the following objectives:

Meeting Heterogeneous Quality-of-service (QoS) Requirements: A crucial aspect of 6G and beyond cellular networks is the need to accommodate highly diverse application requirements, including augmented/virtual reality, high-speed rails, industrial robots, and E-health, to name a few. Network operators are challenged with providing heterogeneous user QoS to effectively meet the unique requirements of various applications. Consequently, it becomes imperative to redesign the traditional one-size-fits-all user-centric cellular architecture to accommodate the heterogeneous nature of application needs which are expected to be demand-driven, elastic, and capable of supporting multiple services.

Facilitating Proactive Network Operation: The current reactive mode of operation in cellular networks, with approximately two thousand tunable configuration and optimization parameters (COPs) and numerous key performance indicators (KPIs) in 5G [5], often leads to sub-optimal performance gains and imposes significant operating expenditures on network operators. This complexity is expected to increase in future generations of cellular networks. To achieve proactive adaptability, agility, and intelligence in cellular network design, it is crucial to leverage artificial intelligence tools. By harnessing artificial intelligence tools, cellular networks can be controlled proactively, effectively catering to the diverse and evolving needs of user applications. This approach enables near-optimal performance by maximizing various KPIs and ensures responsiveness to dynamic network conditions. Furthermore, the proactive operation of the cellular network can result in a reduction of operating expenditures for network operators.

Performing Risk-aware Optimization: Despite the recent popularity of data-driven online learning frameworks, their practical deployment in cellular networks remains limited due to their risk of deteriorating network performance during the exploration phase. These data-driven online solutions require hit-and-trial on live networks before converging to optimal COPs. Unlike other domains where hit-and-trial is acceptable during the training phase, in cellular networks, online solutions can degrade network KPIs. To make data-driven online solutions practical for optimizing desired set of KPIs in complex environments, such as cellular networks, there is the need to make data-driven learning algorithms capable of averting the risk of choosing extreme exploratory COPs that can cause harm to a cellular system (such as low quality of experience at UEs, service unavailability, etc.) during online optimization.

Adopting Open Radio Access Network: The centralized architecture of traditional cellular networks presents challenges in terms of scalability and feasibility, particularly when combined with a data-driven automation solution. The introduction of the Open Radio Access Network (O-RAN) and its disaggregated architecture revolutionizes the radio access network (RAN) vendor ecosystem. It enables multiple vendors to supply commercial off-the-shelf hardware that aligns with open standards and interoperability specifications [6]. This shift empowers network operators to choose the best-of-breed solution rather than being constrained by an inflexible one-size-fits-all approach. The O-RAN framework facilitates flexible cloud-native functionality and interoperability among services offered by different vendors. However, it does come with its own deployment challenges. The traditional base station is transformed into logical E2 nodes, including the radio unit (O-RU), control unit (O-CU), and distributed unit (O-DU), which are supported through open interfaces. These components are optimized through two types of RAN intelligent controllers (RICs): near real-time (near-RT RIC) and non-real-time (non-RT RIC) RAN intelligent controllers. Considering these factors, it is essential to design cellular architecture that is interoperable with the O-RAN specifications. By embracing openness, cellular networks can leverage the advantages of vendor diversity, interoperability, and flexible functionality, paving the way for more scalable and future-proof network

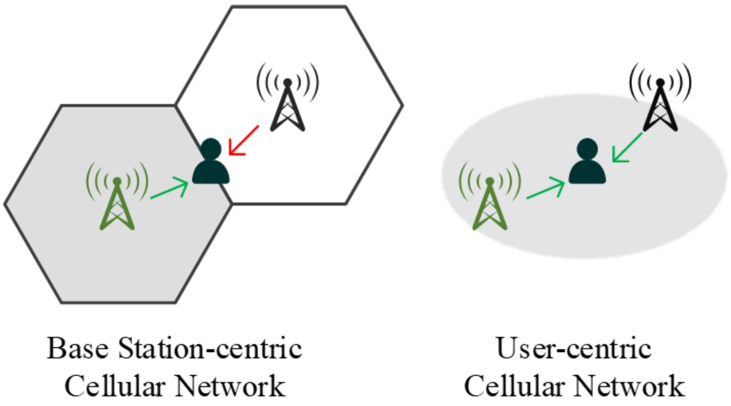


Fig. 1.1: Comparison of base station-centric and user-centric user association.

deployments.

Enabling Universal Coverage: Despite the goal of user-centric networks to provide uninterrupted coverage, there is a tendency for low-priority users to suffer as high-priority users are prioritized, resulting in delays for the low-priority users. This issue is particularly exacerbated in hotspot areas where scheduling delays are amplified. In order to achieve universal coverage and address these challenges, it is imperative to devise innovative cellular architectures that offer additional flexibility to ensure seamless coverage for users of all priorities.

1.2 User-centric Radio Access Network Architectural Overview

UC-RAN has emerged as a promising technology to address the challenges posed by traditional base station-centric UDNs [2–4,7,8]. By forming virtual cells around scheduled users, UC-RAN effectively mitigates inter-cell interference and provides a significant reduction in interference impact. Architecturally, UC-RAN separates the baseband processing unit from the radio access network (RAN), allowing for dense deployment of cost-effective data base stations (DBSs) without incurring high capital and operational costs associated with traditional hardware requirements [9]. The dense deployment of DBSs in UC-RAN also shortens the average distance between users and serving base stations, resulting in relaxed transmission power requirements for both User Equipment (UE) and DBSs.

UC-RAN's ability to mitigate inter-cell interference and reduce deployment/operational costs positions it as the ideal architecture for supporting user-centric services in ultra-dense cellular networks. A typical UC-RAN configuration consists of a tier of low-density control base stations (CBS) with large coverage, complemented by a tier of high-density switchable distributed base stations (DBS) with intermediate coverage. A key feature introduced by UC-RAN is the concept of an elastic degree of freedom known as the Service Zone (S-zone), which determines the minimal separation gap between scheduled users. The S-zone represents the size of a user-centric virtual cell centered around the scheduled UE. In the user-centric architecture, the S-zone size governs the association of DBSs, while in a base station-centric architecture, the association is determined by the base station cell, as illustrated in Fig. 1.1. During each transmission time interval (TTI), the CBS activates the most suitable DBS within S-zone centered on the UE, while ensuring no overlap occurs among adjacent S-zones. The elastic user-centric S-zone around a UE enables (i) efficient interference protection between scheduled UEs; (ii) dynamic coverage extension and shrinkage by activating a single radio unit having the best channel gain among the multiple radio units within the S-zone region; (iii) high energy efficiency due to opportunistic activation of radio units as compared to always on radio units; and (iv) uniform coverage and uninterrupted provision of QoS due to diminishing cell edge users and intercell interference [2–4, 7, 8].

1.3 Related Work

In recent works, researchers have investigated the impact of S-zone size in a UC-RAN using analytical models for both sub-6 Gigahertz and millimeter frequency bands [3,4,10–12]. These studies focus on network design and propose non-overlapping virtual cells (S-zones) around scheduled users, based on their priorities. By employing user-centric cells and macro-diversity techniques, the S-zone size can be optimized as a control parameter to achieve desired KPIs.

For example, the authors in [3] demonstrated through a statistical framework that an optimal user-centric virtual cell size exists, maximizing both area spectral efficiency and energy efficiency in UC-RAN. The authors emphasized that this virtual cell size depends on variations in the density of DBS and UEs, necessitating adaptation with changes in these parameters. In another work by [3], the authors considered Stienen cells in UC-RAN to analyze the signal-to-interference-noise ratio (SINR) distribution, area spectral efficiency, and energy efficiency. They compared UC-RAN with non user-centric architectures and showed that UC-RAN not only improves SINR but also optimizes area spectral efficiency and energy efficiency by adjusting the design parameters. The authors in [10] proposed a user-centric model for combining base stations in millimeter-wave networks, using stochastic geometry to determine coverage probability and optimal area spectral efficiency. They introduced a framework for optimizing the clustering parameter, leading to increased area spectral efficiency.

Additionally, while CoMP solutions have been investigated for over two decades, their incorporation in UC-RAN-based architectures is still in early stages. The authors in [11], derived analytical expressions for coverage probability in the downlink heterogenous network with user-centric architecture and base station cooperation. Although the numerical results were highly accurate, a closed-form expression for the coverage probability was lacking. The authors primarily focused on the analytical model for coverage probability and did not analyze the impact of cooperation on different KPIs in a user-centric network.

Other research works have addressed various optimization problems in UC-RAN. The authors in [13] formulated a max-min rate problem to reduce the power consumption of UC-RAN by joint optimization of beamforming weights and UE association with the access point. The formulated problem is first divided into two subproblems by relaxing the energy efficiency subproblem and power consumption subproblem and then solved separately utilizing the Lagrange duality method. Authors in [14] considered transmission points cluster approaches in massive multiple-input-multiple-output and millimeter-wave aided CRANs to reduce the overhead cost and ensure minimum signal changes at both network and user ends. In [15], the authors have derived approximate analytical expressions for the ergodic capacity and coverage probability for millimeter-wave user-centric dense networks.

Despite the promises of UC-RAN as a flexible architecture, several challenges need to be addressed before its practical adoption in O-RAN-based cellular networks. Firstly, the non-overlapping S-zone criterion in previous UC-RAN architectures allows only one UE to be scheduled per S-zone, leading to inherent rigidity that negatively impacts UE latency satisfaction and scheduling ratio [3, 16]. Secondly, evaluating UC-RAN's impact on latency satisfaction from a system-level perspective requires an evaluation metric that considers the temporal domain, taking into account the time at which the UE requests service and when it receives the service. Additionally, although data-driven solutions have shown effectiveness in optimizing user-centric networks [8], the increase in the number of control and optimization parameters may result in a large optimization space, requiring extensive exploration that can potentially adversely affect the cellular network during optimization.

While online data-driven approaches (such as, DRL) are well understood for their convergence, optimality, and sample data efficiency, risk awareness and minimization during training and execution has received far less attention [17, 18]. A line of previous works has used techniques to ensure risk awareness in DRL-based automation [18–24]. These techniques include: (i) expert advice [19, 20], which requires an accurate probabilistic model of the system; (ii) constrained policy optimization [21–24], which generally requires domain knowledge about which actions will lead to constraint violations; and (iii) human intervention [18], which requires a human to watch the actions coming from DRL algorithm and intervenes when needed. The fundamental issue with these approaches is that without having prior access to the probabilistic model of the real system, safety must be learned through interaction with live network, which can violate reliability (e.g. QoS requirements) during the initial stages of learning.

Moreover, although the UC-RAN architecture demonstrates significant enhancements in QoS for high-priority users, it often has adverse effects on QoS of low-priority users. This is primarily attributed to the preemptive scheduling of high-priority users with large service exclusion zones that delay the scheduling of low-priority users [8]. The impact of this issue is further intensified in hotspot areas, where the delay in scheduling is magnified, leading to coverage holes for low-priority users. To address these coverage gaps in cellular networks, there has been a growing interest in leveraging cell-free ABS-aided wireless communications to extend coverage to areas with limited infrastructure. One such approach was proposed in [25], where the authors introduced an ABS-assisted cell-free network for providing coverage to vehicles on highways with poor cellular infrastructure. They formulated determining ABS trajectories as a Markov decision process, accounting for vehicular network dynamics, and used DRL to optimize ABS trajectories for maximum vehicular coverage. In another study [26], the authors proposed a user-centric ABS swarm network, where multiple ABSs form a swarm around a user to provide customized services. They derived a semi-closed form expression for coverage probability and average achievable data rate with respect to ABS locations.

In [27], the authors considered joint optimization of location, transmit power, altitude, and bandwidth for ABS in an underlaid D2D communications network. They proposed a low-complexity iterative algorithm to obtain closed-form solutions for transmit power allocation and altitude planning subproblems. In [28], the authors considered the problem of user association and bandwidth allocation along with ABS placement in a ABS-assisted cellular network with the aim to minimize the overall average latency ratio while meeting the user quality of service requirements. They formulate a cyclic iterative algorithm to decompose the primal problem into two subproblems, i.e., the bandwidth allocation and user association problem with ABS placement.

1.4 Research Objectives

Building upon the discussions presented in Section 1.1 and Section 1.3, this dissertation introduces four novel approaches for UC-RAN optimal design and operation with the objectives of mitigating interference, reducing power consumption, ensuring diverse user/vertical service quality, facilitating proactive network operation, risk-aware optimization, adopting an open radio access network, and enabling universal coverage.

Objective 1: Investigate the benefits of a UC-RAN in mitigating interference within ultra-dense networks while dynamically optimizing coverage and enhancing the quality of experience (QoE) for users through spatial diversity techniques, such as coordinated multi-point (CoMP) technology [2].

Objective 2: Develop an intelligent, proactive, demand-driven, and elastic architecture based on UC-RAN to effectively mitigate interference and cater to the diverse needs of user applications. Furthermore, leverage the additional degrees of freedom provided by UC-RAN to enable optimal control through an online data-driven optimization framework. This framework should be capable of adapting to dynamic user demands and mobility patterns while simultaneously performing multi-objective optimization of system KPIs.

Objective 3: Design an interoperable architecture based on UC-RAN that leverages the benefits of O-RAN and the architectural flexibility offered by UC-RAN. This architecture aims to harness the full potential of O-RAN while ensuring risk-aware optimization of COPs in a live network environment. The optimization framework should enable accelerated learning of data-driven solutions while simultaneously avoiding any degradation of network performance below the minimum reliability requirements of the system during online optimization.

Objective 4: Develop an integrated aerial network in conjunction with UC-RAN to effectively tackle the problem of coverage holes in UC-RAN. By leveraging the agile deployment of aerial base stations (ABSs), on-demand rapid service provision can be achieved, providing enhanced flexibility in adapting to changing user demands and locations. However, to fully harness the potential of aerial network-assisted communication, optimal control of various parameters, including ABS deployment, transmit power, altitude, and beamwidth, is crucial.

1.5 Contributions

In light of the research objectives discussed above, the key contributions of the dissertation are outlined in the following section.

Contribution 1: We leverage stochastic geometry concepts to derive analytical expressions for coverage probability, area spectral efficiency, and energy efficiency in a CoMP-enabled user-centric architecture. Theoretical analysis and simulation results are presented, focusing on joint optimization of area spectral efficiency and energy efficiency within the user-centric architecture. Additionally, we explore scenarios where CoMP deployment can enhance network-level KPIs. This contribution provides valuable insights into the design and planning of future user-centric networks, offering a comprehensive understanding of the relationship between network parameters and system-level efficiency metrics.

Contribution 2: We propose a user-centric architecture for demand-driven elastic communication, catering to a diverse range of user applications. The proposed architecture allows the elastic virtual user-centered clusters, known as service zones (S-zones), to be malleable to specific vertical requirements. Considering the heterogeneous user requirements in future cellular communications, a multi-objective problem is formulated to optimize KPIs such as area spectral efficiency, energy efficiency, user service rate, and throughput satisfaction as a function of S-zone size for respective verticals. Given the non-stationarity of user application demands and mobility, we propose an online data-drive solution based on deep reinforcement learning framework to accurately learn the mapping of environment state and action instilling intelligence in the demand-driven elastic user-centric architecture. The proposed intelligent deep reinforcement learning framework for UC-RAN networks, named D-RAN, dynamically allocates S-zones to users such that a Pareto-optimal front is found for the formulated multi-objective function. We evaluate the convergence, efficacy, and adaptability of D-RAN to the non-stationary environment of the proposed approach through numerical results. We also compare D-RAN's performance against brute-force and state-of-the-art metaheuristics such as simulated annealing. With the proposed D-RAN framework, the paradigm of traditional cellular networks could be transformed into demand-driven, elastic, user-centric systems in future 6G and beyond networks.

Contribution 3: We present and evaluate a user-centric architecture based on O-RAN architecture to meet the QoS requirements of various verticals. We investigate two key UC-RAN COPs; 1) size of user-centric virtual cells (S-zones), and 2) number of UEs scheduled per S-zone, which are leveraged through an rApp to control latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency. Our extensive investigation show that both COPs indeed offer a powerful, mechanism to control multiple network KPIs in a flexible and scalable fashion. To cope with the performance deterioration risk associated with online network reconfiguration, which has hindered the industry uptake of online learning-based solutions, we propose offline learning leveraging a digital twin instilling risk-awareness in the DRL optimization framework. The proposed digital twin-assisted DRL framework's convergence and accumulated risk are compared against brute force results, showing an impressive performance in reaching the near-optima in a few hundred iterations. Furthermore, the risk-aware optimization framework indicates the viability of online learning techniques in live cellular networks with controlled and safe exploration. This contribution demonstrates a highly flexible O-RAN-based user-centric architecture coupled with a risk-aware DRL optimization framework that can address the fundamental tradeoff between latency, reliability, and throughput in live emerging cellular networks.

Contribution 4: We introduce and evaluate an integrated aerial network with UC-RAN to meet the emerging need of user applications. We present a multi-objective optimization framework that addresses the optimization of ABS COPs. Our analysis considers the tradeoffs between key system parameters, including ABSs location, transmit power, altitude, and beamwidth. Furthermore, we propose an optimization approach using reinforcement learning and demonstrate its superiority over brute-force methods. These findings provide valuable insights for intelligent ABS deployment and optimization, improving service provisioning for low-priority users within user-centric networks.

1.6 Current and Planned Dissemination and Publications

Academic Awards:

A1. ORISE Research Fellowship at U.S. Food and Drug Administration, Oct '21

A2. Third Prize at OU-Tulsa Research Forum 2023, Apr '23

Peer-Reviewed Journal Articles:

J1. **S. K. Kasi**, U. S. Hashmi, M. Nabeel, S. Ekin and A. Imran, “Analysis of Area Spectral & Energy Efficiency in a CoMP-Enabled User-Centric Cloud RAN,” in IEEE Transactions on Green Communications and Networking, vol. 5, no. 4, pp. 1999-2015, Dec. 2021, doi: 10.1109/TGCN.2021.3093390.

J2. **S. K. Kasi**, U. S. Hashmi, S. Ekin, A. Abu-Dayya and A. Imran, “D-RAN: A DRL-Based Demand-Driven Elastic User-Centric RAN Optimization for 6G & Beyond,” in IEEE Transactions on Cognitive Communications and Networking, vol. 9, no. 1, pp. 130-145, Feb. 2023, doi: 10.1109/TCCN.2022.3217785.

J3. **S. K. Kasi**, F. A. Khan, S. Ekin and A. Imran. “Digital Twin Empowered Risk-aware Reinforcement Learning Framework for User-centric O-RAN,” in IEEE Transactions on Cognitive Communications and Networking (2023). [**Submitted**]

J4. **S. K. Kasi**, F. A. Khan, S. Ekin and A. Imran. “User-centric Communication with Aerial Network for 6G: A Reinforcement Learning Approach.” [**Under Co-authors Review**]

J5. U. B. Farooq, **S. K. Kasi**, U. S. Hashmi, F. A. Khan, S. Ekin and A. Imran. “Service Exclusion Zone Based RAN Slicing for 6G: Opportunities, Challenges, and Research Directions,” in IEEE Network Magazine (2023). [**Submitted**]

J6. U. B. Farooq, **S. K. Kasi**, M. Manalastas, C. Zhu, B. Sheen, and A. Imran. “Optimizing Mobility in Cellular Networks: A Risk-Averse Multi-objective Reinforcement Learning Approach.” [**Under Co-authors Review**]

Peer-Reviewed Conference Papers:

C1. **S. K. Kasi**, U. Sajid Hashmi, M. Nabeel, S. Ekin and A. Imran, “Is CoMP Beneficial In User-Centered Wireless Networks?,” 2022 International Conference on 6G Networking (6GNet), Paris, France, 2022, pp. 1-5, doi: 10.1109/6GNet54646.2022.9830168.

C2. **S. K. Kasi**, U. Sajid Hashmi, S. Ekin and A. Imran, "Learning-Aided Demand-Driven Elastic Architecture for 6G & Beyond," 2023 IEEE 97th Vehicular Technology Conference: (VTC2023-Spring), Florence, Italy, 2023. [**Accepted**]

C3. U. B. Farooq, **S. K. Kasi**, M. Manalastas, C. Zhu, B. Sheen, and A. Imran. 'Risk Averse RL-based Multi-objective Mobility Management for Emerging Cellular Networks." 2023 IEEE Global Communications Conference (GLOBECOM), Kuala Lampur, Malaysia, 2023. [**Submitted**]

C4. M. Shaukat, **S. K. Kasi**, and A. Imran. "Deep Graph Reinforcement Learning for Optimization of Emerging User-Centric Radio Access Network." [**Under Co-authors Review**]

1.7 Organization

The dissertation is organized as follows: Chapter 2 addresses the challenge described in research objective 1 by providing an analytical and numerical analysis of the impact of enabling CoMP in a UC-RAN architecture. Chapter 3 addresses the challenge described in research objective 2 by proposing an architecture for demand-driven elastic user-centric communication to provide on-demand services to a diverse set of user applications. This chapter also focuses on developement of DRL-based optimization framework to learn the mapping of environment state and action, instilling intelligence in the demand-driven elastic user-centric architecture. Chapter 4 addresses the challenge described in research objective 3 by introducing a novel framework for accelerating DRL training using a digital twin. Chapter 5 addresses the challenge described in research objective 4 by proposing a solution to improve service provision for low-priority users in user-centric networks using an aerial network. Finally, Chapter 6 presents the conclusion and outlines potential future work.

CHAPTER 2

Analysis of Area Spectral and Energy Efficiency in a CoMP-Enabled User-Centric RAN

2.1 Introduction

2.1.1 Motivation

UC-RAN has emerged as a promising technology to address the high interference challenge introduced by UDNs. UC-RAN with its centralized architecture is considered an ideal architecture to support CoMP [29], which enhances KPIs such as coverage probability, SINR, and area spectral efficiency. In this chapter, we investigate the benefits of UC-RAN with CoMP in reducing interference within UDNs while simultaneously improving coverage and enhancing user QoE.

Unlike existing literature on CoMP that lacks analytical analysis of CoMP-enabled UC-RAN, this chapter focuses on analyzing average aggregate interference, area spectral efficiency, and energy efficiency in UC-RANs with CoMP. We aim to derive closed-form expressions for these KPIs and provide close bounds to numerical results. Additionally, we extend the analysis to different CoMP schemes as described in Section 2.2. By developing analytical and numerical models, we calculate optimal parameters such as S-zone size, density of distributed base stations (DBSs), and other system parameters in terms of area spectral efficiency and energy efficiency.

2.1.2 Contributions

The contributions of this chapter can be summarized as follows:

- First, we extend the user-centric architecture proposed in [4] to include cooperation between DBSs in an S-zone. We characterize the activated DBS density followed by the average interference experienced by a scheduled UE in UC-RAN using the stochastic geometry tools. (Section 2.4).

- In contrast to previous works, we derive a closed-form expression characterizing the lower bound on the probability of coverage for a scheduled UE in a CoMP enabled UC-RAN (Section 2.4). The lower bound is further utilized to derive the area spectral efficiency in UC-RAN (Section 2.4).

- We then proceed to quantify the energy consumption model for UC-RAN to support CoMP communication and the associated overhead for discovering DBS(s) providing the highest channel gains at each scheduled UE. The power consumption model is used to derive the energy efficiency of CoMP-enabled UC-RAN (Section 2.5).

- Next, we provide a comparative performance analysis of different realizations of the joint transmission mode of CoMP in UCRANs. The three realizations are categorized based on the selection strategies of cooperative DBSs in an S-zone (Section 2.6).

- Finally, the derived analytical framework is used to investigate the impact of new degrees of freedom, i.e., the S-zone size and DBS density, on the area spectral efficiency, and energy efficiency of CoMP-enabled UC-RAN. The results indicate that for any number of cooperative DBSs in an S-zone, an optimal operating point for the S-zone size and density of DBS exists that maximizes the area spectral efficiency and energy efficiency. However, the S-zone size optimal for area spectral efficiency does not need to also be optimal for energy efficiency, therefore, we provide an analysis on the tradeoff of these KPIs using the new degrees of freedom (Section 2.6).

2.1.3 Chapter Organization and Notation

Throughout this chapter, the boldface small case letter (such as $\mathbf{x}$) is used to represent a vector, and $||\mathbf{x}||$ is used to denote the $L2$ norm of vector $\mathbf{x}$ in Euclidean space. The symbol / denotes the set subtraction, whereas $\in$ denotes the set membership. The notations $\mathbb{E}_Z(.)$ and $f_Z(.)$ are used to denote the average value and probability distribution of a random variable, respectively. The symbol $Z \sim U(a, b)$ indicates a uniform distribution for values between a and b. The symbol $Z \sim \exp(\mu)$ represents an exponential distribution with average value μ. The symbol $\mathbb{1}(x > y)$ denotes a characteristic function and $b(x, r)$ represents a circle centered at a point x with a radius of size r. Finally, the Poisson point process (PPP) is denoted by Π.

The rest of the chapter is organized as follows. The problem description and research challenges are discussed in Section 2.2. The network model is explained in Section 2.3. The quantification of area spectral efficiency and energy efficiency are derived in Sections 2.4 and 2.5, respectively. System evaluation is conducted in Section 2.6. Finally, the outcomes of the chapter are concluded in Section 2.7.

2.2 Problem Description and Design Issues

In this Section, we first discuss the CoMP-enabled UC-RAN architecture in detail. We then identify the key challenges in CoMP-enabled UC-RAN architecture followed by the discussion of various methods through which CoMP can be employed.

2.2.1 CoMP-enabled UC-RAN Architecture

In a UC-RAN model, the S-zone of a predefined radius is created around all scheduled UEs during each TTI. An arbitrary UE is scheduled depending on its service requirements which is then served by one or more activated DBSs within an S-zone. The set of serving DBSs around a scheduled UE may change across TTIs depending

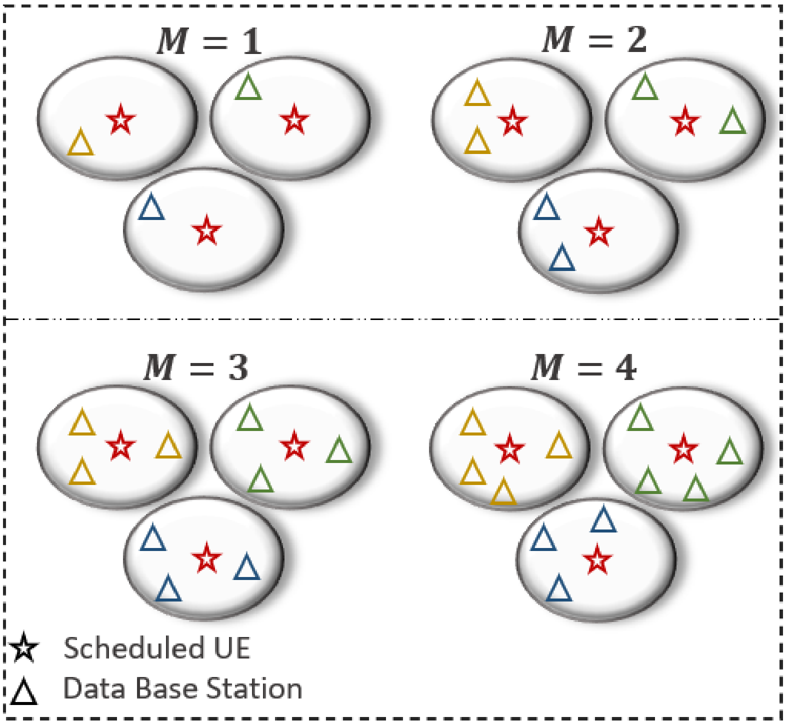


Fig. 2.1: Graphical realization of CoMP for different values of M where the different colors of DBSs correspond to DBSs of different S-zones.

on the spatial distribution of DBSs, user's mobility, and wireless channel conditions. If more than one DBS is activated, then a mechanism for cooperation is required to simultaneously transmit data to the scheduled user.

For LTE-advanced, 3GPP has identified three major downlink coordination techniques based on the complexities of implementation and required backhaul capacity [30]. These techniques can be categorized as (1) joint transmission (JT), (2) dynamic point selection (DPS), and (3) coordinated beamforming/scheduling (CB/CS). In JT, channel state information (CSI) and user data are shared between the coordinated set of transmission points. The mode of operations of DPS is very similar to JT except that the data is transmitted by one transmission point at a specific TTI. Unlike JT and DPS, CB/CS requires only CSI to be shared between the transmission points. Even though the backhaul bandwidth requirement of JT is the highest amongst all the aforementioned coordination techniques, the maximum gain in performance is also offered by JT [29].

In this chapter, we mainly focus on the realization of JT for enabling cooperation between the transmitting DBSs in an S-zone. To make a comparative analysis, we consider that at any specific TTI, the number of cooperative DBSs in an S-zone cannot exceed $M \in Z^+$, where Z^+ is the set of positive integers. An exact bound on the number of cooperative DBSs is not considered because the number of DBSs in an S-zone depends on the density of DBSs, therefore, it is not realistic to assume that each S-zone will have at least a certain number of DBSs. In Fig. 2.1, we show the graphical realization of CoMP enabled UC-RAN for different values of M. An important point to consider here is that in each S-zone, no more than M DBSs coordinate to simultaneously transmit the same data to the scheduled UE. Also, if the number of DBSs is less than M in any S-zone then all of them will be activated by the BBU to serve a scheduled UE (as shown in Fig. 2.1 for $M = 4$ scenario).

2.2.2 CoMP Clustering Challenges in UC-RAN

CoMP is often realized with small clusters of DBSs due to the complexity required for coordination which increases exponentially with the increase in coordinated cells in a cluster [29]. An innate question that arises is whether the benefits of CoMP exceed the complexities involved in enabling it in a UC-RAN. Although CoMP has been widely studied in HetNets, there has been limited work on CoMP in UC-RAN. Therefore, we discuss the challenges faced by enabling CoMP in UC-RAN which can be classified as:

Is it spectral-efficient and energy-efficient to realize CoMP? As briefly discussed in Section 2.1, gains through enabling CoMP are achieved at the expense of energy efficiency. However, the underlying assumption that realizing CoMP increases the area spectral efficiency in any network may be too much of an exaggeration. The two important points to notice in the UC-RAN architecture shown in Fig. 2.2 are that: (i) there are no cell-edge users because virtual cells are created around the users, and (ii) increasing coordinating DBSs in an S-zone leads to more number of activated DBSs in the overall network increasing the number of interferers in the network. By enabling CoMP, the dominant interferers are virtually removed which intuitively should reduce the overall interference experienced by a UE. However, in an architecture such as shown in Fig. 2.2 where S-zones are non-overlapping and DBSs are activated "on-demand" of a UE within a specific user-centric cell area (S-zone), the existence of dominant interferers is unfounded. Both of these observations are directly related to the spectral gains achieved by realizing CoMP because: (i) the major feature of CoMP, which is to mitigate cell-edge user interference, is not applicable in UC-RAN architecture with non-overlapping S-zones, and (ii) interference is minimum only when one DBS is activated in every S-zone. However, ideal CoMP gains can only be achieved if the increase in received signal powers at the scheduled UEs through coordination is much more than the increase in the aggregate interference in the network.

It must be noted that by enabling CoMP, the number of coordinating DBSs increases but so does the number of interferers because more DBSs are activated in other S-zones. In the UC-RAN architecture discussed in this chapter, the selection combining diversity technique is used to activate no more than M DBSs that provide the highest channel gains to a scheduled UE in an S-zone. Because of this reason, even if the number of coordinating DBSs is increased, the dominant impact on received signal power at a scheduled UE (in most cases) will be from the DBS providing the largest channel gain. The disparity in the channel gains from coordinating DBSs increases for higher values of the path-loss exponent. Contrarily, the increase in the number of interferers due to an increase in the number of activated DBSs will increase the overall interference experienced at a scheduled UE. This phenomenon increases the interference roughly by M fold while received signal power is not increased by the same factor leading to a decrease in SIR due to uncoordinated interference out of the S-zone. Based on the discussion, it is quite evident that whether CoMP transmission in UC-RAN architecture will enhance spectral and energy efficiency is not a trivial research problem.

What is the optimal S-zone size? Another key parameter involved in the design of UC-RAN is the size of the cooperative cluster or S-zone created around a scheduled UE. Increasing the S-zone size causes: (i) an increase in the average distance between scheduled UE and interfering DBSs, (ii) an increase in the possible set of DBSs within a cooperative cluster yielding high macro-diversity gain, and (iii) a possible decrease in the total number of activated DBSs because the number of activated DBSs in an S-zone is bounded by M in the UC-RAN architecture shown in Fig. 2.2; meaning that if the total number of S-zones are decreased in a network, then the total number of activated DBSs will also decrease. Hence, larger the S-zone size, lesser the number of scheduled UEs and consequently activated DBSs serving those scheduled UEs. Conversely, larger S-zone sizes may lead to a decrease in the overall DBSs power consumption and increase spectrum reuse such that more number of scheduled UEs can be served simultaneously. Given these insights, we investigate the optimal S-zone size that yields an ideal tradeoff between area spectral efficiency, energy efficiency, or some combination of both in a CoMP-enabled UC-RAN.

Which DBSs to activate/deactivate? Another important design parameter is to decide which DBSs should be kept activated and which DBSs should remain deactivated assuming that measurements of average received signal powers from all DBSs are available at a scheduled UE. In [31], the authors have discussed two schemes by which DBSs in JT can be activated/deactivated. In scheme 1, M DBSs that provide the best average received powers are activated while others are kept deactivated. In scheme 2, a cooperative DBS is only considered for CoMP if the received signal power from the DBS is above some percentage of the maximum received signal power. For example, if the maximum received signal power from all DBSs is P_{max}, then the received signal power from the cooperative DBSs P_{CoMP} should be greater

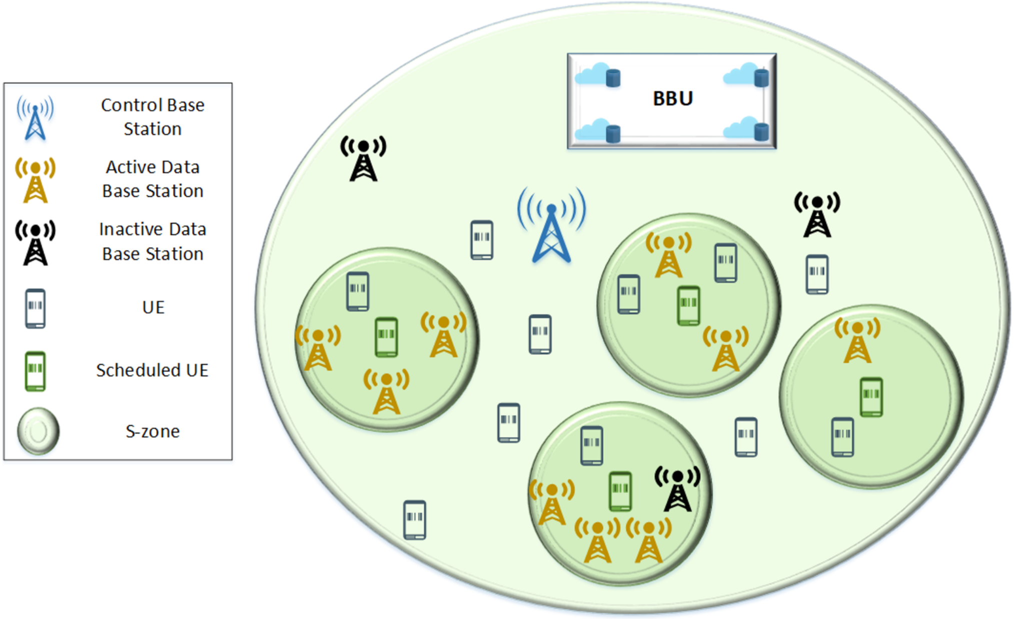


Fig. 2.2: CoMP-enabled UC-RAN architecture with activation region of radius R_{szone} for a scheduled UE.

than $\beta \times P_{max}$, where β can be any value between 0 and 1. Another method by which DBSs are activated/deactivated, termed as a random scheme, is to randomly activate DBS(s) in the S-zone while deactivating other DBSs. Therefore, DBSs activation/deactivation and its impact on area spectral efficiency, and energy efficiency in a CoMP-enabled UC-RAN architecture is another research problem that we investigate in this chapter.

2.3 Network Model

2.3.1 Spatial and Channel Model

In this chapter, we consider an underlaid cloud radio access network with an ultra-dense DBS deployment scenario. The ultra-dense deployment of DBSs is an imaginable scenario in future networks [32]. Both DBSs and UEs are spatially modeled as independent stationary Poisson point processes Π_{DBS} and Π_{UE} with densities λ_{DBS} and λ_{UE}, respectively. The average number of DBSs in an S-zone is given by $\lambda_{DBS}\pi R_{szone}^2$ that is characterized by λ_{DBS} and Lebesgue measure [33] of a disc with radius R_{szone}.

The communication channel between an arbitrary user $x \in \Pi_{UE}$ and DBS $y \in \Pi_{DBS}$ is modeled by $h_{xy}\ell||x-y||$, where $h_{xy} \sim \exp(1)$ is a exponential random variable with unit mean representing the effects of Rayleigh fading and $\ell||x-y||$ is the large-scale path loss model. The large-scale path loss model is given by the frequency-dependent constant K, distance between the UE and DBS $||x-y||$, and path loss exponent $\alpha > 2$ such that $\ell||x-y|| = ||x-y||^{-\alpha}$. All DBSs are assumed to transmit at equal power levels P_{DBS}, and each DBS and UE is equipped with a single antenna. We also assume that the thermal noise is negligible, hence, the communication is interference-limited.

2.3.2 User-centric Clustering in UC-RAN

In this chapter, we use the user-centering cluster mechanism given in Algorithm 1 [4] for UC-RAN. The UEs are scheduled at each TTI by the macro-cell or BBU according to their scheduling priorities which are marked according to a uniform random distribution $p_{UE} \sim U(0,1)$. The lower the mark value of a UE, the higher the scheduling priority it possesses. For the UEs which are not scheduled yet, their scheduling priorities increase in the subsequent TTIs until they are scheduled to be served.

A UE x is scheduled (i.e., $p_{sch}^{x} = 1$) iff its scheduling priority is the highest in its neighborhood which is characterized by the cluster radius R_{szone}. To be succinct, this means that within a circle of radius R_{szone} centered at UE, there is no other UE with a higher scheduling priority, and the minimum distance between any two S-zones should be at least $2R_{szone}$. Note that this circle (S-zone) is commensurate to the size of the cooperative cluster. The dynamic change in S-zone size allows the flexibility to activate DBSs in an S-zone depending on which scheme of joint

```
Algorithm 1: UE Scheduling Algorithm in CoMP-based UC-RAN
Input  : $\Pi_{UE}$, $\Pi_{DBS}$, $R_{szone}$
Output: $\Pi'_{UE}$, $\Pi'_{DBS}$
Initialize the set of UEs and the activated DBS(s) as $\Pi'_{UE} \leftarrow \emptyset$, $\Pi'_{DBS} \leftarrow \emptyset$.;
Assign random priorities to each UE based on $p_{UE} \sim U(0,1)$.;
foreach $x \in \Pi_{UE}$ do
    $p^x_{sch} \leftarrow 1$;
    foreach $y \in \Pi_{UE}$ do
        if $y \in b(x, 2R_{szone})$ and $p_{UE}(y) > p_{UE}(x)$ then
            if $y \neq x$ then
                $p^x_{sch} \leftarrow 0$;
    if $p^x_{sch} = 1$ then
        $\Pi'_{UE} \cup x$;
foreach $m \in \Pi'_{UE}$ do
    $DBS \leftarrow \emptyset$;
    foreach $n \in \Pi_{DBS}$ do
        if $n \in b(m, R_{szone}) \neq \Phi$ then
            $DBS \cup m$;
    Rank $DBS$ in order of smallest path-loss criteria such that path-loss
      $(DBS_i) <$ path-loss $(DBS_j)$, $\forall i < j$;
    if $|DBS| \leq M$ then
        $\Pi'_{DBS} \cup DBS_i$, $\forall DBS_i \in DBS$;
    else
        $\Pi'_{DBS} \cup DBS_i$, $DBS_i \in DBS$, $i \leq M$;
Scheduled users $\Pi'_{UE}$ are served from the coordinating DBSs $\Pi'_{DBS}$;
```

transmission is used to service a scheduled UE. A macro-cell or BBU is responsible for both the activation of DBSs in a cooperative cluster and delegating the size of a cooperative cluster to scheduled UEs.

The on-demand activation of DBSs makes UC-RAN capable of self-organizing its coverage according to the spatiotemporal variation in the user demography. Though, for the UC-RAN architecture to avoid coverage holes in areas where there are no DBSs available to provide coverage to a scheduled UE, DBSs need to be deployed densely so that at any time there is at least 1 DBS available to provide services to a scheduled UE. In the case of a void cluster, which is an unlikely scenario, the scheduled UE can be clustered together to nearby scheduled UEs using clustering strategies discussed in [34].

2.3.3 Signal Model and Probe Cluster

Consider a scheduled UE $x \in \Pi^{'}_{UE}$, where $\Pi^{'}_{UE}$ is the PPP representing scheduled UEs. $\Pi^{'}_{UE}$ unlike Π_{UE} is a non-stationary Poisson point process that can be modeled as a type II Matern hardcore process [33]. The density of scheduled UEs can be approximated by an equidistant stationary Poisson point process [35] given as:

$$\lambda^{'}_{UE} = \frac{1 - \exp(-\lambda_{UE} 4\pi R^2_{szone})}{4\pi R^2_{szone}}. \tag{2.1}$$

For a scheduled UE x, let $\Pi^{'C}_{DBS} = \Pi^{'}_{DBS} \cap b(x, R_{szone})$ be the set of DBSs that are activated by the BBU to serve x based on a scheduling criterion [4]. $\Pi^{'}_{DBS}$ represents the spatial distribution of no more than M activated DBSs in an S-zone. Similarly, let $\Pi^{'I}_{DBS} = \Pi^{'}_{DBS} \backslash \Pi^{'C}_{DBS}$ be the set of DBSs which are simultaneously transmitting to the scheduled UE $u \in Pi^{'}_{UE}$ where $u \neq x$. Given these observations, we model the received signal at a particular scheduled UE as:

$$q_x = \sum_{i \in \Pi^{'C}_{DBS}} \sqrt{P_{DBS} h_{ix} \ell ||x - i||} s_x + \sum_{u \in \Pi^{'}_{UE} \backslash x} \sum_{j \in \Pi^{'I}_{DBS}} \sqrt{P_{DBS} h_{jx} \ell ||x - j||} s_u, \tag{2.2}$$

where s_x is the signal transmitted to a scheduled UE x. Capitalizing on the stationary characteristics of the scheduled UE's PPP, focusing on a typical UE is sufficient. As maintained by Silvnyak's theorem [33], the addition of a single point does not affect the law of stationary PPP, therefore, a probe UE is added at the origin. Additionally, the received signal q_x can be simplified with $\ell ||i - y|| = ||i - y||^{-\alpha} = ||i - o||^{-\alpha}$ where the index o is the location of a typical UE.

2.4 Characterizing the Area Spectral Efficiency of a UC-RAN

A typical UE is served by at most M DBSs in an S-zone centered at origin o with a ball area of $b(o, R_{szone})$ where R_{szone} is the radius of the ball. The cooperative cluster is defined as:

$$C = arg_{r_1, r_2, \ldots, r_n \subset \Pi'_{DBS}} \sum_{i=1}^{n} h_i r_i^{-\alpha}, \tag{2.3}$$

where $n \leq M$, r_i denotes the distance between serving DBS i and scheduled UE, h_i captures the effect of Rayleigh fading, and Π'_{DBS} is the resultant PPP of activated DBS with density λ'_{DBS}. With the joint transmission mode of CoMP, all the DBSs in a cooperation set jointly transmit the same message to a scheduled UE on the same time-frequency resource [36, 37]. Therefore, the signal-to-interference ratio at a typical UE in an interference-limited environment can be expressed as:

$$SIR = \Gamma_{UE} = \frac{\sum_{i \in \Pi'^{C}_{DBS}} h_i r_i^{-\alpha}}{\sum_{j \in \Pi'^{I}_{DBS}} h_j r_j^{-\alpha}}. \tag{2.4}$$

The noise power at a scheduled UE is at much lower levels as compared to the aggregate interference which is why the assumption of an interference-limited environment is valid even with the induced spatial repulsion between scheduled UEs and activated DBSs in other S-zones [11].

2.4.1 Expected Aggregate Interference and Modified Density of Activated DBSs

According to reduced Palm measure and Slivnyak's theorem [33], the expected aggregate interference at a typical UE can be expressed as:

$$\mathbb{E}_{\mathbf{I}}[I] = \mathbb{E}\bigg(\sum_{j \in \Pi'^{I}_{DBS}} h_j r_j^{-\alpha} \bigg). \tag{2.5}$$

According to Campbell's theorem [33], the expectation term in above expression can be simplified to:

$$\mathbb{E}_{\mathbf{I}}[I] = \int_{R_{szone}}^{\infty} 2\pi\lambda^{'}_{DBS}\mathbb{E}[H] r^{1-\alpha} dr, \tag{2.6}$$

where $\lambda^{'}_{DBS}$ is the density of activated DBSs and $\mathbb{E}[H]$ is the expected value of small-scale fading. By integrating and substituting $\mathbb{E}[H] = 1$, we get:

$$\mathbb{E}_{\mathbf{I}}[I] = \frac{2\pi\lambda^{'}_{DBS}}{(\alpha - 2)(R_{szone}^{\alpha-2})}. \tag{2.7}$$

The density of activated DBSs $\lambda^{'}_{DBS}$ can be approximated as $p_{ACT}\lambda_{DBS}$, where λ_{DBS} is the density of original DBSs distribution and p_{ACT} is the activation probability of DBSs in an S-zone.

Theorem 1. The activation probability of DBSs in a CoMP-enabled UC-RAN can be expressed as follows:

$$\begin{aligned} p_{ACT} = \Big(1 - \exp(-\lambda^{'}_{UE}\pi R_{szone}^2)\Big) . \bigg(\frac{\Gamma(M+1, X)}{\gamma(M+1)} + \\ \exp(-X)\bigg[\frac{M(X)^{M+1}{}_2F_2(1, M+1; M+2, M+2; X)}{(M+1)\gamma(M+2)} - 1\bigg]\bigg), \end{aligned} \tag{2.8}$$

where ${}_pF_q(a_1, ..., a_p; b_1, ..., b_q; z)$ is the generalized hypergeometric function, $\gamma(x)$ is the complete gamma function, $\Gamma(x, y)$ is the upper incomplete gamma functions and $X = \lambda_{DBS}\pi R_{szone}^2$ is the average number of DBSs in a circle.

Proof: See Appendix A. ■

From Eq. 2.7 and Eq. 2.8, we make the following remarks:

- Expected aggregate interference increases with the increase in the density of activated DBSs which is a function of M, R_{szone}, $\lambda_{UE}^{'}$ and λ_{DBS}. However, for a fixed density of scheduled UEs and DBSs, the only tunable parameters are M and R_{szone}. Both parameters will have a direct impact on the average aggregate interference experienced at a typical scheduled UE.

- It can be observed that by enabling CoMP, i.e., $M > 1$, the average aggregate interference will increase with the increase in M. Likewise, reducing the S-zone size will increase the overall number of DBSs activated for serving scheduled UEs, thereby leading to an increase in average aggregate interference.

2.4.2 Coverage Probability

A typical UE's probability of coverage can be defined as the probability of received SIR to be greater than a desired SIR threshold value (γ_{th}). The mathematical expression of the coverage probability can be simplified as:

$$P_{cov}(\gamma_{th}, R_{szone}) = P_r(\Gamma_{UE} \geq \gamma_{th}) = 1 - P_r(\Gamma_{UE} < \gamma_{th}). \tag{2.9}$$

Substituting the value of Γ_{UE} from Eq. 2.4 in Eq. 2.9, we obtain:

$$P_{cov}(\gamma_{th}, R_{szone}) = 1 - P_r\left(\frac{\sum_{i \in \Pi'^{C}_{DBS}} h_i r_i^{-\alpha}}{\sum_{j \in \Pi'^{I}_{DBS}} h_j r_j^{-\alpha}} < \gamma_{th}\right). \tag{2.10}$$

$$P_{cov}(\gamma_{th}, R_{szone}) = 1 - P_r\left(\sum_{i \in \Pi'^{C}_{DBS}} h_i r_i^{-\alpha} < \gamma_{th} \sum_{j \in \Pi'^{I}_{DBS}} h_j r_j^{-\alpha}\right). \tag{2.11}$$

Considering the aggregate interference in the above expression as a random variable, we can average the SIR distribution over all instances of interference between non-cooperating active DBSs that will allow us to simplify the above expression as:

$$P_{cov}(\gamma_{th}, R_szone) = 1 - \mathbb{E}_{\mathbf{I}}\Big[P_r\Big(S < \gamma_{th} I\Big)\Big], \tag{2.12}$$

where $S = \sum_{i \in \Pi'^{C}_{DBS}} h_i r_i^{-\alpha}$ and $I = \sum_{j \in \Pi'^{I}_{DBS}} h_j r_j^{-\alpha}$ denote the desired signal power and aggregated interference strength, respectively.

Theorem 2. The lower bound on the coverage probability of the typical user in a CoMP-enabled UC-RAN can be given as follows:

$$P_{cov}(\gamma_{th}, R_{szone}) \geq 1 - \exp\left(- \frac{\lambda_{DBS} \pi^{1-\delta} \delta \gamma \left(\delta, \dfrac{\gamma_{th} 2\pi \lambda'_{DBS} R^2_{szone}}{\alpha - 2} \right)}{(\gamma_{th} 2\lambda'_{DBS})^{\delta} (R_{szone})^{-\delta(\alpha-2)} (\alpha - 2)^{-\delta}} \right), \tag{2.13}$$

where $\delta = \frac{2}{\alpha}$ and $\gamma(a, b) = \int_a^b t^{a-1} \exp(-t) dt$ is the lower incomplete Gamma function.

Proof: See Appendix B. ■

2.4.3 Area Spectral Efficiency

Building on the coverage probability metric obtained, we define the area spectral efficiency (ASE) performance metric in this Section. The average area spectral efficiency can be defined as [38],

$$ASE = \lambda'_{UE} \log_2(e) \int_0^{\infty} \frac{P_{cov}(\gamma_{th}, R_{szone})}{1 + \gamma_{th}} d\gamma_{th}, \tag{2.14}$$

where λ'_{UE} is the modified density of the PPP representing scheduled users. Under the assumption that all users transmit at the same rate $\log_2(1 + \gamma_{th})$ and the transmission is considered successful only if the received SIR is above the desired threshold γ_{th}, the ASE metric can be lower bounded [39] as:

$$ASE = \lambda'_{UE} \log_2(1 + \gamma_{th}) P_{cov}(\gamma_{th}, R_{szone}). \tag{2.15}$$

Given a desired threshold, the average area spectral efficiency metric represents the sum of the maximum of average bits transmitted per unit Hertz bandwidth per unit area. It is worth noting that ASE is dependent on the density of scheduled UEs, γ_{th} and coverage probability. The bound for coverage probability and density of scheduled UEs will be tight for any value of γ_{th}, however, the multiplication of term $\log_2(1+\gamma_{th})$ with coverage probability and density of scheduled UEs is expected to slightly loosen the bound of ASE values for higher values of γ_{th}.

Similar to the coverage probability, ASE is also coupled with the size of the S-zone and the density of activated DBSs. While an increase in the cluster size reduces the density of scheduled UEs, it also improves SIR due to a reduction in the number of interfering DBSs. Similarly, by enabling CoMP, an increase is expected in the received signal power at a typical UE. However, CoMP also increases the number of interfering DBSs. Therefore, both these parameters can be treated as the design parameters of a UC-RAN architecture for which there exist optimal values which maximize the network-wide ASE.

2.5 Characterizing The Energy Efficiency Of A UC-RAN

In this Section, we quantify the energy efficiency (EE) performance of the proposed CoMP-enabled UC-RAN architecture. Enabling CoMP and exploiting spatial diversity gain by activating DBS(s) with maximum channel gains will increase the energy consumption cost. At the same time, only activating some DBSs will improve the energy efficiency compared to the mechanism in which all the DBSs are kept ON [40, 41]. The energy efficiency can be formulated as [3]:

$$EE = \frac{\log_2(1+\Gamma_{cran})}{P_{cran}}, \tag{2.16}$$

Table 2.1: CoMP-enabled UC-RAN simulation parameters.

Symbol	Parameter Name	Parameter Value
-	Dimensions of Simulation Region	$100\,m \times 100\,m$
λ_{UE}	UE's average density	$10^{-1}\backslash m^2$
λ_{DBS}	DBS's average density	$3 \times 10^{-2} - 1.3 \times 10^{-1}\backslash m^2$ (Variable)
$\alpha, \alpha_{near}, \alpha_{far}$	Path-loss exponents	$3, 3, 6$
R_{szone}	S-zone Size	$1\,m - 10\,m$ (Variable)
M	Maximum Number of Cooperative DBSs	$1 - 5$ (Variable)

where P_{cran} is the average power consumption of the whole network and Γ_{cran} is the effective SIR [3]. In the power consumption model, we focus on the overhead associated with enabling CoMP and discovering the best DBS(s) for the scheduled user association. During the discovery process, each DBS estimates the channel gain from the scheduled UE which will contribute to the energy consumption of the network.

The power consumption model proposed in this chapter is inspired from [42], wherein the authors proposed an accurate model for power dissipation considering parameters such as cooling, power amplifiers, baseband processing, and antenna interface. A related but modified model designed specifically for C-RAN was provided in [43] that uses parameterization specific to C-RAN efficiency.

The average power consumption can be modeled as:

$$P_{cran} = \omega_{cran}(N, \theta) P_O + P_{sp} + \Delta_u P_u + P_{ou}, \tag{2.17}$$

where P_O is the power consumption of the DBS allowing it to operate in listening mode, P_u is the transmission power of a UE, Δ_u is a factor for radio frequency module of power consumption at the UE, P_{ou} is the circuit power consumed at the UE, and P_{sp} is the power consumption due to signal processing overhead. The UC-RAN coefficient is directly proportional to the average number of cooperative

Table 2.2: CoMP-enabled UC-RAN power consumption parameters.

Symbol	Parameter Name	Parameter Value
P_u	UE transmit power	1 W
P_O	DBS fixed power consumption	6.8 W
Δ_u	Radio frequency component's power consumption	4 W
P_{ou}	UE device discovery circuit power consumption	4.3 W

DBSs in each S-zone (represented by N) and a parameter θ which characterizes the implementation efficiency. By using a simple linear parameterization, the UC-RAN coefficient can be modeled as $\omega_{cran}(N, \theta) = \theta N$ where $0 \leq \theta \leq 1$. In this chapter, θ is set to 1 to realize the least efficient UC-RAN implementation in terms of fixed power consumption of activated DBSs.

Further, enabling CoMP requires additional power consumption for signal processing. The signal processing overhead required for CoMP is calculated as [44],

$$P_{sp} = 58(0.87 + 0.03N^2), \tag{2.18}$$

where N is the average number of cooperative DBSs activated in each S-zone and is given by $N = \lambda'_{DBS}/\lambda'_{UE}$. The network power consumption can now be given as:

$$P_{cran} = NP_O + 58(0.87 + 0.03N^2) + \Delta_u P_u + P_{ou}. \tag{2.19}$$

From the expression given in Eq. 2.15, it can be observed that EE is a function of the size of the S-zone, density of activated DBSs, and the number of cooperating DBSs. However, the optimal values of these parameters will be different for ASE and EE leading to an important design question as to what values should be chosen to maintain a balance between the system's area spectral efficiency, and energy efficiency. In the next Section, we discuss the system evaluation considering the above-mentioned design question.

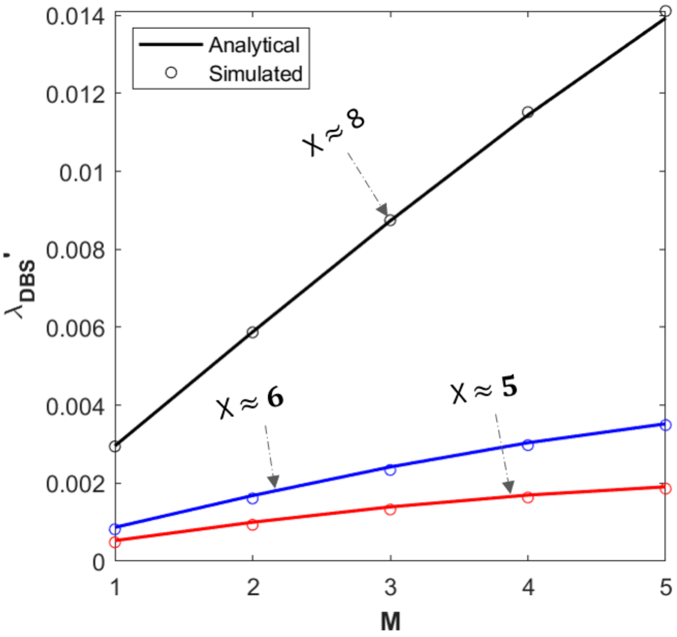


Fig. 2.3: Average activated DBS density (λ'_{DBS}) in the network for different numbers of maximum cooperating DBSs (M) within an S-Zone.

2.6 System Evaluation

In this Section, we evaluate the performance of the proposed CoMP-enabled UC-RAN using MATLAB simulations by setting the simulation parameters as shown in Table 2.1. The service area under consideration is a square of 100 $m \times 100$ m. In the service area, UEs and DBSs are distributed through Poisson point processes with densities λ_{UE} and λ_{DBS}, respectively. At each TTI, UEs are scheduled according to the algorithm initially proposed in [4] and discussed in Section 2.3. The size of the virtual cell (S-zone) and density of DBSs are varied across different experiments to study their impact on ASE and EE of a CoMP-enabled UC-RAN. The transmission power value of each activated DBS is set to 1 Watt and the path-loss exponent is set to 3. The maximum size of cooperative DBSs is set to M and Monte-Carlo simulations are employed for 10^4 realizations in each experiment.

2.6.1 Validation of the Modified Density of DBSs

Fig. 2.3 presents the validation of the analytical model for the modified density of DBSs expressed in Eq. 2.8. For different values of M, the theoretical values are consistent with the simulated values of DBSs modified density. As expected, the density of activated DBSs λ'_{DBS} increases with an increase in M due to more number of activated DBSs in each S-zone.

Similarly, the impact of average number of DBSs within a circular region (calculated as $X = \lambda_{DBS}\pi R^2_{szone}$) can be observed by varying the DBS deployment density (λ_{DBS}) and the radius of the S-zone (R_{szone}). For X approximately equal to 8, 6, and 5, the density of activated DBSs decreases as the value of X is decreased. This is mainly because X is the average number of DBSs in a circle that can vary across different S-zones depending on the random distribution of the Poisson point process. However, if $M << X$ then for different values of X, there will be little or no impact on the density of activated DBSs as each S-zone will probably have at least M DBSs.

2.6.2 Validation of Coverage Probability for JT scheme 1

Fig. 2.4 compares the analytical and simulated results of coverage probability with different values of desired SIR thresholds γ_{th} for both CoMP-enabled and no-CoMP scenarios. It can be observed that with the increase in the value of γ_{th}, coverage probability is decreased. Moreover, the analytical coverage probability curves provide a lower bound to the simulated curves (as discussed in Section 2.4).

It is also important to note that there is a slight offset in the analytical and simulated curves for different values of the maximum number of cooperative DBSs in an S-zone, i.e., M. This offset, also observed by authors in [45], can be explained by recalling the derivation of expected aggregate interference expression in Section 2.4.

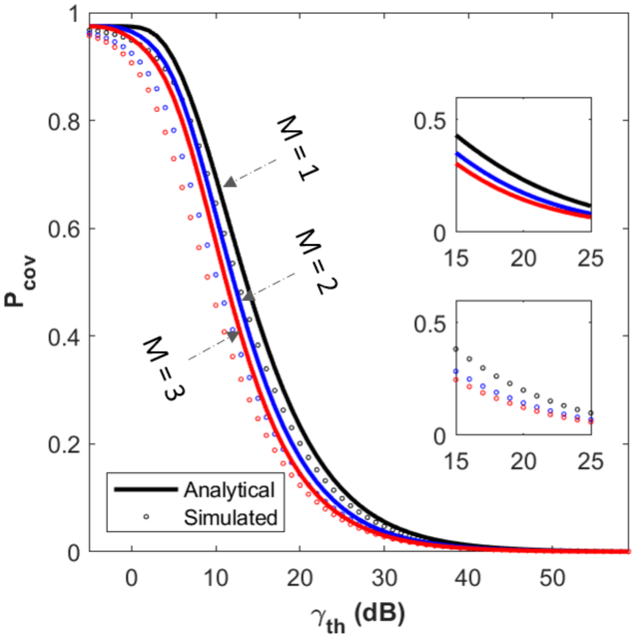
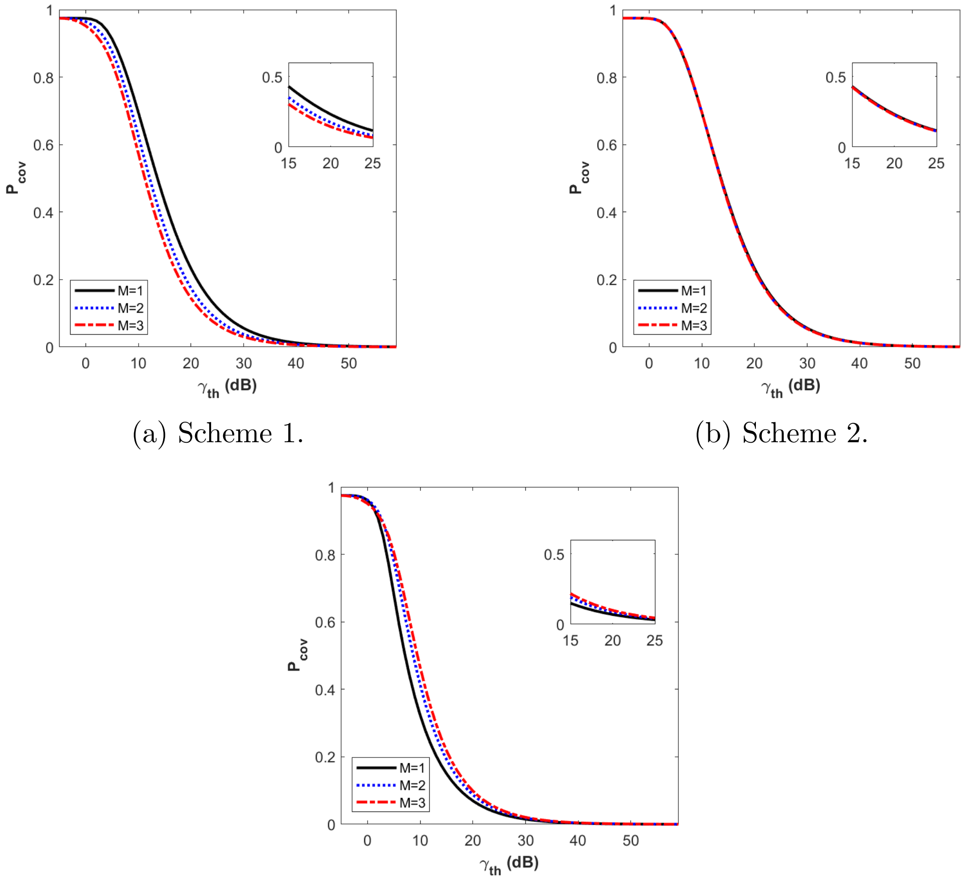


Fig. 2.4: Coverage probability for different SIR requirements (γ_{th}) and numbers of maximum cooperating DBSs (M) within an S-Zone.

In the devised analytical model, the Campbell theorem assumes an infinite number of interferes in the network. However, in our simulations, we can only consider a finite service region consequentially resulting in a finite number of interferers. The difference in the analytical and simulated aggregate interference contributes to the offset observed in the coverage probability curves.

Another interesting observation in Fig. 2.4 is the difference in the coverage probabilities of CoMP-enabled and no-CoMP UC-RAN architectures. As discussed in Section 2.2, the UC-RAN architecture with non-overlapping S-zones removes the possibility of cell-edge UEs, therefore, the major feature of CoMP to alleviate cell-edge interference is not applicable. Without any cell-edge UE, the only constructive impact of CoMP is the increase in accumulated signal powers due to coordination in an S-zone. However, by activating DBSs according to scheme 1 of JT (spatial diversity technique), the DBS chosen first will always be a dominant contributor towards the accumulated signal power. Therefore, when CoMP is enabled, the signal powers will only increase by a small fraction due to the random deployment of



(a) Scheme 1.

(b) Scheme 2.

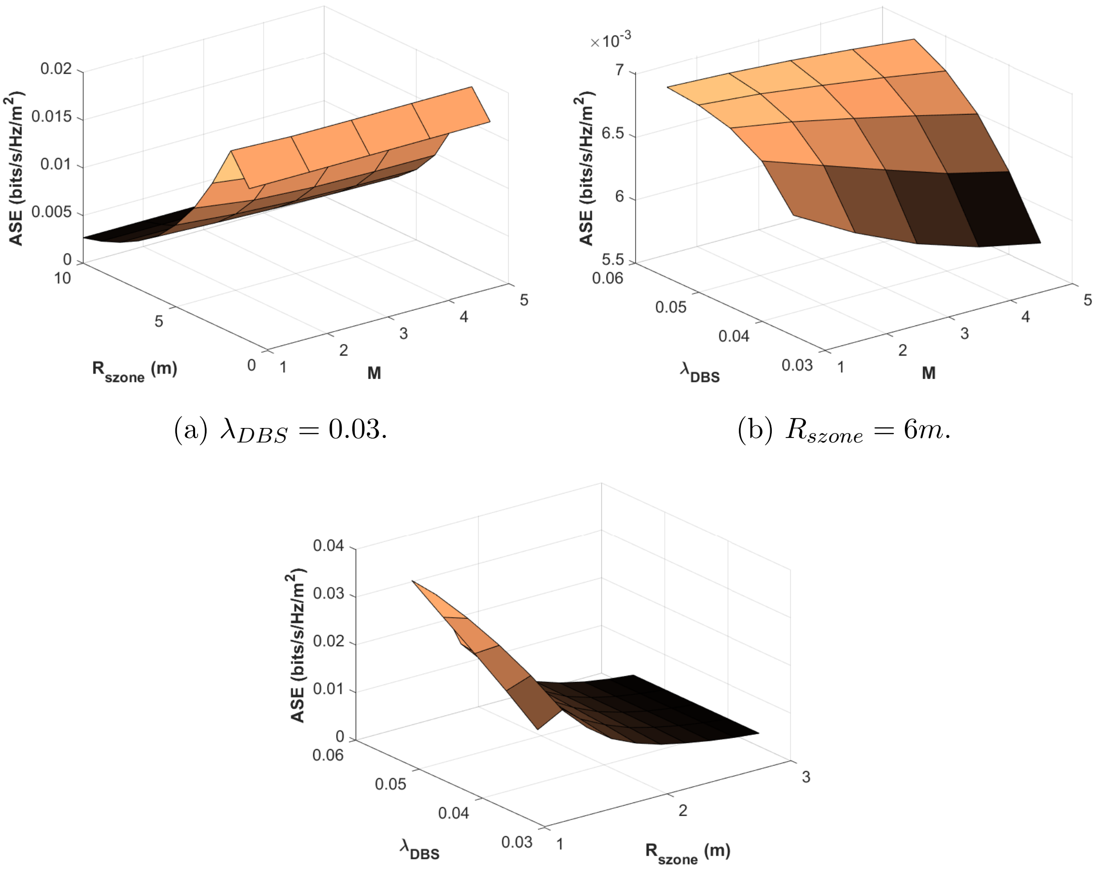
(c) Scheme 3.

Fig. 2.5: Performance comparison of coverage probabilities for scheme 1, scheme 2 and scheme 3 of joint transmission mode.

DBSs and the path-loss model. On the other hand, the aggregate interference will increase linearly resulting in the degradation of coverage probability at a typical UE.

2.6.3 Comparison of JT Schemes Coverage Probability

In Section 2.2, we discussed different schemes of joint transmission mode based on which DBSs are activated. In Fig. 2.5, we compare the performance of different schemes in terms of coverage probabilities. By employing scheme 2 of joint transmission, we only choose DBSs for cooperation if $P_{CoMP} > 0.9P_{max}$, where P_{CoMP} is the received signal power of cooperative DBS and P_{max} is the maximum of all



(a) $\lambda_{DBS} = 0.03$.

(b) $R_{szone} = 6m$.

(c) $M = 2$

Fig. 2.6: Area spectral efficiency of the CoMP-enabled UC-RAN with varying S-zone radius, DBS density and M for $\gamma_{th} = 4\ dB$.

received signal powers in an S-zone.

As discussed in the previous subsection that the dominant contributor in signal power is always the DBS that provides the maximum channel gain, therefore, occasionally, the signal power of the second maximum or third maximum DBS satisfies the criterion of scheme 2. For this reason, in Fig. 2.5 (b), it can be observed that the coverage probabilities for CoMP or no-CoMP are approximately the same. Enabling CoMP in UC-RAN architecture using schemes 1 and 2 does not improve the coverage probabilities at an arbitrary UE. Hence, it buttresses our claim that enabling CoMP in UC-RAN with the proposed architecture degrades the performance in terms of coverage probability, ASE, and EE (as further shown in Fig. 2.6 and

7). However, employing a random scheme of DBS selection where no more than M DBSs are randomly selected in each S-zone will show the improvement in terms of coverage probability in a CoMP-enabled UC-RAN (as shown in Fig. 2.5 (c)). Nevertheless, the random selection of base stations is not a realistic scenario since in most cellular networks the base station offering the strongest channel link between UE and base station is selected for communication. Though for the test of concept, we show the results for the random scheme of JT in Fig. 2.5 (c).

JT with random selection of DBSs is one such instance where employing CoMP in a UC-RAN architecture with non-overlapping S-zones will not degrade the performance in terms of coverage probability. However, notice that the coverage probability for $M = 1$ in a random scheme is much less than the coverage probability for $M = 1$ in scheme 1 and scheme 2 because spatial diversity is not utilized in the random scheme.

From the results shown in Fig. 2.5, we can establish that the joint transmission technique of CoMP will not benefit the user-centric network, however with the requirement that: (i) the user-centric virtual clusters (S-zones) are non-overlapping, (ii) DBSs are activated using selection combining in an S-zone, and (iii) the communication path between scheduled UE and activated DBSs are not blocked by blockages. If any of the three conditions are not met, then the one-to-one link between the transmission technique of CoMP and its benefits in a user-centric scenario cannot be established.

2.6.4 Optimal S-zone Radius and DBS Density for ASE

In Fig. 2.6 (a), the area spectral efficiency is plotted for different values of S-zone radius and M. We anticipate the existence of an optimal S-zone size at which the network-wide ASE is maximum. Further, the optimal S-zone radius to maximize ASE is expected to be smaller in magnitude as compared to S-zone size which maximizes EE. With the increase in the S-zone radius, the density of scheduled UEs reduces, thereby affecting the network ASE. Similarly, a decrease in the S-zone radius increases the density of scheduled UEs at the risk of spatially closer S-zones that increases interference levels. Therefore, as mentioned previously, the optimal S-zone radius should be a small value but not too negligible.

In Fig. 2.6 (a), we note that the optimal S-zone radius is 1.5 m which is slightly larger than the minimum considered S-zone radius of 1 m, hence, supporting our hypothesis. We also observe that the performance in terms of ASE is consistent across different M. This is mainly because ASE is dependent on the density of scheduled UEs, desired SIR threshold, and coverage probability. The desired SIR threshold does not change and the change in coverage probability is almost negligible when CoMP is employed. Therefore, the optimal S-zone size is not overly sensitive to the value of M unless the coverage probability change is significant between CoMP and no-CoMP scenarios.

In addition to S-zone radius, the density of DBSs that can maximize ASE is also an important design parameter from the perspective of a network operator. In Fig. 2.6 (b), the area spectral efficiency is plotted for different values of DBS density and M. From the figure, we can see that ASE increases monotonically with an increase in DBS deployment density for a fixed S-zone size. This is mainly because with larger DBS density, (i) the chances of coverage holes where no DBSs are available to provide service to a scheduled UE decrease, and (ii) there exist more options to activate the DBS(s) with strongest channel gains to further exploit spatial diversity. Also, for lower DBS density, employing CoMP significantly reduces the ASE, whereas, for higher DBS density, the performance in terms of ASE is consistent across different M.

Finally, the network ASE is plotted for different values of S-zone size and DBS density. For an optimal S-zone radius, ASE improves with the increase in the den-

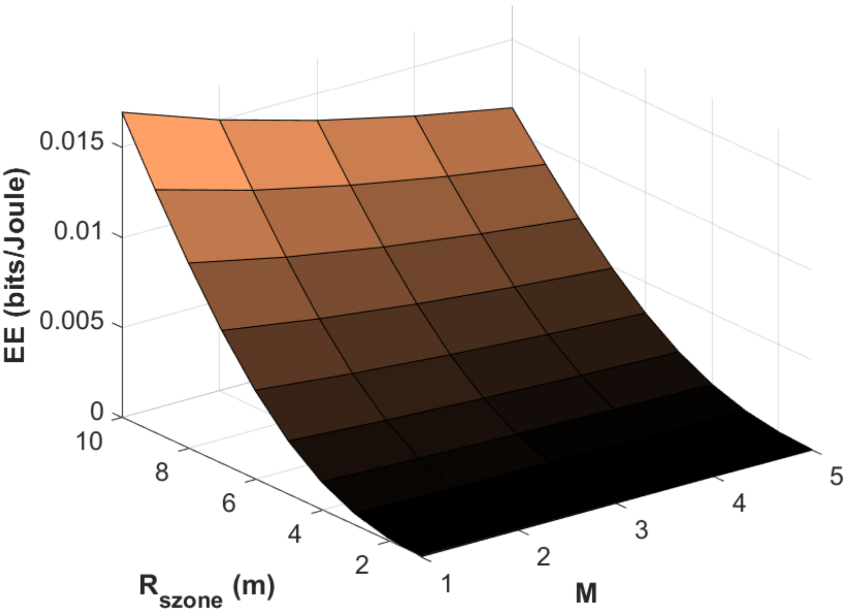


Fig. 2.7: Energy efficiency of the CoMP-enabled UC-RAN with varying S-zone radius and M for $\gamma_{th} = 4\ dB$.

sity of DBSs. Contrarily, given a fixed DBS density, ASE after an initial jump at $R_{szone} = 1.5\,m$ is decreased with the increase in S-zone radius. From these observations along with the observations reported in Fig. 2.6 (a) and (b), we can conclude that the radius of S-zone, DBS density, and the maximum number of cooperative DBSs greatly impact ASE. To maximize ASE, all of these inter-linked parameters should be jointly optimized through a self-organizing framework in subsequent chapters.

2.6.5 Optimal S-zone Radius for EE

In Fig. 2.7, the energy efficiency is plotted for different values of S-zone radius and M. The power consumption parameters are summarized in Table 2.2. Similar to ASE, the existence of an optimal S-zone size for which the EE will be maximum is obvious. However, the S-zone radius that will maximize EE is expected to be different than the S-zone radius that would maximize ASE. Intuitively, the S-zone radius that maximizes EE should be larger because by increasing the S-zone radius, the density of activated DBSs reduces resulting in lesser power consumption. From Fig. 2.7, we observe a similar trend where the EE is maximum for the largest S-zone radius. Also, for a fixed S-zone radius, enabling CoMP increases the power consumption of the network due to increased signal processing and additional power consumption overhead resulting in the degradation in EE of the system. Thus, we can easily conclude that EE degrades by enabling CoMP in UC-RAN architecture with non-overlapping S-zones.

These results further second the need for an AI-assisted self-organizing framework that can capture the tradeoff of ASE and EE to jointly maximize both KPIs given the new degrees of freedom such as S-zone radius and density of DBSs. Also, the results support the hypothesis presented in Section 2.1 that employing CoMP will not only affect the system's energy efficiency but will also negatively impact the area spectral efficiency as well as coverage probability in UC-RAN.

2.6.6 Performance Comparison for Mean Serviced UEs' ratio

Fig. 2.8 shows the comparison of the average number of UEs that are serviced out of the total scheduled UEs for varying values of S-zone radius, SIR desired threshold, and M. A UE is offered service iff: (i) there is at least 1 DBS present in the S-zone, and (ii) the average received SIR at the UE is greater than the desired SIR threshold. From the result shown in Fig. 2.8, a similar trend can be observed in the mean serviced UEs' ratio as observed in the coverage probability (scheme 1) with the enabling of CoMP. The decrease in the mean serviced UEs' ratio for $M > 1$ can be attributed to the increased interference in the network that affects the SIR received at a typical UE. We also observe that with an increase in SIR requirement threshold values (γ_{th}), the mean serviced UEs' ratio also decreases.

Additionally, the impact of R_{szone} on the mean scheduling UEs' ratio is demon-

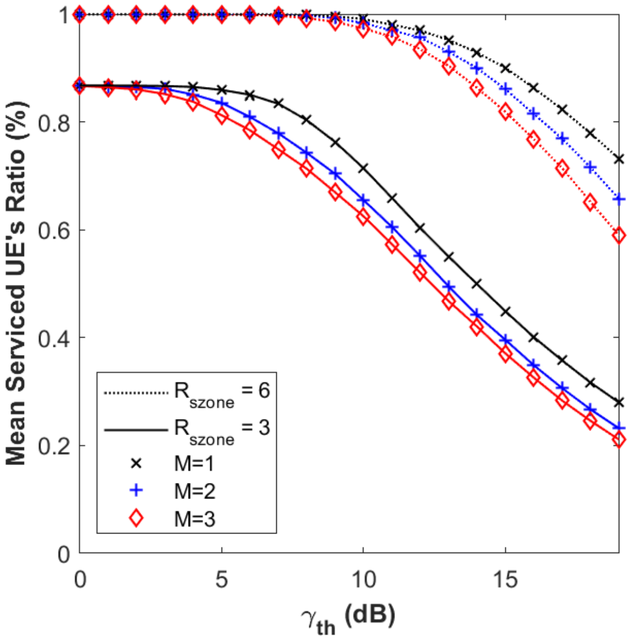


Fig. 2.8: UE servicing ratio comparison of proposed UC-RAN approach.

strated in Fig. 2.8. Intuitively, for a larger S-zone radius: (i) the inter-cell separation increases effectively reducing the interference at a typical UE, and (ii) the average number of DBSs available in an S-zone is increased effectively increasing the probability that there will be at least 1 DBS within the S-zone. For both of the above reasons, we observe a trend shown in Fig. 2.8 in which the mean service UEs' ratio increases when R_{szone} is increased from 3 to 6.

2.6.7 Performance Comparison with Traditional HetNet

Fig. 2.9 shows performance improvement in area spectral efficiency for the proposed CoMP-enabled UC-RAN architecture (with S-zone radius = $2\,m$) in comparison to a traditional heterogeneous network architecture discussed in [46]. From the figure, we can observe that there is a massive increase (of the order of x100 and more) in the ASE of the proposed UC-RAN approach as compared to the traditional HetNet. This is mainly because the traditional HetNet architecture can experience extreme inter-cell interference in a dense network due to its cell-centric architecture.

To overcome inter-cell interference, the proposed UC-RAN architecture not only provides a certain minimum separation between scheduled UEs but also provides dynamic coverage to each UE effectively resulting in higher ASE.

Another interesting observation from Fig. 2.9 is the existence of an optimal SIR threshold γ_{th} for which the ASE is maximized in CoMP-enabled UC-RAN. The reason is more mathematical rather than conceptual and can be better explained by referring to the expression given in Eq. 2.14. For the same R_{szone} and user density, the density of scheduled UEs remains constant. The ASE thus changes with fluctuations in the SIR threshold or coverage probability. Now as the SIR threshold increases, two interactions are happening simultaneously which affect the ASE. First, with the increase in the SIR threshold, the term $\log_2(1 + \gamma_{th})$ also increases resulting in increased ASE. Secondly, an increase in the SIR threshold, as observed in Fig. 2.4, reduces the coverage probability. Due to these contrasting effects, we observe a high jump in ASE values which then plummets as coverage probability approaches zero.

2.6.8 CoMP-enabled UC-RAN with Dynamic Blockages

After reviewing the simulation analysis hitherto, a reader may question the applicability of CoMP in user-centric networks where cell-edge users are diminished. Till now, we have shown that enabling CoMP for UEs that maintain a certain minimum repulsion with other UEs degrades the network performance in terms of important KPIs such as coverage probability, area spectral efficiency, and energy efficiency. However, there are many possible scenarios in which CoMP may be able to provide better reliability effectively enhancing the system's area spectral efficiency.

In this Section, we briefly discuss one such scenario where DBSs operate on high-frequency bands that are highly sensitive to blockages. We assume a simplified scenario in which the line of sight (LOS) link between DBSs and UEs is affected

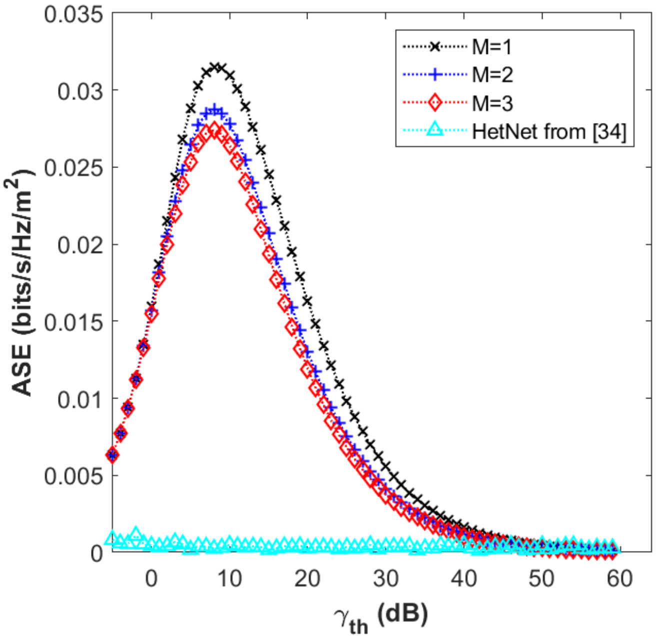
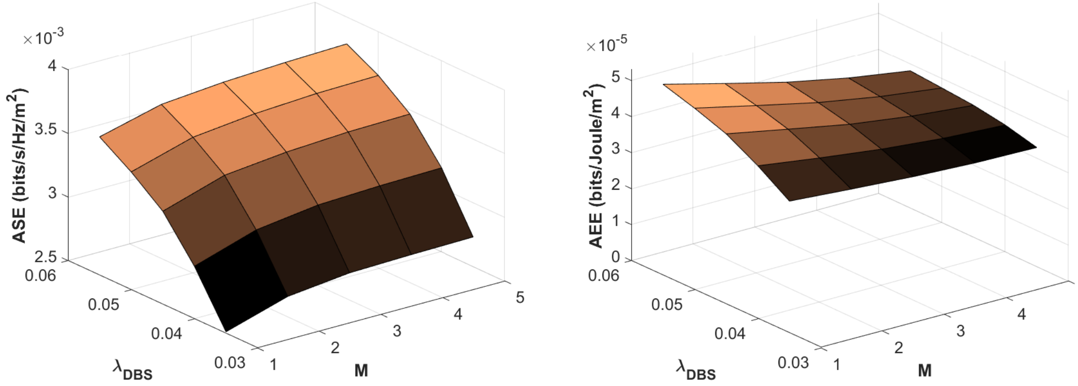


Fig. 2.9: ASE comparison of proposed UC-RAN approach and traditional HetNet for different values SIR requirements.



(a) Area spectral efficiency.

(b) Area energy efficiency.

Fig. 2.10: Performance comparison of CoMP-enabled UC-RAN with dynamic blockages with varying DBS densities and M for $\gamma_{th} = 4\ dB$, $R_{szone} = 6m$, $\alpha_{near} = 3$ and $\alpha_{far} = 6$.

by the presence of blockages. The presence of blockages divides the network region into two parts, near-field, and far-field regions. The near-field and far-field regions can be best described by a dual-slope path loss model (DSPM) that have different path-loss exponents α_{near} and α_{far} for near-field and far-field, respectively, where $\alpha_{far} \geq \alpha_{near} > 2$ [47]. Note that higher values of α_{far} will result in sufficiently large path-loss, effectively causing the interference caused by DBSs beyond a critical distance (d_c) to approach zero. Also, we assume that a non-LOS link within an S-zone may not be able to provide sufficient signal strength to the UE, thus changing the state of the DBS-UE link to outage when affected by an obstacle. The standard dual path-loss model is given as [48]:

$$DSPM(x) = \begin{cases} ||x||^{-\alpha_{near}}; & \text{with } 1 - p_{blockage}(x) \\ d_c^{\alpha_{diff}} ||x||^{-\alpha_{far}}; & \text{with } p_{blockage}(x), \end{cases} \tag{2.20}$$

where $\alpha_{diff} = \alpha_{far} - \alpha_{near}$, $d_c > 0$ is the critical distance assumed to be equal to 1 meter and x is the distance (in meters) between the DBS and UE.

The blockage/non-LOS probability proposed by 3GPP is given as [49]:

$$p_{blockage}(x) = 1 - \left(0.5 - \min\left(0.5, 5\exp\left(-\frac{156}{x}\right) + \min\left(0.5, 5\exp\left(-\frac{x}{30}\right)\right)\right)\right). \tag{2.21}$$

We expect that due to random blockages, the communication between serving DBSs and UEs will be highly affected. In such a scenario, CoMP will be able to provide second-tier protection from service degradation due to blockages. In Fig. 2.10 (a), the area spectral efficiency is plotted for different values of DBS densities and M. We observe that the network performance in terms of ASE improves by enabling CoMP. For instance, as the network shifts from a non-CoMP mode to a CoMP enabled UC-RAN (i.e. from $M = 1$ to $M = 2$), we observe approximately 8% increase in ASE. This is mainly because if the closest UE-DBS link within an S-zone is affected by a blockage, there is a very low probability that the second closest UE-DBS link will also be affected by a blockage, thus ensuring the service reliability constraints at a typical UE. Besides, the dual-slope path loss model ensures that most of the interference dies out beyond the S-zone region. These two notions in parallel provide improvement in the ASE of the system with the enabling of CoMP.

An interesting observation is the consistency of area spectral efficiency values for any value of $M > 1$. There is an almost unnoticeable improvement when M is increased from beyond 2 as also noticed in [50]. This is mainly linked with the blockage probability considered in the network. In an environment, where it is highly probable that 2 or more serving DBSs can be simultaneously affected by blockages, we may be able to observe a noticeable increase in the ASE with an increase in M beyond 2. The change in the area spectral efficiency values concerning DBS density is due to the reasons discussed in Section 2.6. Since $M = 1$ reflects the scenario of UC-RAN with single DBS activation, the result in Fig. 2.10 (a) shows that when random blocking is considered, the CoMP-based UC-RAN with multiple DBS activations within an S-zone outperforms our earlier works [3, 4].

The performance in terms of area energy efficiency (AEE), which is defined as the ratio of area spectral efficiency and power consumption, is shown in Fig. 2.10 (b) for different values of DBS density and M. Intuitively, by enabling CoMP, the AEE should decrease because of higher power consumption by the activated DBSs. The results reveal that AEE decreases as more DBSs are enabled for cooperation. However, due to the dependence of AEE on ASE, the major drop in AEE occurs when M is increased from 1 to 2. The decrease in AEE values when M is changed from 1 to 2 is approximately 3.5%, whereas the increase in ASE is approximately 8%. Therefore, there is a visible trade-off between ASE and AEE values based on the number of cooperative DBSs that the network operators can utilize to design the network based on the service requirements of the users.

2.7 Conclusion

In this chapter, we provided an analytical and numerical analysis on the impact of enabling CoMP in a UC-RAN architecture. Contrary to the existing literature on analytical models of CoMP in UC-RAN, we derived closed-form analytical expressions for activated data base station density, aggregate interference, coverage probability, area spectral efficiency, and energy efficiency. For CoMP, we presented a comparative analysis of three joint DBS transmission methods via cooperative DBSs in a user-centered virtual cell or an S-zone.

Through our analysis which was supported by extensive Monte Carlo simulations, we showed that employing CoMP in a UC-RAN architecture, with non-overlapping S-zones that uses spatial diversity for DBS activation with no blockages in the wireless communication channel, not only reduces the energy efficiency of the network but also degrades the coverage probability at a typical UE and consequentially the network-wide area spectral efficiency. However, we also discussed scenarios, in particular highly blockage sensitive propagation, where the proposed design offered an improved area spectral efficiency. The analysis presented in this chapter provides a baseline on the design and planning of futuristic UC-RAN based cellular networks.

We also investigated the impact of new degrees of freedom such as S-zone size and density of data base stations on the mean serviced UEs, area spectral efficiency, and energy efficiency of the network. The numerical results based on the derived analytical model revealed an interesting interplay between S-zone size and the density of data base stations. It is observed that for any number of cooperative data base stations in an S-zone, there exists an optimal size of S-zone and DBS density that maximizes the area spectral efficiency, and energy efficiency. However, the values of optimal S-zone size and data base station density that maximizes area spectral efficiency are quite different from the values that maximize network-wide energy efficiency. Therefore, there is a need for an artificial intelligence-assisted self-organizing framework proposed in the next chapter that is capable of dynamically orchestrating these network design parameters to offer the ideal tradeoff between these KPIs for a network operator.

CHAPTER 3

D-RAN: A Deep Reinforcement Learning-based Demand-Driven Elastic User-Centric RAN Optimization for 6G and Beyond

3.1 Introduction

3.1.1 Motivation

6G networks are envisioned to cater to a wide range of user services with assorted throughput and latency requirements [51]. In order to meet this requirement, there is a need for an elastic architecture that can tailor to the needs of each service, as opposed to traditional one-size-fits-all architecture. This, along with the interference-limited nature of UDNs, has prompted a shift to a UC-RAN paradigm from traditional networks. [52–54]. Although the existing literature on UC-RAN (discussed in the related work section in 1.3) provides some useful information, it has two shortcomings. First, it deals with static S-zone size for all UEs with the assumption that all UEs will have similar throughput and latency requirements which is not a practical assumption. Second, although the analytical models in above studies are highly detailed, they lack the interaction of controlling parameters (S-zone) with the spatiotemporal changes in the wireless network such as dynamic user application demands and mobility.

Recognizing the significance of user-centric services in future cellular communications, particularly in 6G, this chapter introduces an elastic and demand-driven UC-RAN model, as outlined in Section 3.2. We formulate a multi-objective optimization problem to maximize important KPIs such as area spectral efficiency, network energy efficiency, user service rate, and throughput satisfaction. The S-zone size serves as a control parameter to form a Pareto-optimal trade-off among these KPIs.

The core research objective of this chapter is to develop a solution that can dynamically solve this multi-objective optimization problem in UC-RAN to achieve a Pareto-optimal solution in real-time based on changes in the varying application demands and user mobility. Inspired by our earlier work on utilizing wireless network telemetric big data for enabling zero touch optimization in future wireless networks [55], we propose a deep reinforcement learning (DRL)-based framework to solve this problem. This framework is hereafter referred to as D-RAN: a deep reinforcement learning-based user-centric RAN optimization framework under dynamic user application demands and network conditions. D-RAN uses the massive amount of control, signaling, and contextual data in UC-RAN network to update network parameters dynamically to optimize the KPIs of interest in real-time.

Driven by the above motivations, this chapter studies the deep reinforcement learning approach owing to its ability to adapt to dynamic environments to determine the optimal S-zone size for each QoS category intelligently so that network KPIs such as area spectral efficiency, energy efficiency are maximized as well as throughput, and latency requirements of each QoS category are met.

3.1.2 Contributions

Specifically, the contributions of this chapter are summarized as follows.

- An architecture for demand-driven elastic user-centric communication is proposed with the aim of providing on-demand services to a diverse set of user applications ranging from augmented/virtual reality to industrial robots to E-health applications, and more. The proposed architecture allows the elastic user-centered S-zone to be malleable to specific QoS category requirements.

- Considering the heterogeneous user requirements in future cellular communications, a multi-objective problem is formulated to optimize KPIs such as area spectral efficiency, energy efficiency, user service rate, and throughput satisfaction as a function of S-zone size for respective QoS categories. Given the stringent requirement of very high throughput and ultra-low latency, the multi-objective problem is geared towards meeting users' throughput and latency requirements while also maximizing the area spectral efficiency and network energy efficiency.

- Given the non-stationarity of user application demands and mobility, we propose a deep reinforcement learning framework to accurately learn the mapping of environment state and action instilling intelligence in the demand-driven elastic user-centric architecture. The proposed intelligent deep reinforcement learning framework for UC-RAN networks, named D-RAN, dynamically allocates S-zones to users such that a Pareto-optimal front is found for the formulated multi-objective function.

- We evaluate the convergence, efficacy, and adaptability of D-RAN to the non-stationary environment of the proposed approach through numerical results. We also compare D-RAN's performance against brute-force and state-of-the-art metaheuristics such as simulated annealing. The simulation results show that D-RAN can achieve a gain of up to 45% in the network-wide utility compared to an simulated annealing-based solution. The proposed framework has the potential to change network mode from rigid cell-centric to elastic user-centric through the use of an intelligent module (D-RAN) that allows optimization of S-zones in real-time, resulting in enhanced user experience, greater system capacity, and improved energy savings.

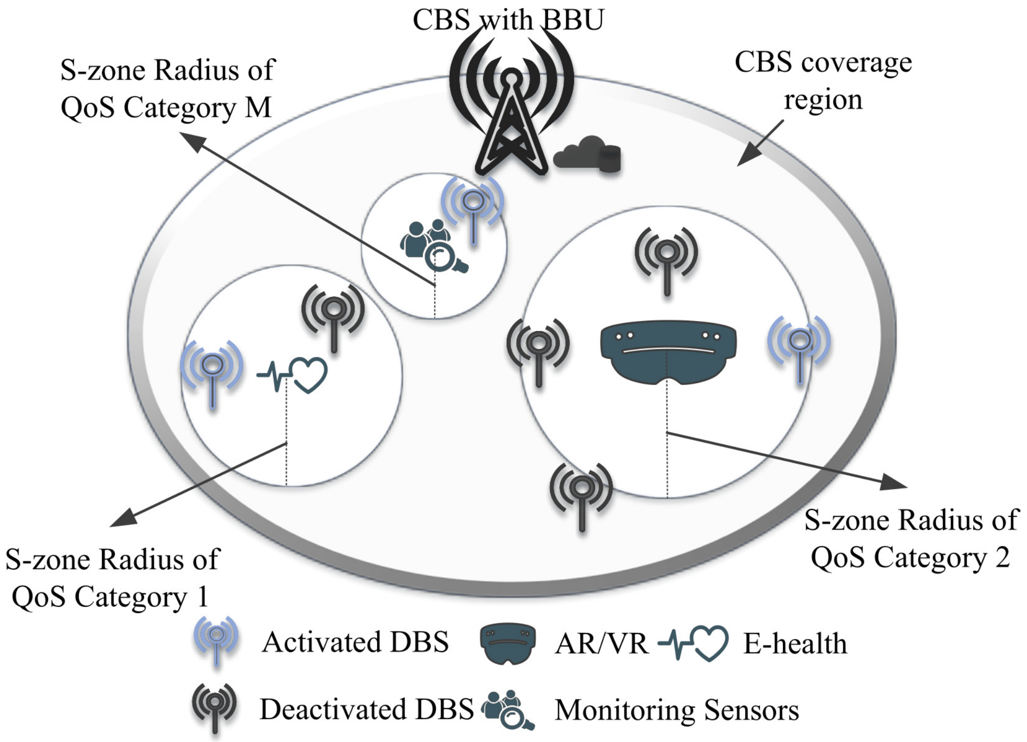

CBS with BBU
S-zone Radius of QoS Category M
CBS coverage region
S-zone Radius of QoS Category 2
S-zone Radius of QoS Category 1
Activated DBS
AR/VR
E-health
Deactivated DBS
Monitoring Sensors

Fig. 3.1: Dynamic S-zone UC-RAN architecture with M different S-zone region of radius R_c for scheduled UE's.

3.1.3 Chapter Organization

The remainder of the chapter is organized as follows. The system model is discussed in Section 3.2. A multi-objective optimization problem as a function of S-zone size is formulated in Section 3.3.2. A brief summary of deep reinforcement learning and simulated annealing algorithms are presented in Section 3.4. The details of the proposed approach and the results of the numerical analysis are presented in Section 3.5 and Section 3.6, respectively. Finally, the chapter is concluded in Section 3.7.

3.2 System Model

This section presents the UC-RAN architecture, S-zones scheduling algorithm, network model, and channel model.

3.2.1 UC-RAN Architecture with QoS Category Specific S-zones

Fig. 3.1 provides a graphical illustration of a UC-RAN network with virtual user-centric cell boundaries for UEs belonging to different QoS categories. These categories are classified according to the UEs' latency and throughput requirements as illustrated in Fig. 3.1.

A critical design parameter in UC-RAN is the size of S-zone which is defined by the radius of circular disk around the UE. In the proposed model, the DBSs falling within the S-zone of a UE are only allowed to associate with that UE in a given TTI. Increasing the S-zone size ensures (i) larger distances between a UE and interfering DBSs resulting in high link-level SINR (hence, link-level high throughput and spectral efficiency); (ii) yields high macro diversity gain through selection among the larger number of DBSs in the S-zone and (iii) offers high energy efficiency as large S-zones keep more DBSs deactivated as compared to small S-zones. However, larger S-zones also yield low user scheduling ratio and low spectrum reuse resulting in negative impact on the system-level capacity. Given these insights, the S-zone size serves as a controlling parameter that yields an ideal tradeoff between area spectral efficiency, energy efficiency, and other system-level KPIs.

In UC-RAN, a scheduled user in each TTI is allocated the full bandwidth of the system for two reasons: i) to make the system capable of providing maximum throughput to a user that the total system bandwidth allows; ii) to keep the radio resource scheduling at DBS simple and thus keep DBS cost and energy consumption low. The spectrum waste is avoided by managing the temporal scheduling where a user needing a lower throughput is scheduled after larger number of TTIs. The temporal gap in TTIs after which a user is scheduled is inversely proportional to user bandwidth/throughput requirement.

Besides, to make the spectrum allocation more efficient, there is a need to intel-

Algorithm 2: UE Scheduling Algorithm

```
Initialize the set of UEs and the DBS(s) ;
Assign priorities to UEs based on their latency requirements ;
Sort UEs in the descending order according to their priorities ;
for each UE in the sorted list do
    if DBS available in S-zone region of UE and UE is not overlapping with
      other scheduled UEs then
        Schedule UE
```

ligently allocate both physical resource blocks and S-zone size to scheduled users according to their needs. Since the D-RAN framework is proposed mainly to establish that the S-zone size of multiple QoS categories (with varied QoS demands) can be intelligently controlled to optimize the desired KPIs, the joint optimization of S-zone size and physical resource blocks will be addressed in future research. It is also important to mention that the intelligent allocation of physical resource blocks in 5G cellular systems has already been proposed in several publications [56, 57].

3.2.2 UE Scheduling Algorithm

In this chapter, we propose a scheduling mechanism to meet the heterogeneous latency requirement of UEs in UC-RAN. Latency requirements of UEs are drawn from a uniform distribution and rounded off to specified bins of latency requirements corresponding to the QoS categories. Each UE x is marked with $p_x^{latency} \sim U(a, b)$ by the BBU where a and b are measured in milliseconds (ms) and are determined by the minimum and maximum latency of the considered QoS categories. The lower the value of mark $p_x^{latency} \sim U(a, b)$, the higher will be the scheduling priority.

The BBU based on these scheduling priorities schedules a UE x if and if only the scheduling priority of UE x is highest in the neighborhood which is characterized by the S-zone size R_c for a specific QoS category. This means that within a circle of radius R_c centered at UE x, no other UE has a higher priority than UE. For example, the scheduled UEs shown in Fig. 3.2 have a lower latency requirement

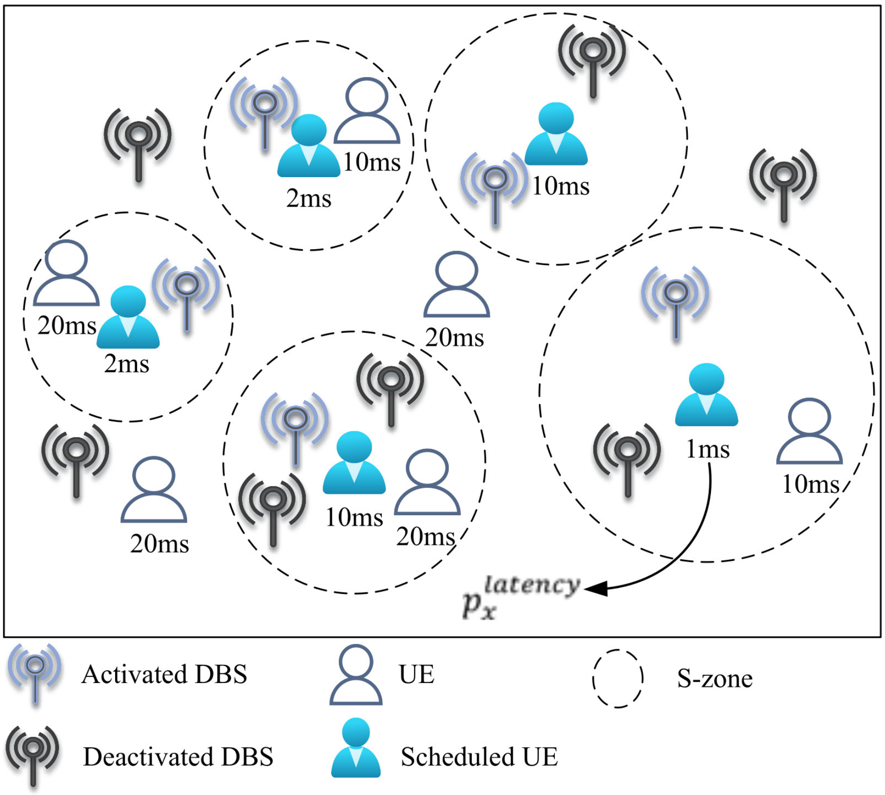


Fig. 3.2: Graphical illustration of UEs scheduling with varied latency requirements.

than any other UE in the S-zone of the respective QoS category. Note that larger the S-zone size of QoS categories, lesser the number of UEs will be scheduled with non-overlapping S-zones.

Once the UE is scheduled, a single DBS providing the highest channel gain within the S-zone of the respective UE is activated by the BBU to serve the UE. It is important that the DBSs are deployed densely, so at least one DBS is available within an S-zone to provide coverage to a scheduled UE and thus avoid coverage holes in areas where no DBSs are available within the user-centric circular disk.

3.2.3 Network Model

A downlink of a two-tier UDN is considered consisting of a CBS and DBSs operating on sub 6 GHz frequencies. The DBSs and UEs are randomly distributed following two independent and homogeneous Poisson point processes Π_{DBS} and Π_{UE} with intensities λ_{DBS} and λ_{UE} respectively. The location of each UE acts as a centering point for the user-centric virtual cell (S-zone) which bounds the UE to be associated with DBS only within the S-zone region. This implies that each DBS can at most serve a single UE. This chapter defines the S-zone as a disk of radius R_c, where $c \in C$ is a QoS category present in the network model. The network model in Fig. 3.1 for example, includes three QoS categories: augmented/virtual reality, E-health, and monitoring sensors.

3.2.4 Channel Model

The communication channel between an arbitrary user $x \in \Pi_{UE}$ and activated DBS $i \in \Pi'_{DBS}$ is modeled to experience both large-scale and small-scale fading given by hl^{-PLE}, where h is an exponential random distribution with unit mean, l_{xi} represents the propagation distance between x and i, PLE is the pathloss exponent, and Π'_{DBS} is the Poisson point process of activated DBSs. UE and DBS are equipped with a single antenna and the transmission power of DBS is assumed to be equal. Each scheduled user is served by a DBS providing the highest channel gain within an S-zone of radius R_c whose SINR (Γ_x) is given as:

$$\Gamma_x = \frac{h_{xi} l_{xi}^{-PLE}}{\sum\limits_{j \in \Pi'_{DBS}} h_{xj} l_{xj}^{-PLE} + n_o}, \tag{3.1}$$

where $i \neq j$ and n_o denotes the additive white Gaussian noise.

3.3 Problem Formulation

This section characterizes the KPIs, followed by the formulation of a multi-objective optimization problem.

3.3.1 Characterizing Key Performance Indicators

This chapter measures system performance in terms of area spectral efficiency, network energy efficiency, user service rate, and throughput satisfaction as the desired set of KPIs. We selected these KPIs to reflect that the objective is to meet throughput and latency requirements while maximizing area spectral efficiency and network energy efficiency.

Area Spectral Efficiency

The area spectral efficiency refers to the amount of information that can be transmitted from a DBS per unit bandwidth channel per unit area to a UE, which can be defined as follows for each QoS category c:

$$\mathsf{A}_c = \frac{\sum_{x \in N_c} \log_2(1 + \Gamma_x)}{\mathring{\mathrm{A}}}, \tag{3.2}$$

where N_c is the set of UEs belonging to QoS category c, and $\mathring{\mathrm{A}}$ is the target area considered in the simulations model. The formulation of area spectral efficiency in this chapter differs from that in Chapter 2 as we consider the area spectral efficiency for each QoS category separately.

There is a strong relationship between the QoS category's S-zone size and area spectral efficiency [4, 7]. Intuitively, increasing the S-zone size decreases the scheduling ratio of UEs. Nevertheless, decreasing the S-zone size increases the SINR (due to the higher number of neighboring interfering DBSs). There is, therefore, an optimal size for S-zones that balances these two opposing effects to maximize the attainable area spectral efficiency. To optimize the area spectral efficiency, intelligent real-time optimization is needed to calibrate the S-zone size of multiple QoS categories simultaneously.

Table 3.1: D-RAN power consumption parameters.

Symbol	Parameter Name	Parameter Value
P_f	DBS fixed power consumption	1.932 W
P_{DBS}	DBS transmit power	10 W
Δ_{DBS}	Radio frequency component's power consumption at DBS	23.22 W
P_{UE}	UE transmit power	1 W
Δ_{UE}	Radio frequency component's power consumption at UE	4 W
P_{disc}	UE cell discovery circuit power consumption	4.3 W

Energy Efficiency

According to [4, 58], the network-wide energy efficiency is defined as the ratio of area spectral efficiency and total power consumed for all scheduled UE's. The power consumption model in this chapter is inspired by project Earth [42], in that it represents the power consumption of CBS and DBSs as a linear combination of fixed power and load-dependent power consumption components. Since energy efficiency is measured network-wide, these power consumption values are summed for all scheduled users. The total power consumption can be mathematically calculated as follows:

$$P = \lambda_{DBS} P_f + \lambda'_{DBS} \Delta_{DBS} P_{DBS} + \lambda'_{UE} (\Delta_{UE} P_{UE} + P_{disc}), \tag{3.3}$$

where λ_{DBS} is the density of all deployed DBSs, λ'_{DBS} is the density of activated DBSs, λ'_{UE} is the density of scheduled UEs, P_f is the fixed DBS power consumption required for DBS to operate in listening mode, P_{DBS} is the DBS transmission power, Δ_{DBS} is the radio frequency component power at DBS, P_{UE} is the UE transmission power, Δ_{UE} is the radio frequency component power at UE, P_{disc} is the power required at UE for discovery of the DBS with the highest channel gain. The typical values of these variable are summarized in Table 3.1 [4]. The energy efficiency therefore can be given as:

$$\mathsf{E} = \frac{\text{Å} \times \sum_{c \in C} \mathsf{A}_c}{P}. \tag{3.4}$$

In a cellular DBS, radio frequency components and data transmission account for the majority of total power consumption [59]. DBSs can save significant amounts of energy when they are dynamically activated, particularly in dense deployments. The direct relationship between energy efficiency and area spectral efficiency mandates that the S-zone size of QoS categories will also influence network energy efficiency. Intuitively, increasing the S-zone size decreases the number of activated DBSs (decreasing the average power consumption). The contrasting trends of area spectral efficiency and power consumption raise an important design question: what S-zone size should be selected for QoS categories to optimize network-wide energy efficiency.

UE Service Rate

The UEs' heterogeneous latency requirements necessitate scheduling more UEs within each TTI while meeting UE quality of experience requirements. The mean UE service rate (user service rate) for any QoS category c can be calculated as:

$$\mathsf{U}_c = \frac{\lambda_{UE_c}^{service}}{\lambda_{UE_c}}, \tag{3.5}$$

where λ_{UE_c} is the density of all UEs belonging to QoS category c and $\lambda_{UE_c}^{service}$is the density of UEs belonging to QoS category c whose minimum throughput requirement is met.

The S-zone size of QoS categories influences the user service rate in two different ways. A decrease in the S-zone size leads to the scheduling of more users. However, decreasing the S-zone size also increases the average distance between UE and DBS, thus, affecting the average SINR. Due to these contrasting results with the change in S-zone size, we anticipate that optimizing user service rate will require intelligent optimization of S-zone sizes of QoS categories.

Throughput Satisfaction

There can be a wide variety of throughput requirements for UEs belonging to different QoS categories. Operators must satisfy the minimum throughput requirements of each QoS category as part of their objective. Moreover, network operators must ensure that they are utilizing their resources efficiently by avoiding scenarios in which excess throughput is allocated to a few UEs (or categories of UEs) while other UEs' minimum requirements are not met. For this reason, this chapter uses the difference between required and obtained throughput, a metric we define as throughput satisfaction (throughput satisfaction), to measure system performance. Throughput satisfaction for a specific QoS category c is given as:

$$\mathsf{T}_c = \prod_{x \in N_c} \left| tp_x^{\star} - tp_x^{\Diamond} \right|^{|N_c|}, \tag{3.6}$$

where $tp_x^{\star}$ and $tp_x^{\Diamond}$ are the obtained and required throughput for an arbitrary UE x respectively. The required throughput values for UEs are drawn from a uniform distribution and rounded off to specified bins of throughput requirement of QoS categories. While the obtained throughput values are obtained by mapping the SINR values of UEs to its physical layer throughput given in [60].

Intuitively, the increase in S-zone size of QoS category is expected to improve the average SINR (and throughput obtained) at the UE. However, the mere increase in throughput of a few users is not the desired behavior. Instead, the S-zone size should be adjusted such that the throughput achieved at UEs belonging to a QoS category float near the throughput requirement of that specific QoS category. This entails that the S-zone size of QoS categories should be carefully calibrated to ensure satisfaction is achieved throughput across all QoS categories.

3.3.2 Multi-objective Optimization Problem Formulation

Hitherto, the above definition of KPIs demonstrate the need for optimizing S-zone size of QoS categories to maximize area spectral efficiency, energy efficiency, UE service rate and throughput satisfaction individually. The challenge from a network operator's perspective is that all these KPIs should be optimized simultaneously, leading to a Pareto-optimal tradeoff between them. To account for this tradeoff, this chapter defines the multi-objective optimization problem as follows:

$$\max_{R_c, c \in C} \quad \frac{\left(\sum_{c \in C} \mathsf{A}'_c\right)^{\alpha} \left(\sum_{c \in C} \mathsf{U}'_c\right)^{\beta} \left(\mathsf{E}'\right)^{1-\alpha-\beta}}{\sum_{c \in C} \mathsf{T}'_c} \tag{3.7}$$
$$\text{s.t.} \quad R_{min} \leq R_c \leq R_{max}; \forall c,$$

where $0 \leq \alpha$, $\beta \leq 1$, $\alpha + \beta \leq 1$, A'_c is area spectral efficiency normalized between $[0, 1]$, E' is energy efficiency normalized between $[0, 1]$, U'_c is UE service rate normalized between $[0, 1]$, T'_c is throughout satisfaction normalized between $[1, 2]$, R_{min} and R_{max} are the minimum and maximum allowable size for S-zone of QoS categories. To bring it to the reader's attention, throughput satisfaction is included in denominator to ensure that the increase in the difference between required and obtained throughput of UEs reduces the utility of the solution.

The rationale behind the proposed objective function formulation is to optimize holistic system-level performance by combining network operators' four most important and common KPIs of interest. However, these KPIs have different scales/units. This issue makes combining the multiple KPIs in a single objective function far from a straightforward problem. In this chapter, we address this problem by normalizing each KPI value with its minimum and maximum value. These minimum and maximum KPI values are determined through pseudo brute force method. The pseudo brute force method sweeps the solution space (with a pre-defined step size) in numerous independent runs. Given the step sizes are large enough to explore the possible extrema in the search space within an affordable computational effort, this pseudo brute force method gives values of KPIs that can be taken as approximation of minimum and maximum values for the normalization purposes. This way of approximating the true Pareto optimal front is quite common in general reinforcement learning problems [61].

The solution obtained from the pseudo brute force search is then used to linearly scale/normalize the value of each KPI, allowing the effective KPIs to be unitless and combined in a multi-objective optimization problem. The real goal of the system is to maximize the area spectral efficiency, energy efficiency, and user service rate while keeping the gap between target and achieved throughput values minimum. To be reflective of the real goals of the system, Eq. 3.7 is designed such that the normalized values of area spectral efficiency (between 0 and 1), energy efficiency (between 0 and 1), and user service rate (between 0 and 1) are multiplied in the numerator to jointly maximize these KPIs while the normalized value of throughput gap (between 1 and 2) is included in the denominator to minimize the difference between throughput obtained and achieved by the users. This gap-based formulation to model user satisfaction, instead of simple threshold based KPI where throughput is maximized for some users without a cap, is used as a clever way to avoid wasteful resource allocation. Compared to alternative simpler formulation where all KPIs are maximized as linear sum or product, this formulation is chosen to minimize intrinsic conflict QoS KPIs has with other two KPIs of area spectral efficiency and energy efficiency.

These four KPIs are representative of one of the four key aspects of network performance, either at the network level or user level. For instance, area spectral efficiency is representative of network spectral efficiency, energy efficiency is representative of network energy efficiency, user service rate is representative of scheduling maximum users while satisfying a certain data rate requirement, and throughput satisfaction is representative of meeting specific user throughput requirements. Note that using these many KPIs is not common in academia due to the intractability of the analytical models with complex multi-objective optimization functions. However, optimizing tens of KPIs simultaneously is a standard practice in real-time network optimization.

With the formulated optimization problem, a BBU controls the S-zone size of QoS categories such that desired KPIs (area spectral efficiency, energy efficiency, user service rate, and throughput satisfaction) are optimized while keeping the S-zone size within a specified range of R_{min} and R_{max}. The problem in Eq. 3.7 is a mixed-integer nonlinear programming problem with complexity of the order of $\mathcal{O}\big((R_{max} - R_{min} + 1)^{|C|}\big)$. It is computationally difficult to achieve an optimal solution for a non-convex multi-objective problem in a dynamically changing network, which makes its application in real-time optimization systems impossible.

The legacy approaches to address such problem relies either on analytical modeling, or simulation-based modeling or more recently data-driven modeling. Our choice to leverage deep reinforcement learning instead of aforementioned approaches is motivated by its superiority to all three alternatives for the particular problem in hand. This superiority stems from the following reasons. Deep reinforcement learning-based framework is better than analytical model-based framework due to its ability to capture network dynamicity and complexity that analytical models miss to achieve due to the abstraction needed to obtain tractility. Compared to a simulator model-based offline optimization approach, deep reinforcement learning can tune optimization parameters of interest using live responses that reflect real-network behavior instead of an offline simulator behavior. Finally, deep reinforcement learning is advantageous compared to pure data-driven model-based optimization (e.g., using deep learning) as deep reinforcement learning does not require deluge of data that would be required to train a complex system-level network behavior data-driven model for performing the optimization. To this end, this chapter proposes a D-RAN (DRL-based) framework that is capable of determining the optimal S-zone size for all QoS categories with the objective of maximizing network KPIs.

3.4 Preliminaries

The following section gives a primer on deep reinforcement learning and simulated annealing algorithms.

3.4.1 Deep Reinforcement Learning

In a general reinforcement learning (RL) problem, an agent takes an action by observing the state from the environment and receives a scalar reward in an iterative manner. An RL agent aims to maximize the future cumulative rewards for different states of environments to learn the best course of action. Based on the specified set of actions, the RL algorithm generates a mapping between these actions and environment states. An implementation of RL includes these elements:

- *Observations*: Observations $\mathbf{O} \in \mathbb{R}^p$ are a set of measurements provided by the environment where p indicates the number of measurements observed.

- *States*: States $\mathbf{s}^t \in \mathcal{S}$ are a subset of observations vector observed at each epoch t either through handcrafted or non-handcrafted features where an epoch is a discretized time interval, signifying a single forward or backward pass of training samples.

- *Actions*: Actions $\mathbf{a}^t \in \mathcal{A}$ are a discrete/finite set of allowed choices that an RL agent can send to the environment as an input at each epoch t. Ideally, the choice of action should have an influence on the state of the environment such that the input of action changes the state of the environment from $\mathbf{s}^t$ to $\mathbf{s}^{t+1}$.

- *Policy*: A policy $\pi(\mathbf{s}, \mathbf{a})$ is the mapping between the state of the environment and an agent's action.
- *Value function*: The value function (also called Q-function) under a given policy is given as $Q_\pi(\mathbf{s}, \mathbf{a})$ which represents the discounted future expected return for a state-action pair. The value function determines the value of being at a particular state and taking a specific action at that state [62].
- *Rewards*: The reward signal $r^{t+1} \in \mathbb{R}$ is a scalar value returned by the environment when an action $\mathbf{a}^t$ influences the state of the environment from $\mathbf{s}^t$ to $\mathbf{s}^{t+1}$.

These elements in conjunction drive the RL agent to maximize the future cumulative reward which is given as:

$$G = \sum_{t=0}^{\infty} \gamma^t r^{t+1}, \tag{3.8}$$

where $\gamma \in [0, 1]$ is the discount factor. Through iterative updates, the Q-function values are estimated using the Bellman equation in a traditional Q-learning algorithm:

$$Q^{t+1}(\mathbf{s}^t, \mathbf{a}^t) = (1 - \kappa)Q^t(\mathbf{s}^t, \mathbf{a}^t) + \kappa(r^{t+1} + \gamma \max_{\mathbf{a}} Q^t(\mathbf{s}^{t+1}, \mathbf{a}^{t+1})), \tag{3.9}$$

where $\kappa \in (0, 1]$ is the learning rate.

It is theoretically proven that Q-Learning algorithms converge under certain conditions [62]. However, the drawback of Q-learning is that it requires the agent to store a matrix of the size of state space times the size of action space, which is impossible for most real-world problems. To assuage that, deep neural networks are utilized in RL algorithms, to act as universal Q-function approximators and learn the handcrafted features representation. The input dimension of deep reinforcement learning represents the number of states in the state space $|\mathcal{S}|$, while the output dimension represents the number of possible actions $|\mathcal{A}|$. The loss function with θ_Q as the trainable weights is used to train deep reinforcement learning is given below [63]:

$$\mathcal{L}(\theta_Q) = \mathbb{E}\big[r^{t+1} + \gamma \max_{\mathbf{a}^{t+1}} Q^t(\mathbf{s}^{t+1}, \mathbf{a}^{t+1}|\theta_Q) - Q^t(\mathbf{s}^t, \mathbf{a}^t|\theta_Q)\big]^2. \tag{3.10}$$

3.4.2 Simulated Annealing

The simulated annealing technique approximates the global optimum of nonlinear and non-convex objective functions by a series of iterative searches. Simulated annealing methodology is cognate to metallurgical annealing in which a metal is heated to a specific temperature before slowly cooling it down. Simulated annealing begins its global optimum search with a very high-temperature parameter $Temp$, which enables it to explore a relatively wide area and then decreases the temperature, progressively narrowing the exploration area as it iteratively follows the steepest descent.

A fitness function associates a fitness value to each solution depending on the objective function. In each iteration, simulated annealing compares the fitness value of the current solution to the solutions that are available in the local neighborhood W. If the neighboring solution has a higher fitness value than the current solution, then the neighboring solution is chosen for the next iteration. The simulated annealing uses an acceptance probability to avoid adhering to a local optimum. The acceptance probability is given as follows [64]:

$$\text{Acceptance Probability} = \exp\Bigg(-\frac{F_{curr} - F_{neig}}{Temp}\Bigg), \forall \ neig \in W, \tag{3.11}$$

where F_{curr} represents the fitness value of current solution.

3.5 Proposed Solution

This section discusses the design of the proposed D-RAN framework. The multi-objective problem formulated in Eq. 3.7, even though a mixed-integer nonlinear programming problem with high complexity, can be solved using various optimization techniques including DRL-based approaches and meta-heuristics such as simulated annealing. To compare the effectiveness of proposed D-RAN framework (DRL-based approach) to a meta-heuristic approach, we have included a simulated annealing solution. As simulated annealing is also known to yield near optimal solutions for optimization problems [65], it offers a benchmark to evaluate the performance of the proposed D-RAN framework. A BBU implements the optimization agent, which collects the network parameters and specifies the S-zone size for each QoS category. This centralized implementation facilitates the independence of processing times from UE and DBS densities, thus allowing for practical realizability and scalability of the optimization framework.

3.5.1 D-RAN Framework

A D-RAN framework is described in detail in terms of state space, action space, reward function, and the procedure of agent training and testing.

State Space

Section 3.3.1 establishes the linkage between the S-zone size of QoS categories and KPIs considered in this chapter. These KPIs define the state of the environment which if probed further can be decomposed into three parts:

- The average SINR of each QoS category is impacted by the change in S-zone size of QoS categories as divulged in Eq. 3.1, which has an impact on the

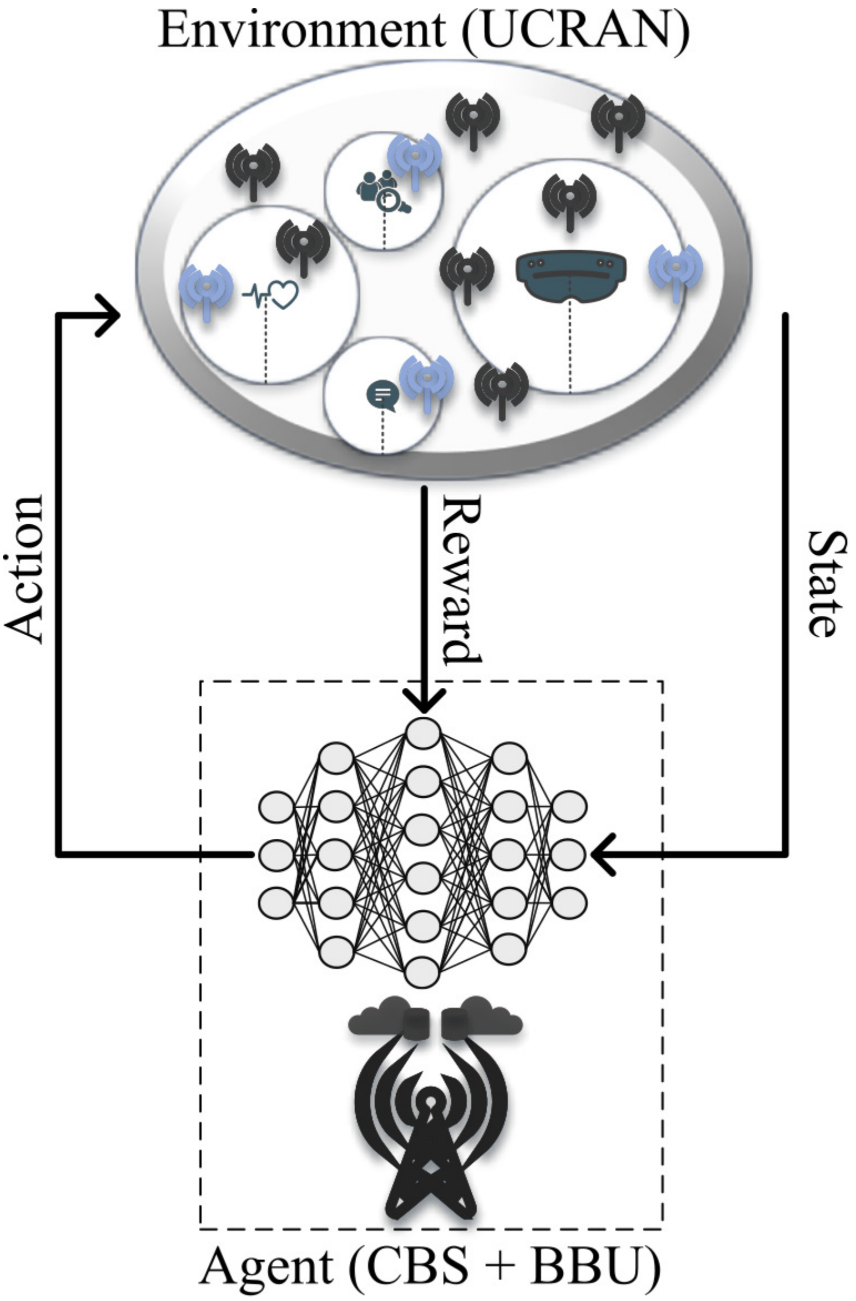


Fig. 3.3: Block diagram of the proposed D-RAN framework.

area spectral efficiency, energy efficiency, user service rate, and throughput satisfaction. Increasing the S-zone size is expected to increase the average SINR inherently for two reasons: (i) a large S-zone yields a large minimal separation gap and hence reduction in interference between a scheduled UE and nearest interfering DBS; and (ii) a larger S-zone should lead to a higher macro-diversity gain due to selection among the larger number of DBSs in the S-zone. However, average SINR's impact on the listed KPIs makes it a suitable choice for defining environment state. The average SINR of each QoS category can be given as:

$$\varphi_c = \frac{\sum\limits_{x \in N_c} \Gamma_x}{|N_c|}, \forall c \in C. \tag{3.12}$$

- The user service rate of each QoS category given in Eq. 3.5 determines the ratio of UEs from each QoS category that gets served, thus directly impacting the learning objective.
- The throughput satisfaction of each QoS category given in Eq. 3.6 relates to how well the achieved throughput compares to the throughput demanded by UEs in each QoS category.

In conjunction, the state vector of the proposed D-RAN framework with the cardinality of $3|C|$ is defined as:

$$\mathbf{s}^t = \{\varphi_1^t, ..., \varphi_{|C|}^t, \mathsf{U}_1^t, ..., \mathsf{U}_{|C|}^t, \mathsf{T}_1^t, ..., \mathsf{T}_{|C|}^t\}. \tag{3.13}$$

Action Space

For each QoS category, the action is to either increase or decrease the S-zone radius by d unit (measured in meters) or to keep it the same, that is, $\mathbf{a}_c^t = \{-d, 0, d\}$. Having a centralized agent responsible for adjusting the S-zone size for all QoS categories in the network will result in a combined action set.

The incremental action space has been selected to circumvent the combinatorically large action space that can be obtained by considering each combination of the QoS categories as an individual action, affecting the learning and convergence of the deep reinforcement learning agent greatly. Even with the incremental action space, the size of combined action space is $3^{|C|}$ for all QoS categories, which grows exponentially with QoS categories.

Motivated by the method to reduce deep reinforcement learning's large action space in [66, 67], the action space of each QoS category in D-RAN is considered as a separate action branch that controls an individual degree of freedom for each QoS category. By allowing individual action dimensions to operate independently, this approach ensures a linear increase in the size of combined action space with the number of QoS categories, of the order of $2|C| + 1$. For example, if $|C| = 2$, the following binary coding with $|C| + 1$ bits is used to represent the action space:

$$\mathbf{a} = \begin{cases} 101; & \text{increase } R_1 \text{ by } d \text{ meters.} \\ 001; & \text{decrease } R_1 \text{ by } d \text{ meters.} \\ 110; & \text{increase } R_2 \text{ by } d \text{ meters.} \\ 010; & \text{decrease } R_2 \text{ by } d \text{ meters.} \\ 000; & \text{keep } R_1 \text{ \& } R_2 \text{ unchanged.} \end{cases} \tag{3.14}$$

In a similar way, the combined action space dimensionality reduction approach is scalable to networks with a greater number of QoS categories.

Reward Function

The reward function in D-RAN primarily focuses on two aspects for the S-zone size estimation in a dynamic environment: 1) finding the optimal trade-off between system-wide KPIs formulated as a multi-objective function given in Eq. 3.7, and 2) penalizing the agent for failure to satisfy the S-zone radius constraint given in Eq. (3.7). The utility function (u^t) at each TTI t is given as the objective function given in Eq. 3.7. Subsequently, the reward is calculated as follows:

$$r^t = \begin{cases} e^{\zeta(u^t - 1)} & \text{if constraint given in Eq. 3.7 is met.} \\ Z & \text{otherwise,} \end{cases} \tag{3.15}$$

where $\zeta > 1$ in the exponential term is used to amplify the difference between values of the utility function and $-1 < Z < 0$ is a negative constant to punish the agent for choosing an S-zone size that is not within the specified bounds of R_{min} and R_{max}. The exponential scaling of the reward against utility values allows the deep reinforcement learning agent to give a much higher reward when it achieves higher utility values and much lesser when it achieves lesser or mid-range utility values. The reward function is designed to obtain values between -1 and 1 to accelerate the stochastic gradient descent algorithm in the deep neural network [68, 69].

Agent Training & Testing Procedure

The schematic diagram of the proposed D-RAN framework is shown in Fig. 3.3. The learning agent located in BBU collects state information from the environment and aims to find the optimal action policy (S-zone size for all QoS categories) such that the reward function given in Eq. 3.15 is maximized. The deep neural network includes four fully connected layers, and three rectified linear unit activation functions with input layer neurons equal to the number of state variables $3|C|$ and output layer equivalent to the number of actions $2|C| + 1$.

As part of the training process, the agent stores the experience tuple $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$ in the experience pool with buffer size $\mathcal{D}$ and updates the deep neural network weights in Eq. 3.10 by applying the stochastic gradient descent algorithm to a mini-batch of data at each epoch t (equivalent to a TTI) as detailed in Algorithm 3. As part of the execution/testing process, the agent collects the state information from the environment and outputs the action in each TTI. In every episode, consisting of T epochs/TTIs, the agent is initialized at R_{init} for all QoS categories, and the

```
Algorithm 3: D-RAN Framework
Data: $\mathcal{A}, P, T, \eta, \epsilon, \epsilon_{max}, \epsilon_{min}, \epsilon_{decay}, E, R_{init}$
Initialize state, action, reward, and experience replay buffer $\mathcal{D}$ ;
while converged or aborted do
    $violate := 0$;
    Initialize S-zone size of QoS categories as $R_c := R_{init}\ \forall c \in C$ ;
    while $t \leq T$ do
        Observe environment state $\mathbf{s}^t$;
        $\epsilon := \max(\epsilon_{min}, \epsilon - (\epsilon_{max} - \epsilon_{min})\epsilon_{decay})$;
        if $z^t \sim U(0,1) < \epsilon$ then
            Select an action $\mathbf{a}^t \in \mathcal{A}$ randomly;
        else
            Select an action $\mathbf{a}^t = \arg\max_{\mathbf{a}^t} Q^t(\mathbf{s}^t, \mathbf{a}^t|\theta_Q)$;
        if $\mathbf{a}^t$ violate $R_{min}$ and $R_{max}$ for any $R_c$ then
            Assign penalty $P$;
            $violate := violate + 1$;
            if $violate > \eta T$ then
                Abort the episode;
        Compute reward using Eq. 3.15;
        Observe next environment state $\mathbf{s}^{t+1}$;
        Store experience tuple $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$ in the experience pool;
        Prioritize experiences using Eq. 3.16;
        Sample experiences in minibatch from $\mathcal{D}$ $e_y \triangleq \{\mathbf{s}^y, \mathbf{a}^y, r^y, \mathbf{s}^{y+1}\}$;
        Perform stochastic gradient descent on $\mathcal{L}(\theta_Q)$ given in Eq. (3.10);
        Update weight parameter $\theta_Q$;
        $\mathbf{s}^t := \mathbf{s}^{t+1}$;
```

environment is initialized with different random seeds to generate different mobility patterns. An episode is ended prematurely only if the agent chooses S-zone of any QoS category that is beyond the allowed limits of S-zone size (R_{min} and R_{max}) for more than ηT times, where $0 \leq \eta \leq 1$ is a design parameter used to limit the proportion of wrong actions to ensure that the agent learns "what not to learn" [70].

The experiences drawn from experience replay during training are prioritized according to the importance of the tuple, which is dependent on the temporal difference that measures the unexpected deviation from the state transition value [71]. The prioritized experience replay algorithm stores the subsequent temporal difference error with each state transition and assigns high priority to experiences that have high temporal difference error and are recent. A stochastic sampling method is used in the D-RAN framework to interpolate experience samples between greedy and uniform random sampling by using the following formula:

$$Y = \frac{p_y^{\upsilon}}{\sum_z p_z^{\upsilon}}, \tag{3.16}$$

where $p_y > 0$ is the priority of transition y and the exponent υ determines the prioritization weightage, with $\upsilon = 0$ corresponding to the uniform random sampling. The prioritized experience replay model ensures stability and avoids local minimum convergence. To further assist stability in D-RAN training, a target deep neural network is used to predict the target Q-values that are updated after every U steps.

D-RAN adopts an exploration algorithm with the exploration variable ϵ initialized at ϵ_{max} and decayed linearly at a rate of ϵ_{decay} until ϵ_{min} is reached. If the current exploration rate ϵ is greater than a random uniform distribution sample, then the deep reinforcement learning agent chooses a random action. Learning is deemed to have converged when the average reward function is flat and no longer increases in the last E episodes. The Algorithm 3 steps can be summarized as follows:

- Initialize the environment and agent parameters.
- Observe the state of the environment at TTI t.
- Select the action at TTI t.
- Compute the reward for the action taken based on Eq. 3.15.
- Train the prioritized experience replay with the experience tuples.
- Repeat the above steps until learning has converged or aborted.

Algorithm 4: Simulated Annealing Framework

```
Data: $\mathcal{K}, N, R_{init}$
Initialize S-zone size of QoS categories as $R_c := R_{init}\ \forall c \in C$ ;
while $t \leq T$ do
    $current := \{R_1, R_2, ..., R_{|C|}\}$;
    Compute utility using Eq. 3.7 for $curr$;
    Append $N$ neighboring solutions of $curr$ to neigh by choosing the adjacent
      combinations in $\mathcal{K}$;
    Compute utility using Eq. 3.7 for neigh;
    Compute acceptance probability $AP$ using Eq. 3.11;
    if acceptance probability of $i^{th}$ neigh $>$ $curr$ then
        $curr :=$ neigh$(i)$;
    else
        $curr := curr$
```

3.5.2 Simulated Annealing Framework

Implementing a meta-heuristic such as simulated annealing for S-zone optimization in principle is similar to implementing a D-RAN framework, as the optimization agent is embedded in the BBU that adjusts the size of S-zones for all QoS categories. Instead of observing the environment state, the simulated annealing algorithm takes into account the current solution, defined as the concatenation of S-zones sizes of all QoS categories; thus, $curr = \{R_1, R_2, ..., R_{|C|}\}$. The simulated annealing algorithm traverses several neighboring solutions at each TTI and calculates the fitness of each of them. The neighboring solution space is derived from the entire solution search space $\mathcal{K}$ that includes the combinations of allowed S-zone size of all QoS categories such that its size will be $(\frac{R_{max}-R_{min}+1}{d})^{|C|}$. As such, the neighboring search space will be defined as the S-zones combinations that are adjacent to the current solution in $\mathcal{K}$. If the utility value of the neighboring solution is greater than the current solution or its acceptance probability is greater than a certain threshold, the neighboring solution is accepted. The acceptance probability is calculated using the formula given in Eq. 3.11 which is sensitive to temperature parameter $Temp$ with the fitness function is equivalent to the utility function given in Eq. 3.7 as

Table 3.2: D-RAN simulation and training parameters.

Symbol	Parameter Name	Parameter Value
λ_{UE}	UE average density	$10^3\backslash km^2$
λ_{DBS}	DBS average density	$10^3\backslash km^2$
PLE	Path-loss exponent	3
R_{min}	Minimum S-zone size	$10\,m$
R_{max}	Maximum S-zone size	$80\,m$
R_{init}	Initial S-zone for each QoS category	$(R_{max} + R_{min})/2\,m$
d	Action space stepsize	$3\,m$
α, β	Weightage parameters in Eq. 3.7	$0.4, 0.4$
P	Penalty for wrong action	-1
T	Number of epochs/TTIs	1000
η	Percentage of wrong actions allowed	5
ϵ_{max}	Maximum exploration rate	1.0
ϵ_{min}	Minimum exploration rate	0.1
ϵ_{decay}	Exploration rate decay	$0.0002/\lvert C\rvert$
U	Target deep neural network update epochs	50
E	Convergence episodes	50
N	Number of neighbor solutions	2

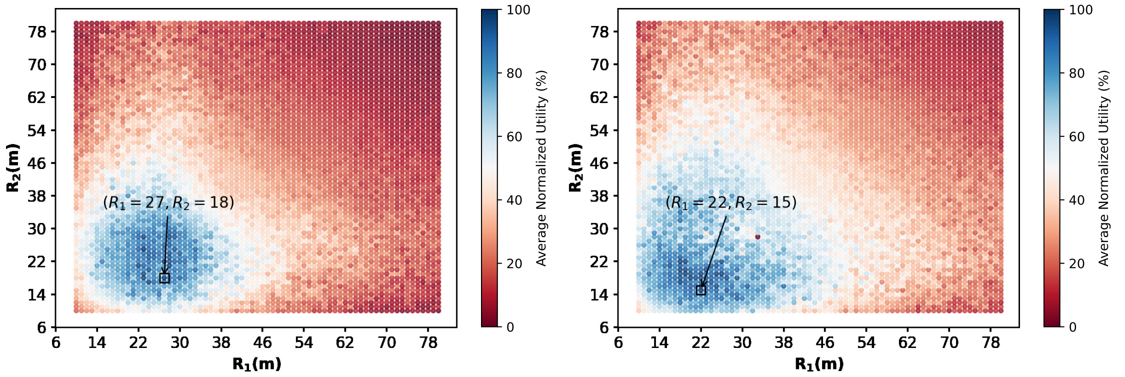
detailed in Algorithm 4.

3.6 Experimental Evaluation

Unlike the physical layer, not much data can be gathered to build pure data-driven models for system-level optimization problems. This is mainly because: 1) network operators cannot afford to try all the parameter ranges in a live network for empirical data generation, and 2) real-network data is not currently available because novel architectures, such as the user-centric architecture investigated in this chapter, are still a concept that will be implemented in 6G and beyond networks. While D-RAN does not require deluge of data from live network before-hand for training an explicit and static network behavior model, it does require some interaction on the live network, or some data from the network to build an implicit dynamic sketch of the model. As, no UC-RAN-based 6G or beyond network yet exist, we resort to a system-level simulator to meet this requirement. Although we use simulator-generated data in this chapter to train D-RAN, the insights gained remain valid for real scenarios, when the proposed D-RAN will be eventually built using data from a live network, once the proposed architecture shows benefit and is deployed in real-networks. Even in that case, pre-training the D-RAN using synthetic data from a simulator and then fine-tuning the model from live network data might be needed to address the data scarcity challenge, making the proposed synthetic data-aided deep reinforcement learning training approach worthy of investigation.

This section presents the performance of proposed D-RAN framework with system model presented in Section 3.2. The target coverage area of CBS is 1 square kilometer. The UEs and DBSs are distributed through an homogeneous Poisson point processes within the CBS coverage region. This chapter considers a maximum of three QoS categories with throughput and latency requirements of 1: virtual/augmented reality, 2: E-health, and 3: monitoring sensor networks, respectively. The number of QoS categories is determined by the network operator depending on the dominant traffic types in a specific CBS coverage area. The minimum and maximum S-zone size considered in this chapter are 10 meters and 80 meters, respectively with the action space step size of 3 meters. The choice of these user-centric cells (S-zone) size limits are inspired by industry standards [72].

Python 3.6 and Pytorch are utilized to conduct these experiments. The number of maximum epochs / TTIs (T) in each training and evaluation episode is set to 1000, where each TTI's duration is set to 1 ms. Both deep neural networks used in the main and the target network have three hidden layers containing 128-256-128 neurons. A careful choice of depth and width of these deep neural networks is made to avoid underfitting or overfitting of the nonlinear mapping between inputs and outputs. The size of the minibatch for deep neural network training is set to 64, and the target network is updated after every 50 TTIs. The rest of the network parameters and hyperparameters required to tune deep reinforcement learning-assisted and



(a)

(b)

Fig. 3.4: Comparison of brute-force solution for different UE placement realizations in a two-dimensional S-zone space.

simulated annealing-assisted frameworks are shown in Table 3.2.

3.6.1 Brute-Force Solution

The brute-force S-zone selection attempts to solves the optimization problem given in Eq. 3.7 by exhaustively searching the S-zone space of size $(\frac{R_{max}-R_{min}+1}{d})^{|C|}$. With the considered values of R_{max} and R_{min}, the brute-force solution may be a feasible option if the size of search space is less than a million combinations ($|C| < 4$). However, the size of S-zone space is not the only deterrent in making a brute-force solution infeasible. UE mobility have a direct effect on SINR, which in turn impacts the KPI values used in the utility function in Eq. 3.7, making a static solution for S-zone selection infeasible due to its complexity of the order of $\mathcal{O}\bigg((\frac{R_{max}-R_{min}+1}{d})^{|C|\times T}\bigg)$.

Fig. 3.4a and Fig. 3.4b shows the averaged normalized utility function for the different realizations of UEs positions for $|C| = 2$. While the concave envelope of maximum utility is somewhat maintained in the Fig. 3.4 (blue region), the individual utility values corresponding to each S-zone size combination as well as the apex of the utility function is shown to change. For example, the maxima of utility function (black square) changes from $(R_1 = 27, R_2 = 18)$ in Fig. 3.4a to $(R_1 = 22, R_2 = 17)$ in Fig. 3.4b. Because UE mobility follows random way point model, these values

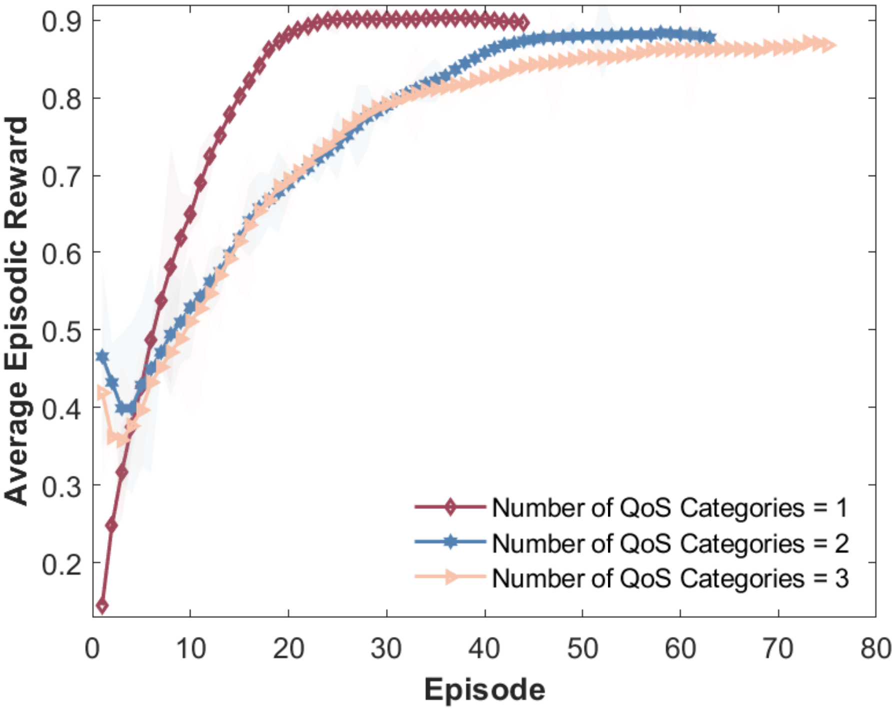


Fig. 3.5: Convergence of the average episodic reward values for varying number of QoS categories. To improve readability, these curves are smoothed with a moving average taken over 20 episodes. The shade represents the standard deviation.

may change in each TTI, which makes it necessary to assign S-zone sizes to the QoS categories dynamically and intelligently by interacting with the environment. To this end, deep reinforcement learning is a more appropriate choices in solving non-deterministic and real-time optimization of S-zone sizes.

3.6.2 Convergence Comparison for Varying Number of QoS Categories

The convergence of the proposed D-RAN framework with dynamicity in the network due to heterogeneous user application demands is shown for different numbers of QoS categories in Fig. 3.5. The value of the utility function is normalized with the upper and lower limits, determined by the brute-force solution so that the reward function can have a maximum and minimum value of 1 and -1, respectively.

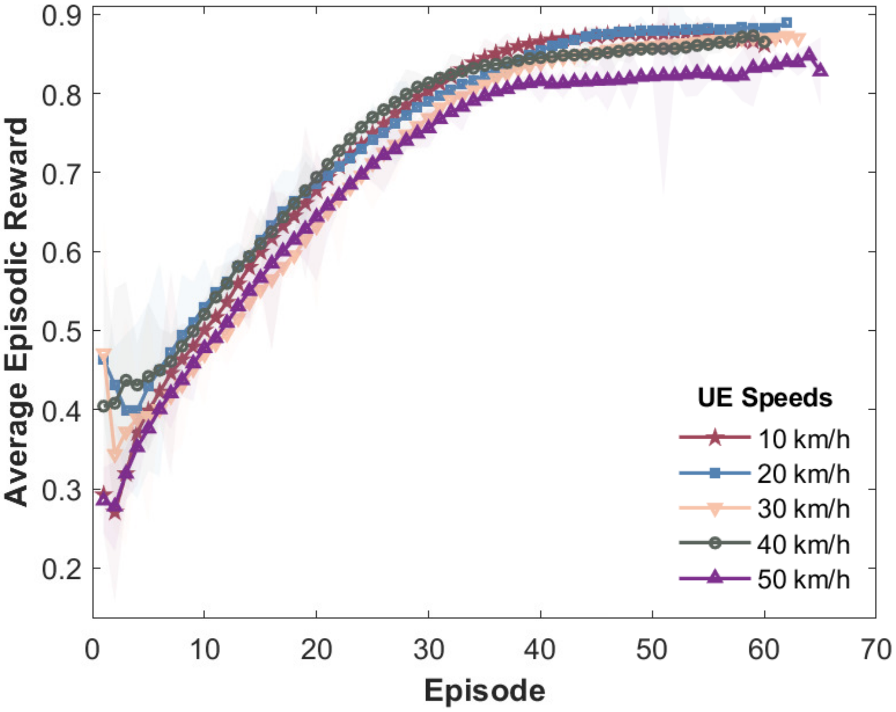


Fig. 3.6: Convergence of the average episodic reward values for varying maximum UE speeds. To improve readability, these curves are smoothed with a moving average taken over 20 episodes. The shade represents the standard deviation.

For each of the considered cases in Fig. 3.5, the learning converges towards higher reward function values after a certain amount of training episodes. The greater the number of QoS categories, the longer it takes to converge due to a larger state space, action space, and search space, requiring more TTIs to explore the environment. Additionally, as the number of QoS categories increases, the reward function tends to converge to a lower reward value. This is mainly due to the expansion of S-zone space and the increase in the minimum required number of TTIs to reach to optimal S-zone (R_c^*) from the initial S-zone (R_c^{init}) for each QoS category.

3.6.3 Convergence Comparison for Non-stationary UEs

Fig. 3.6 shows the convergence of D-RAN framework with varying maximum UE mobility speeds for $|C| = 2$. In each episode, a different random seed is used in

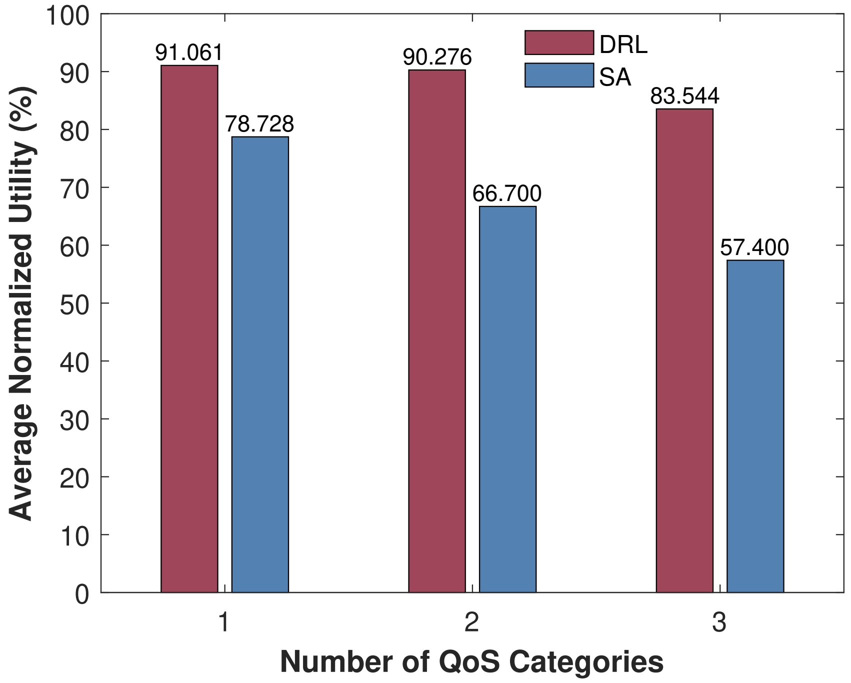


Fig. 3.7: Evaluation of the proposed D-RAN framework against simulated annealing framework for maximum UE speeds equal to 10 km/h.

the random waypoint mobility model, effectively changing the mobility pattern of UEs allowing the agent to learn the dynamics of the environment. The purpose of training with non-stationary UEs distribution is to determine whether the D-RAN framework can dynamically adjust S-zone size as the distribution of UEs changes. In the figure, it can be seen that the reward function tends to converge for each of the considered cases with a decrease in steadiness as the UE speed increases. This is mainly because the higher the UE speed, the more significant the change in the user distribution, leading to highly non-stationary maxima of the utility function causing the oscillations in convergence. However, the reward function on average converges to higher reward values, with the gap between converged reward values and maximum possible reward depicting the minimum required number of TTIs to reach to optimal S-zone (R_c^*) from the initial S-zone (R_c^{init}) as discussed in Section 3.6.2.

3.6.4 Evaluation for Different QoS Categories

In this experiment, the proposed D-RAN and simulated annealing-assisted frameworks are tested for varying number of QoS categories. The D-RAN framework is evaluated using the trained weights (state-action mapping), while the simulated annealing-assisted framework is evaluated using heuristic optimization. Performance is measured by averaging 1000 TTIs for 100 testing scenarios based on the utility function given in Eq. 3.7. To compare the performance in relative terms to maximum achievable utility, the utility values are normalized from maximum and minimum utility values obtained from the brute-force solution.

Compared to a brute-force solution that requires large computations and cannot scale, the D-RAN framework exhibits better adaptability to changing environmental conditions and maintains utility at a near-optimal level, as shown in Fig. 3.7. Additionally, the D-RAN framework surpasses the performance of the simulated annealing-assisted framework due to the slow convergence of simulated annealing optimization and the high sample complexity required to reach a reasonable solution if the search space is too large. Fig. 3.7 illustrates this phenomenon, where simulated annealing performances decrease as the number of QoS categories and the combinatorial search space increase. On the other hand, D-RAN framework manage to maintain a level of uniformity in terms of average utility scores across the QoS categories due to their ability to solve combinatorial optimization problems. Note that the D-RAN framework is not learning on the channel fading values directly since predicting/learning channel fading is too complex a task for any learning framework, particularly at a short time scale at which fast fading changes. However, the channel fading does introduce randomness in the state values and reward function of the D-RAN agent, which it considers as a random perturbation of the environment. The deep reinforcement learning agents have generally been shown to better explore the environment with random perturbations caused due to the slight imperfection of the state values or reward function. The authors in [73] have also made a similar observation where the deep reinforcement learning agent observing noisy reward sometimes even outperforms the case with the true reward, which they attribute to the implicit exploration introduced by the perturbations in the reward.

3.6.5 Proactive Real-time S-zone Optimization

The epoch / TTI-wise S-zone size optimization is shown in Fig. 3.8. To maximize the utility function, the proposed D-RAN framework adjusts the S-zone size for each QoS category to obtain the Pareto-optimal solution for area spectral efficiency, energy efficiency, user service rate, and throughput satisfaction. The S-zone size for each QoS category begins with an initial S-zone size of $\frac{R_{max}-R_{min}}{2} = 45\,m$ and then move towards the near-optimal S-zone size for each category as shown in Fig. 3.8. It can be observed that D-RAN continuously adjusts S-zone size of each QoS category with the changing network dynamics resulting in the maximization of the normalized utility. Fig. 3.9 shows the changes in S-zone size for $|C| = 1$ with associated utility scores during the exploration and exploitation stages of D-RAN. In the exploration stage, the D-RAN agent explores the environment by executing random actions so as to gain knowledge of it as shown in Fig. 3.9a. While in the exploitation stage, the agent uses its current knowledge (deep neural network weights and state-action mapping) to change S-zones size to gain higher rewards as shown in Fig. 3.9b. The results in Fig. 3.9b show that the utility function is higher (near-optimal) in the exploitation stage, indicating good learning of state-action mapping of the environment.

3.6.6 S-zone's Elasticity Impact on area spectral efficiency

Fig. 3.10 compares the one-size-fits-all S-zone size (green circles) and elastic S-zone size for $|C| = 2$. This result supports the claim in Section 3.1.2 that assigning

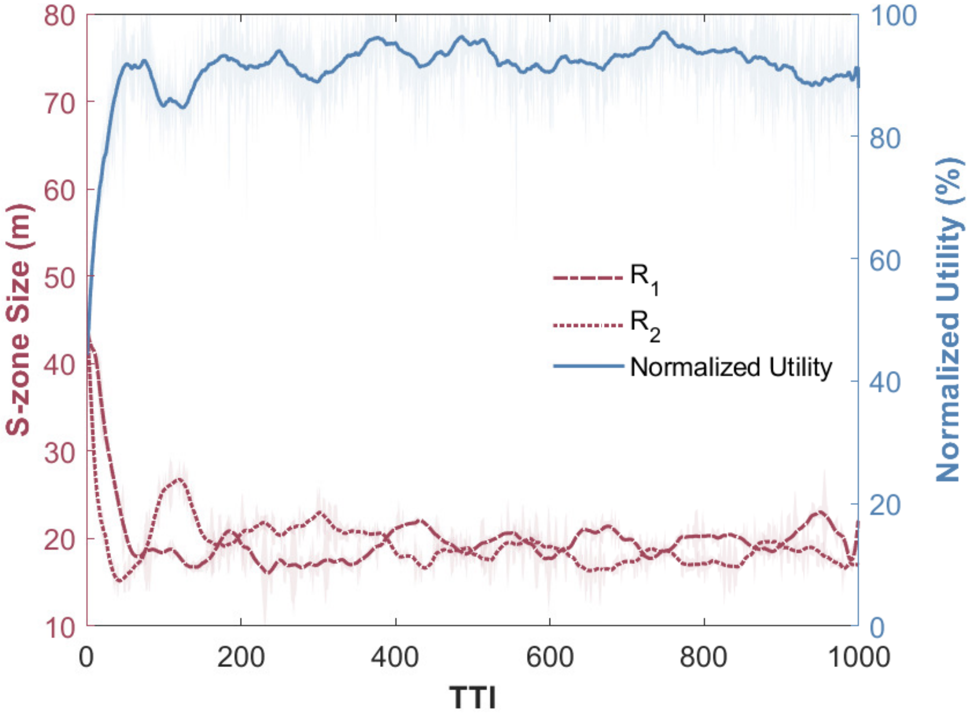
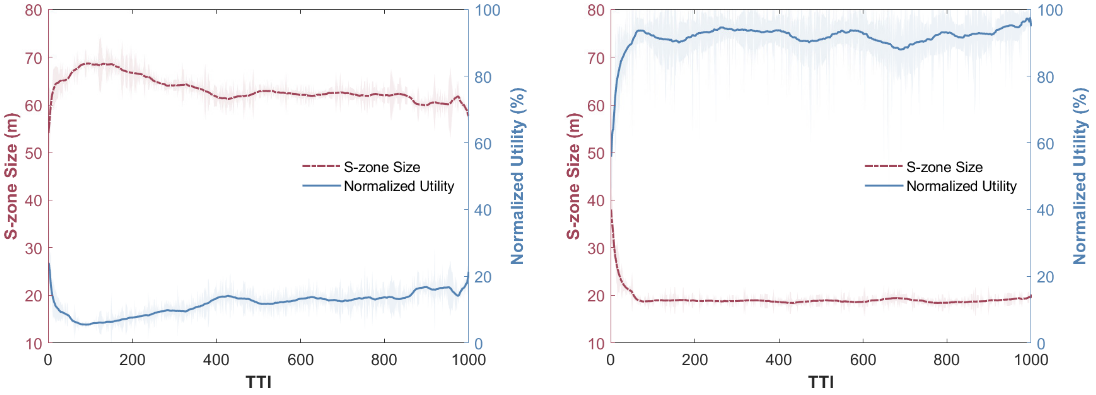


Fig. 3.8: Proactive real-time S-zone size optimization for $|C| = 2$. To improve readability, these curves are smoothed with a moving average taken over 50 TTIs. The shade represents the standard deviation.



(a) Exploration stage.

(b) Exploitation stage.

Fig. 3.9: TTI-wise normalized utility comparison for exploration and exploitation stages of D-RAN training for $|C| = 1$. To improve readability, these curves are smoothed with a moving average taken over 50 TTIs. The shade represents the standard deviation.

the same S-zone to all categories may not be optimal for accommodating heterogeneous throughput and latency requirements. The figure shows that the maximum

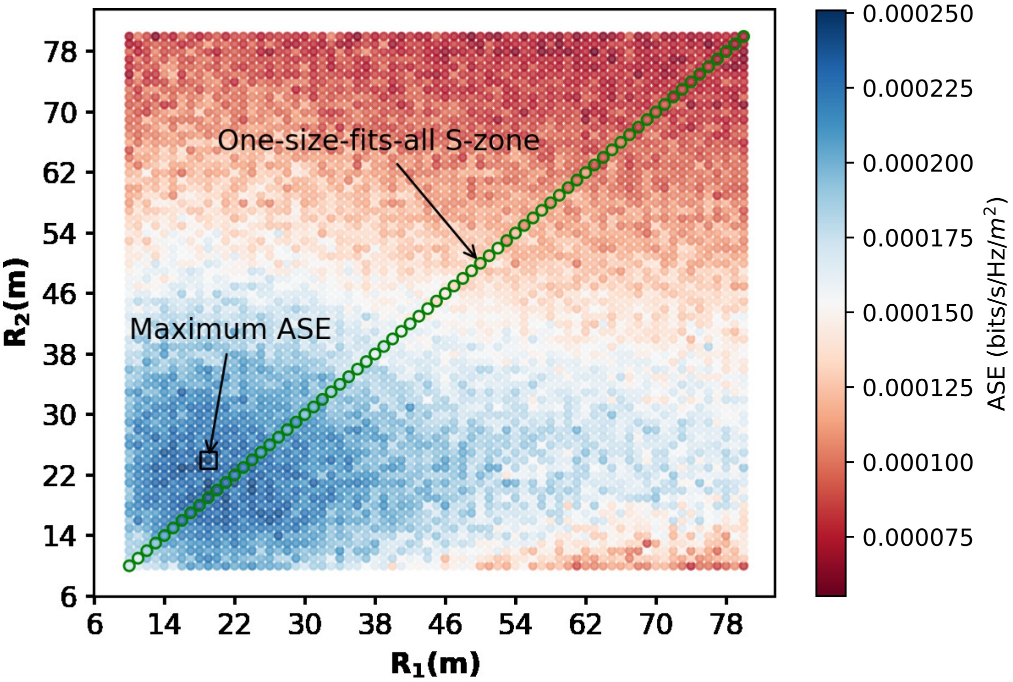
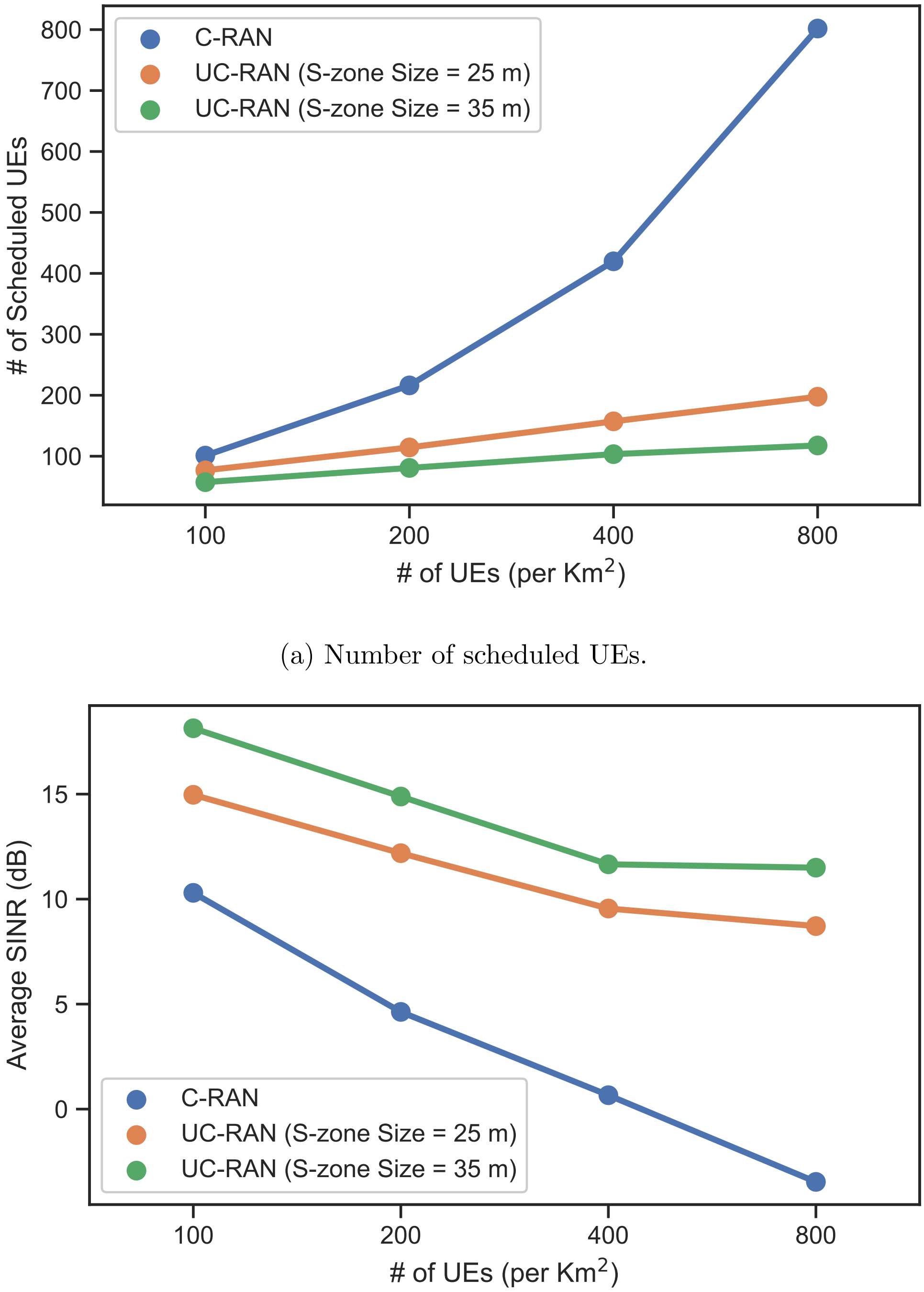


Fig. 3.10: One-size-fits-all versus elastic user-centric cell size comparison for area spectral efficiency.

achievable area spectral efficiency (black square) does not even lie within a one-size-fits-all S-zone space. The elastic S-zone architecture, however, allows for adaption to heterogeneous QoS requirements, which maximizes area spectral efficiency for the whole network.

3.6.7 Comparison of User-centric with Non-user-centric architecture

To compare the performance of the proposed user-centric approach with a non-user-centric approach, we simulate a Cloud Radio Access Network (C-RAN) model which considers similar assumptions as taken for a user-centric architecture to ensure a fair comparison between the two architectures. These assumptions are: (i) the DBSs are deployed in high density, (ii) each UE is allocated the full bandwidth of the system, (iii) there is a one-to-one association between UE and DBS, and (iv) the UE is associated with a DBS providing the maximum channel gain. With these assumptions, the only contrasting factor in C-RAN and UC-RAN architectures is



(a) Number of scheduled UEs.

(b) Average SINR (dB).

Fig. 3.11: Comparison of user-centric (UC-RAN) and non-user-centric (C-RAN) networks.

the S-zone parameter which ensures minimal separation between the scheduled UEs.

Fig. 3.11 shows the average SINR and number of scheduled UEs plots for varying UE densities. It can be observed that the average SINR in the case of C-RAN falls drastically with the increase in the density of UEs in the network. At the same time, UC-RAN architecture with the additional degree of freedom (S-zone size) is able to achieve much higher average SINRs at the cost of lesser scheduled UEs. The S-zone size controls the separation between the scheduled UEs, impacting the average SINR and the number of scheduled UEs. From Fig. 3.11, it can be hypothesized that the C-RAN (traditional Heterogenous network) architecture will not be able to perform better in a network with dense DBS deployment, which is envisaged for 6G and beyond networks. On the other hand, the UC-RAN architecture can provide an effective solution to this problem by incorporating an additional degree of freedom (S-zone size). Manually selecting the S-zone size will only be applicable if the environment is not dynamic and the solution space is too small. Therefore, intelligent control of S-zone size is needed to optimally choose the S-zone size in a dynamic environment with more than one QoS category.

3.7 Conclusion

In this chapter, we proposed D-RAN: a deep reinforcement learning-based user-centric RAN optimization framework under dynamic user application demands and network conditions. Unlike previous cellular network approaches, D-RAN employs a concept of elasticity within user-centric systems that employ non-uniform virtual cells (also called S-zones) for different QoS categories (e.g., Augmented/Virtual Reality and E-health applications). To avoid searching exhaustively using brute-force or meta-heuristics, a D-RAN framework has been developed to adjust S-zone sizes based on changing network dynamics such as user mobility. D-RAN introduces a less complex approach than brute-force or meta-heuristic techniques by accurately learning the mapping of environmental conditions to S-zone size of corresponding QoS categories. A multi-objective problem is optimized in real-time in the proposed architecture based on KPIs like area spectral efficiency, energy efficiency, UE service rate, and throughput satisfaction. Simulated results indicate that D-RAN framework is nearly as effective as brute-force and surpasses meta-heuristics like simulated annealing, but with lower complexity and is adaptable to dynamic changes in the network. In general, this chapter aims to introduce intelligence into user-centric elastic networks to accommodate user applications' non-uniform throughput and latency requirements. Even though the proposed D-RAN framework is shown to intelligently control UC-RAN COPs, many challenges must be addressed to incorporate user-centric architecture with an online optimization framework in practical cellular networks. To that end, the following chapter will discuss the user-centric architecture based on Open radio access network specifications that is able to minimize the risk associated with online optimization to make it practical for implementation.

CHAPTER 4

Digital Twin Empowered Risk-aware Reinforcement Learning Framework for User-centric O-RAN

4.1 Introduction

4.1.1 Motivation

With approximately two thousand tunable COPs and numerous KPIs in 5G [5], which are expected to increase further in future generations of cellular networks, the current data-driven methods to automate cellular networks online (such as DRL) will almost certainly break the system due to its unreliable exploratory behavior [18]. To address the safe optimization challenge in cellular architectures based on O-RAN specifications, we propose a digital twin (DT)-empowered risk-aware optimization framework in which we train a DRL on DT and use it to expedite the learning of DRL and reliably tune COPs during online optimization of live cellular networks (UC-RAN).

Recently DT paradigm has gained traction as a promising tool for cellular network design, experimentation, and optimization [74]. The core concept of DT is to create an accurate digital replica of the cellular network that incorporates wireless channel environment, antenna patterns, mobility models, antenna power consumption models, user traffic patterns, and many more, along with the capability to support emerging cellular network features such as user-centric radio access network (UC-RAN), vehicle-to-everything, to name a few. While DTs can be used for synthetic data generation or offloading computational burden from physical cellular network [75], we propose the incorporation of DT in a DRL framework to ensure accelerated risk-aware tuning of COPs in live cellular networks.

By utilizing DTs to perform DRL training which will be used during online DRL optimization of cellular networks, we can significantly minimize risks associated with COP optimization without compromising on convergence guarantees. This approach leverages the virtualized digital platform of DTs to tune cellular network COPs without any associated risk to real user satisfaction and network operator revenue. The use of DT technology can lead to efficient and safe optimization of cellular networks, enabling data-driven online solutions to become practical to tune COPs for optimizing complex KPIs in live cellular networks. To investigate the DT-empowered risk-aware optimization framework in O-RAN compatible flexible architecture suitable for future generations of cellular communications, we utilize UC-RAN for its ability to cater to the varying needs of wide-ranging quality-of-service (QoS) demands of user applications and verticals suitable for 6G networks, including telemedicine, virtual/augmented reality, industry automation, intelligent transportation, public safety networks, and metaverse [76].

4.1.2 Contributions

Specifically, the key contributions of this chapter include the following:

- To serve UEs belonging to different verticals (with eclectic QoS requirements), we propose a set of COPs (scheduled users per S-cluster and S-zone sizes for each vertical). Through experimental analysis, we establish the impact of these COPs on system-level KPIs representative of latency, reliability, and capacity tradeoff along with network energy efficiency. We formulate a multi-objective optimization problem to jointly optimize a set of KPIs by controlling proposed COPs.

- Owing to the cellular network's complexity and scarcity of real-network data, we propose a DRL-assisted solution based on soft actor-critic algorithm [77] to solve the formulated multi-objective optimization problem in real-time. The proposed DRL-assisted solution is designed as rApp (O-RAN network automation applications) hosted at the non-real-time radio intelligence controller (Non-RT RIC) to provide data-driven policy-based guidance to xApps hosted at the near-real-time radio intelligence controller (near-RT RIC).

- Data-driven or DRL-based optimization approaches, as highlighted in Section 4.1.1, can compromise the minimum level of reliability required for optimizing real-world environments, as they tend to search the solution space for optimal control policies by performing unreliable explorations [17]. To address this issue, we propose a novel framework that leverages the concept of training a base DRL model on DT, which is then used to accelerate the training of live cellular networks reliably via DRL online optimization. We define a metric that assesses the risk associated with DRL action in live networks based on the uncertainty in the choice of its actions (measured as high entropy policy) and the divergence of its action policy from that of a DT action policy (measured using Manhattan distance). We also introduce an adjustable hyperparameter (α) that can control the exploration/exploitation trade-off and can be fine-tuned based on the environment and fidelity of DT. We refer to DT enhanced DRL-assisted optimization framework as "risk-aware" rApp because it is designed to be aware of the risks involved in online optimization. In contrast, the standard DRL-assisted optimization framework, "risk-oblivious" rApp, lacks the capability to learn from DT and is, therefore, more prone to risky actions.
- To realize the system-oriented view of latency and reliability for different verticals in UC-RAN, we propose a new evaluation metric that measures the latency satisfaction and reliability satisfaction of UE's belonging to a partic-

ular vertical. We conduct manifold experiments to evaluate the convergence and risk of proposed risk-oblivious and risk-aware rApps against the brute force results. Our results show that the proposed novel risk-aware rApp can deliver impressive performance by converging to the near-optimal in less than a few hundred iterations. In addition, the risk-aware rApp significantly improves the convergence and associated risk by a factor of ten compared to risk-oblivious rApp by leveraging offline learning from the DT.

4.1.3 Chapter Organization

The rest of this chapter is organized as follows. In Section 4.2 and Section 4.3, the DT system model is presented, and a multi-objective optimization problem is formulated consisting of a set of COPs and KPIs, respectively. In Section 4.4, we discuss the novel interaction of the DRL agent and DT in O-RAN architecture to control the UC-RAN COPs optimally. Section 4.5 demonstrates the proposed frameworks' effectiveness in reliably optimizing a cellular network. Finally, this chapter is concluded in Section 4.6.

4.2 System Model

4.2.1 O-RAN-based UC-RAN Architecture

Fig. 4.1 shows the UC-RAN architecture based on O-RAN specifications where we consider a set of low-power ultra-dense open radio units (O-RUs) connected via a fronthaul link to a group of open distributed units (O-DUs). O-RU is a logical node that hosts lower physical layer functions such as fast Fourier transform, inverse fast Fourier transform, physical random-access channel, and radio-frequency operations. O-DU is a logical node that hosts higher physical layer, medium access control, and radio link control functions. O-DUs are connected via a midhaul link to the open central unit (O-CU), which is a logical node that hosts packet data convergence protocol and radio resource control functions. The management, configuration, security, fault resolution and performance assessment of these components (O-RU, O-DU, and O-CU) is performed by a logical node called service management and orchestration (SMO) [6].

In addition, intelligence is integrated into the O-RAN architecture through the RAN Intelligent Controller (RIC) composed of two vital inter-communicating modules; 1) near-RT RIC and 2) Non-RT RIC. The near-RT RIC is a logical node that hosts one or more xApps to provide control/optimization of RAN elements by collecting near real-time information. Non-RT RIC is a logical function within SMO that hosts one or more rApps to provide policy-based guidance, learning model management, and enrichment information at a granularity of greater than one second to the near-RT RIC via the A1 interface [6]. Network operators, vendors, or developers can create and deploy these artificial intelligence-enabled rApps and xApps to automate and optimize network performance.

In this chapter, we propose an rApp controlling two key UC-RAN COPs; the number of scheduled users per *S-cluster* and vertical specific S-zone size. The vertical specific S-zones (virtual user-centric cells for UEs with similar QoS requirements such as latency, throughput, etc.) are formed around UEs in each transmission time interval (TTI). Unlike the traditional UC-RAN, [3,7,8,16], where a single user was scheduled per S-zone, we introduce flexibility in the UC-RAN architecture by serving more than one users within an *S-cluster*. *S-cluster* can be of arbitrary shape and contains no more than M scheduled users, where M is a tunable parameter. In Section 4.3.2, we discuss the impact of UC-RAN COPs on a set of KPIs, where it can be observed that an appropriate choice of M can significantly improve latency satisfaction, area spectral efficiency, and energy efficiency.

Fig 4.1 shows a graphical representation of the proposed UC-RAN architecture

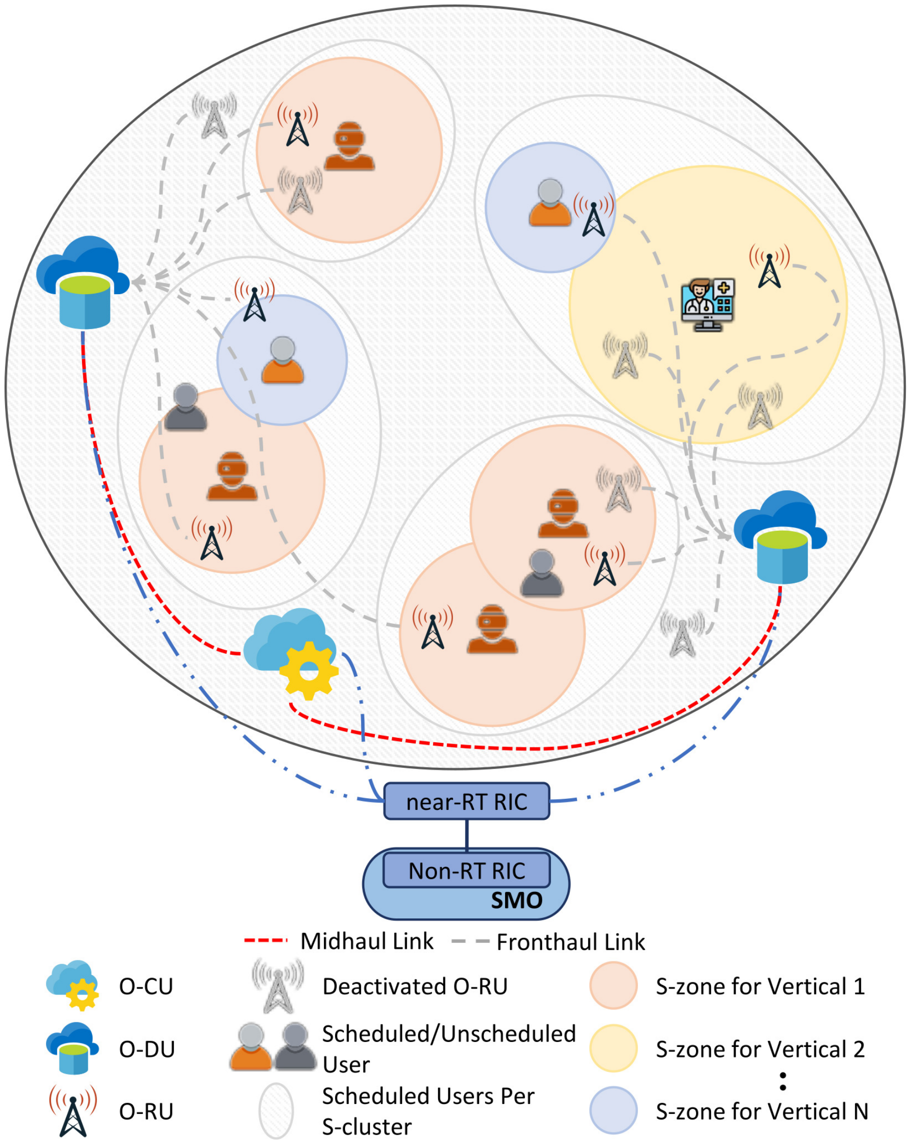


Fig. 4.1: UC-RAN architecture aligned with O-RAN specifications. The number of scheduled users per *S-cluster* is set to two in this graphical illustration. Three different verticals are defined for normal, meta-verse, and telemedicine users.

based on O-RAN abstraction where different S-zone sizes are allocated to UEs belonging to $N = 3$ different verticals and the scheduled UEs per *S-cluster* is set to 2, i.e., $M = 2$. The size of S-zone (defined by the radius of the virtual circle) determines the UE and O-RU association; in other words, only the O-RUs that fall within the S-zone region of UE are permitted to provide it service. The O-RU, which offers the best cell coverage (typically measured in terms of the reference signal's received power (RSRP)), serves the UE within its S-zone region by allocating its full bandwidth. O-DU deactivates the O-RUs which are not associated with any UE. UEs are scheduled in each TTI by the O-CU based on its scheduling priorities if and only if: (i) there are less than M scheduled users per *S-cluster*; and (ii) there is at least one O-RU available in the S-zone region with which it can associate. The UE priorities are derived from its latency requirements, i.e., lower the latency requirement, higher the scheduling priority. UEs which are not scheduled at a certain TTI are scheduled in the subsequent TTI's. Note that these UC-RAN COPs are adjusted by the network operator depending upon the requirement of the enterprise customer and the corresponding vertical.

4.2.2 Digital Twin Model

With no real cellular network implementation of O-RAN based UC-RAN architecture, we rely on our DT to accurately model two crucial aspects: the 3GPP-compliant propagation model and UE/O-RU association. While more detailed modeling is necessary for a DT to accurately mimic a cellular network, such as mobility models and traffic patterns, the implemented steps lay a solid foundation for developing a DT-empowered risk-aware optimization framework for a cellular network. For the DT, we consider the O-RAN-based UC-RAN architecture illustrated in Fig. 4.1, where the UEs and O-RUs are distributed following two independent Poisson point processes Φ_{UE} and Φ_{RU} with densities λ_{UE} and λ_{RU}, respectively. Downlink is considered and the O-RUs serve the UEs on the sub-6 GHz frequency bands. The RSRP at UE $u \in \Phi_{UE}$ from serving O-RU $r \in \Phi_{RU}$ can be mathematically expressed as:

$$\mathrm{RSRP}_u^r = P_r G_u G_r \delta_u^r \upsilon_u^r PL(d_u^r), \tag{4.1}$$

where P_r is the transmit power of serving O-RU r, G_u is the receiver antenna gain of UE u, G_r is the transmitter antenna gain of serving O-RU r towards UE u, δ_u^r is the shadowing observed from O-RU r at the location of UE u modeled as a Gaussian random variable with standard deviation of 4 dB (see [78]), v_u^r is the small-scale fading observed from O-RU r at the location of UE u modeled as exponential distribution with unit mean, d_u^r is the distance between the serving O-RU r and UE u, and $PL(d_u^r)$ is the linear dual slope path loss model derived from the Third Generation Partnership Project (3GPP) Technical Report 38.901 UMi Street Canyon line-of-sight model [78]. The non-linear dual slope path loss (in dB) is expressed as follows:

$$PL(d_u^r) = \begin{cases} PL_1; & 10m \leq d_u^r \leq d_{BP} \\ PL_2; & d_{BP} < d_u^r \leq 5km \end{cases}, \tag{4.2}$$

where $PL_1 = 32.4+21\log_{10}(d_u^r)+20\log_{10}(f_c)$, $PL_2 = 32.4+40\log_{10}(d_u^r)+20\log_{10}(f_c)-9.5\log_{10}(d_{BP}^2+(h_r-h_u)^2)$, d_{BP} is the breakpoint distance, f_c is the carrier frequency, h_r is the height of serving O-RU r, and h_u is the height of UE u. From the above derivation, the signal-to-interference-plus-noise-ratio (SINR) experienced at UE $u \in \Phi_{UE}$:

$$\text{SINR}_u^r = \frac{\text{RSRP}_u^r}{N_0 + \sum_{\substack{r' \in \Phi'_{RU}, \\ r' \neq r}} \text{RSRP}_u^{r'}}, \tag{4.3}$$

where Φ'_{RU} is the Poisson point processes of activated O-RUs, and N_0 denotes the noise power.

4.3 Problem Formulation

This section first provides a detailed account of KPIs used in this chapter followed by the formulation of a multi-objective optimization problem.

4.3.1 Key Performance Indicators

With the goal of addressing the latency, reliability, and capacity tradeoff while optimizing network energy efficiency, we assess system performance in terms of latency satisfaction, reliability satisfaction, area spectral efficiency, and network energy efficiency.

Latency Satisfaction

3GPP defines user plane latency as the time span for unidirectional data transfer from the access point's radio protocol layer to the UE's radio protocol ingress point, assuming the UE is in active state [79]. To quantify 3GPP's definition of latency, we measure latency satisfaction as the weighted sum of the percentage of UEs (from each vertical) served with the required data rate within its latency constraint. The data rate is measured in each TTI when the UE requests service. If the sum of the measured data rate across TTIs satisfies the latency bound of UE, then the latency requirement of that user is said to be served. Mathematically,

$$\text{Latency Satisfaction} = \sum_{i=1}^{N} \dot{w}_i \left(\frac{\sum_{j=1}^{|\Phi_i|} \mathbb{1}_{\left\{\left(\sum_{\tau=T_{ij}}^{T_{ij}+l_i} \Gamma_{ij\tau}\right) \geq \gamma_i \right\}}}{|\Phi_i|} \right), \tag{4.4}$$

where $\mathbb{1}_{\{.\}}$ is the characteristic function, N is the number of verticals, $|\Phi_i|$ represent the number of UEs belonging to vertical i, T_{ij} is the TTI at which UE j belonging to vertical i requests service, l_i is the latency requirement for each vertical, γ_i is the data rate requirement for each vertical during each TTI, $\Gamma_{ij\tau}$ represents the measured data rate at UE j belonging to vertical i during each TTI, and $\dot{w}_i \geq 0$, $\forall i$ and $\sum_{i=1}^{N} \dot{w}_i \leq 1$ are network operator-defined weights assigned to prioritize latency requirements of specific verticals. Note that latency satisfaction metric is calculated for every L TTIs where $L = \max(l_i),\ \forall i \leq N$.

Reliability Satisfaction

3GPP defines reliability as the capability of transmitting a given amount of traffic to UE from the application server within the required time constraint with high success probability [79]. To quantify 3GPP's definition of reliability, we measure reliability satisfaction as the weighted sum of the averaged probability of packets received correctly for UEs belonging to each vertical. The Block Error Ratio (BLER) is a measure of the ratio of erroneous blocks received to the total number of blocks sent, calculated at each Transmission Time Interval (TTI). The UE reports a Channel Quality Indicator (CQI) which is used in conjunction with a SINR defined in Eq. 4.3 to map the CQI to a corresponding BLER value [80]. Mathematically,

$$\text{Reliability Satisfaction} = \sum_{i=1}^{N} \ddot{w}_i \mathbb{E}_\tau \left[\frac{\sum_{j=1}^{|\Phi_i'|}(1-\beta_{ij\tau})}{|\Phi_i'|} \right], \tag{4.5}$$

where $|\Phi_i'|$ represent the number of scheduled UEs belonging to vertical i, $\beta_{ij\tau}$ represents the BLER at UE j belonging to vertical i, $\mathbb{E}_\tau[.]$ represents averaging over several TTIs, and $\ddot{w}_i \geq 0$, $\forall i$ and $\sum_{i=1}^{N} \ddot{w}_i \leq 1$ are network operator-defined weights assigned to prioritize the reliability satisfaction of specific verticals.

Area Spectral Efficiency

The area spectral efficiency is the time-averaged total throughput of active UEs per bandwidth channel per unit area. Mathematically,

$$\text{Area Spectral Efficiency} = \frac{\mathbb{E}_\tau \left[\sum_{i=1}^{N} \sum_{j=1}^{|\Phi_i'|} \Gamma_{ij\tau} \right]}{\mathfrak{A} \times B}, \tag{4.6}$$

where $\mathfrak{A}$ is the area of the target region and B is the channel bandwidth.

Network Energy Efficiency

The network energy efficiency is defined as the ratio of area spectral efficiency and network-wide power consumption, which includes: (i) power required to operate O-RU in listening mode; (ii) O-RU transmission power; (iii) power required for the UE to discover O-RUs; and (iv) UE transmission power. Mathematically,

$$\text{Energy Efficiency} = \frac{\mathfrak{A} \times \text{Area Spectral Efficiency}}{\lambda_{RU} P_f + \lambda'_{RU} \Delta_{RU} P_{RU} + \lambda'_{UE} \Delta_{UE} P_{UE}}, \tag{4.7}$$

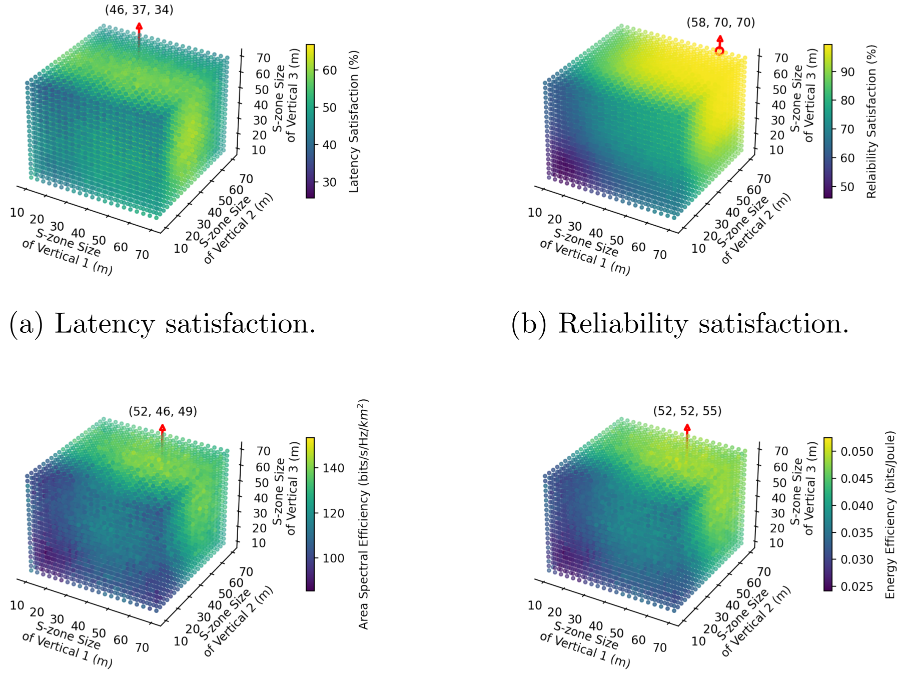
where λ'_{RU} is the density of activated O-RUs, λ'_{UE} is the density of scheduled UEs, P_f is the fixed power required to operate O-RU in listening mode, P_{RU} is the O-RU transmission power, Δ_{RU} is the O-RU radio frequency component power, P_{UE} is the UE transmission power, and Δ_{UE} is the UE radio frequency component power. The typical values of these variables are such that: $P_f = 1.932$ Watts, $P_{RU} = 10$ Watts, $\Delta_{RU} = 23.22$ Watts, $P_{UE} = 1$ Watts, and $\Delta_{UE} = 4$ Watts [3].

4.3.2 Impact of UC-RAN COPs on KPIs

In this section, we discuss UC-RAN COPs (scheduled users per *S-cluster* and S-zone size of different verticals) impact on KPIs (latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency) discussed above.

S-zone Size of Verticals

We begin by analyzing the impact of the S-zone size of different verticals on KPIs by fixing the number of verticals and number of scheduled UEs per *S-cluster* to three. The latency and throughput requirement for each vertical is set differently, such that vertical 1 has high-throughput but relaxed-latency requirement, vertical 2 has low-throughput and ultra-low latency requirement, and vertical 3 has low-



(a) Latency satisfaction.

(b) Reliability satisfaction.

(c) Area spectral efficiency.

(d) Energy efficiency.

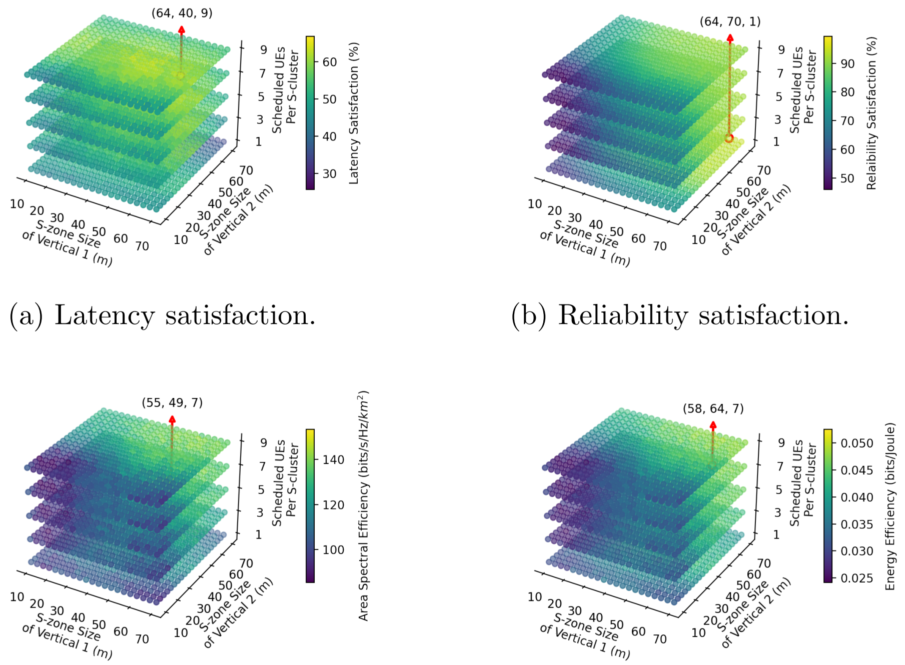
Fig. 4.2: Impact of S-zone size of verticals on latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency with maximum COP combination labeled as red circle.

throughput and relaxed-latency requirement. Intuitively, the decrease in S-zone size enables scheduling more users at the expense of high interference from neighboring O-RUs activated for other scheduled UEs. These contrasting trends impact the KPIs differently.

For instance, Fig. 4.2a shows the impact of three verticals S-zone size on latency satisfaction metric, where the optimal region is represented in yellow shade and the optimal COP combination $(46, 37, 34)$ is annotated with a red circle. The optimal S-zone size is different for each vertical because of the throughput and latency requirement we set for UEs belonging to respective verticals. UEs belonging to a vertical with low-latency requirement demand a smaller S-zone size such that UEs can be scheduled without any delay, whereas UE belonging to a vertical with high-throughput requirement demand a larger S-zone size such that the average distance between scheduled UE and activated O-RU to increase leading to decrease in the interference (hence, increase in SINR and throughput).

The relation of larger S-zones leading to high SINR is more evident in Fig. 4.2b, where the optimal region for reliability satisfaction metric is at larger S-zone sizes signifying the positive impact of large S-zone size on low BLER (hence, high SINR). As the reliability satisfaction metric is calculated for scheduled UEs only, it aims to minimize the BLER of scheduled UEs by maximizing the S-zone size without considering the negative impact of increased S-zone size on latency.

Fig. 4.2c shows the impact of S-zone size on area spectral efficiency. Area spectral efficiency is a metric dependent on the number of scheduled UEs and SINR (throughput). Therefore, the optimal S-zone region will balance these two opposing effects to maximize area spectral efficiency, as shown in the yellow shaded region in Fig. 4.2c. The direct relationship between network energy efficiency and area spectral efficiency mandates that the S-zone size of verticals will influence network energy efficiency similarly to area spectral efficiency. However, the large S-zone



(a) Latency satisfaction.

(b) Reliability satisfaction.

(c) Area spectral efficiency.

(d) Energy efficiency.

Fig. 4.3: Combined impact of scheduled UEs per *S-cluster* and S-zone Size of verticals on latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency with maximum COP combination labeled as red circle.

size of verticals leads to less power consumption (high energy efficiency) since fewer O-RUs will be activated in the network. These trends can be observed in Fig. 4.2d, where the optimal region of network energy efficiency is similar to area spectral efficiency with peaks at the maximum S-zone size of verticals.

Scheduled UEs Per *S-cluster* and S-zone Size of Verticals

We analyze the combined impact of the scheduled users per *S-cluster* and S-zone size of verticals on the system performance. Intuitively, with the increase in the scheduled UEs per *S-cluster*, more UEs are allowed to be scheduled in a specific TTI. However, higher scheduling would diminish the interference protection between scheduled UEs and activated O-RUs of neighboring scheduled UEs, leading to decreased SINR.

These contrasting trends impact can be observed in Fig. 4.3 where the impact of the S-zone size of vertical 1, S-zone size of vertical 2, and scheduled users per *S-cluster* is shown on the above-mentioned KPIs. From the plots shown in Fig. 4.3a, 4.3c, 4.3d, it can be observed that latency satisfaction, area spectral efficiency, and energy efficiency's optimal region in terms of verticals S-zone size remains the same as observed in Fig. 4.2. However, with the increase in scheduled UEs per *S-cluster*, the values of these KPIs increase signifying the positive impact of the increase in the number of scheduled UEs per *S-cluster* on latency satisfaction, area spectral efficiency, and energy efficiency metrics.

Opposite trends can be observed for reliability satisfaction, where the increase in the number of scheduled UEs per *S-cluster* negatively impacts the reliability satisfaction metric. Revisiting the inherent rigidity of previous UC-RAN architectures [3, 7, 8] to limit the scheduled UEs per *S-cluster* to one, it is apparent from Fig. 4.3 that scheduling a single UE does not give the best performance in terms of latency satisfaction, area spectral efficiency, and energy efficiency.

4.3.3 Multi-Objective Optimization Problem

The contrasting impact of UC-RAN COPs (scheduled users per *S-cluster* and S-zone size of different verticals) on KPIs, discussed in Section 4.3.2, raises an important design question; how to jointly optimize these COPs to maximize KPIs? From the network operator's point of view, each of these KPIs should be maximized while considering the network's dynamicity, such as varying UE and O-RU distributions, QoS requirements, and number of verticals.

Optimization Objective Function

To achieve this, we formulate a multi-objective optimization problem given in Eq. 4.8, which minimizes the difference between latency satisfaction, reliability sat-

$$\begin{aligned} \underset{M,\mathbf{C}}{\text{minimize}} \quad & \left|\sqrt{\kappa_1(\zeta-\zeta_{target})^2+\kappa_2(\varsigma-\varsigma_{target})^2+\kappa_3(\rho-\rho_{target})^2+\kappa_4(\xi-\xi_{target})^2}\right| \\ \text{subject to} \quad & M_{min} \leq M \leq M_{max}, \\ & C_{min} \leq C_i \leq C_{max}; \forall C_i \in \mathbf{C} = \{C_1, C_2, ..., C_N\}, \\ & 0 \leq \kappa_1, \kappa_2, \kappa_3, \kappa_4 \leq 1, \\ & 0 \leq \kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 \leq 1 \end{aligned} \tag{4.8}$$

isfaction, area spectral efficiency, and energy efficiency KPIs with their respective target values set by the network operator. In Eq. 4.8, the optimization variables are; number of scheduled UEs per *S-cluster*, denoted as M, and S-zone size of N verticals, $\mathbf{C} = \{C_1, C_2, ..., C_N\}$. (M_{min}, M_{max}) is the optimization range for the number of scheduled UEs per *S-cluster*, (C_{min}, C_{max}) is the optimization range for the S-zone size of verticals, ζ, ς, ρ, and ξ denotes normalized values of latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency, respectively, $_{target}$ subscript correspond to the target values for ζ, ς, ρ, and ξ, and $\kappa_1, \kappa_2, \kappa_3$ and κ_4 are network operator-defined weights that can be used to adjust the relative importance of these KPIs.

With the formulated optimization problem in Eq. 4.8, the rApp hosted at Non-RT RIC aims to solve multiple objectives. First, it addresses the latency, reliability, and throughput tradeoff along with network energy efficiency from a system-level perspective by controlling the UC-RAN COPs. Second, it enables priority access to specific vertical(s) through network operator-defined weights in individual KPIs such as $\dot{\mathbf{w}}$ in latency satisfaction and $\ddot{\mathbf{w}}$ in reliability satisfaction. Third, it combines a set of KPIs in a rather sophisticated way where all KPIs aim to get nearer to the target values set by the network operator minimizing the inherent conflict between latency, reliability, area spectral efficiency, and energy efficiency KPIs. Fourth, these multi-dimensional KPIs are combined by normalizing each KPI value with its utopia and nadir values. We determine utopia and nadir approximate values for each KPI by a psuedo brute force method which sweeps the solution space (with a predefined step size) in multiple independent simulation iterations, a standard method in general reinforcement learning problems [61].

Complexity

Optimizing multiple COPs to maximize diverse and multifaceted KPIs is all but a trivial task. The multi-objective problem formulated in Eq. 4.8 is a mixed-integer nonlinear programming problem with the complexity of the order of $\mathcal{O}\Bigg((M_{max} - M_{min} + 1) \times (C_{max} - C_{min} + 1)^N\Bigg)$. The computational difficulty required to solve such a complex multi-dimensional problem with dynamically varying UE and O-RU distribution with diverse QoS requirements is a mammoth task. The conventional methods to solve such problems rely on analytical or heuristics-based models. These methods either fail to capture network complexity by making simple assumptions about underlying network conditions or utilize a deluge of offline data for optimization, which is not representative of real-world network behavior and is often unavailable.

4.4 Digital Twin Empowered Risk-aware Optimization Framework

The DRL approach, as opposed to the conventional methods discussed above, allows network operators to optimize COPs online by learning from the live responses of real networks, which embodies the dynamic and complex nature of cellular networks. However, DRL's powerful decision-making ability without utilizing offline data comes at a cost. This cost is paid by compromising on the system's reliability by learning through interactions with the environment which often includes actions that compromise on the required safety [17]. For example, consider an inefficient air traffic control system deployed in a city. Using DRL to improve air traffic control system performance is highly desirable; however, DRL should never allow the system to perform worse than the existing suboptimal solution as this might result in anomalous cascading effects that can cause severe problems in air traffic control.

To ensure the DRL performance does not fall below the necessary reliability required for optimizing live cellular networks, we leverage UC-RAN's DT to reliably accelerate the learning process of the DRL model. To this end, we propose risk-oblivious and risk-aware rApps based on DRL framework in O-RAN architecture to solve the optimization problem formulated in Eq. 4.8. The risk-oblivious rApp learns to optimize the set of COPs based on the real network responses. On the other hand, the risk-aware model utilizes real network responses and offline learning from UC-RAN's DT to diminish the risk of choosing extreme exploratory actions in optimizing live cellular networks.

We use the state-of-the-art soft actor-critic algorithm [77] as an enabler to implement DRL in the proposed risk-oblivious and risk-aware rApps. Our choice of soft actor-critic is mainly motivated by its ability to efficiently explore large action spaces. In addition to its applicability to problems with large action space, the soft actor-critic method offers high sample efficiency and avoids brittleness to hyperparameters. Details of soft actor-critic method with policy and value function networks are included in Appendix C. Ergo, we first discuss the proposed rApps deployment architecture aligned with the O-RAN specifications, followed by the discussion on risk-oblivious and risk-aware rApps framework.

4.4.1 Deployment Architecture in O-RAN

Fig. 4.4 shows the deployment framework for proposed risk-oblivious and risk-aware rApps in the O-RAN architecture, which can be summarized as:

- The low-level performance indicators, such as RSRP, SINR, measured data

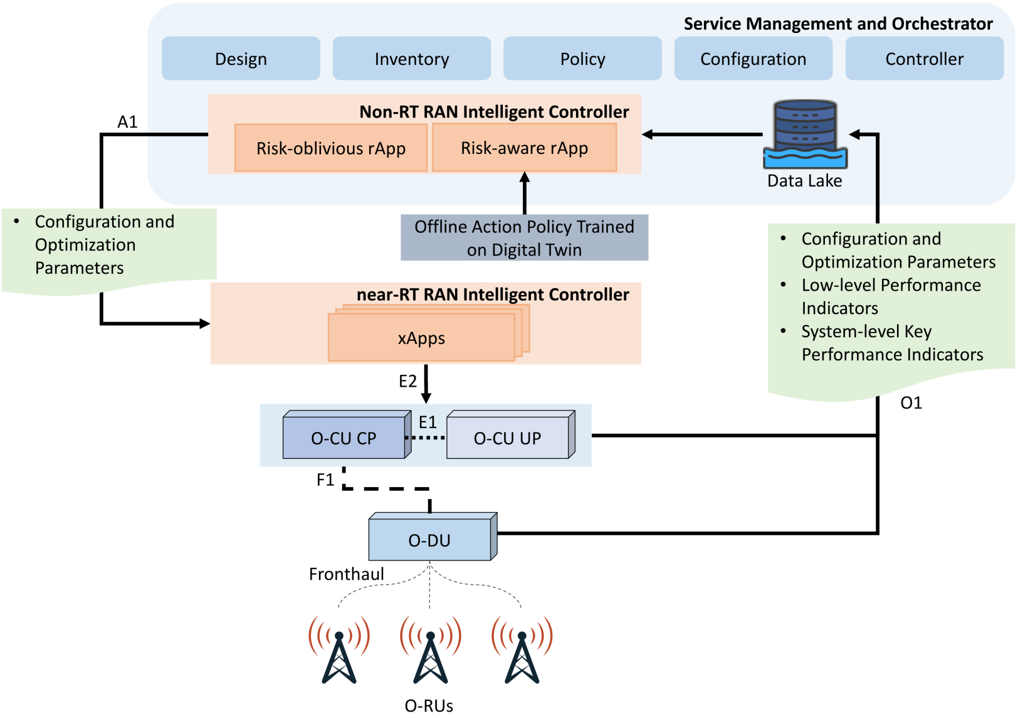
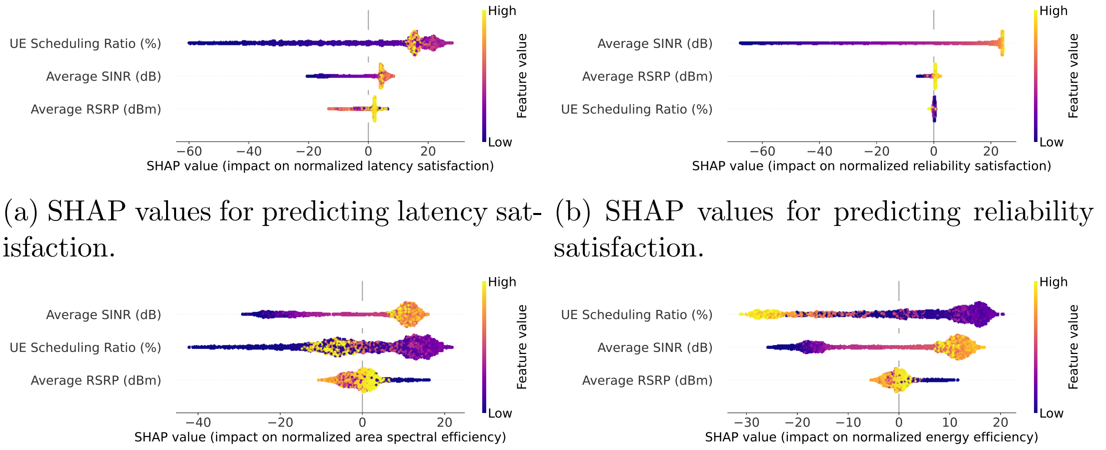

Service Management and Orchestrator
Design
Inventory
Policy
Configuration
Controller
Non-RT RAN Intelligent Controller
Risk-oblivious rApp
Risk-aware rApp
Data Lake
A1
• Configuration and Optimization Parameters
Offline Action Policy Trained on Digital Twin
• Configuration and Optimization Parameters
• Low-level Performance Indicators
• System-level Key Performance Indicators
near-RT RAN Intelligent Controller
xApps
E2
O1
O-CU CP
E1
O-CU UP
F1
O-DU
Fronthaul
O-RUs

Fig. 4.4: High-level deployment framework for the proposed risk-oblivious and risk-aware rApps in the O-RAN architecture.

rate, BLER, number of scheduled UEs, etc., along with vertical-specific identifiers and system-level KPIs such as latency satisfaction, reliability satisfaction, area spectral efficiency, energy efficiency, etc., are collected via the O1 interface at the data lake residing in SMO.

- The data lake collects low-level performance indicators, system-level KPIs, and COPs to perform any required data cleaning operations. Note that the low-level performance indicators data is utilized to calculate system-level KPIs discussed in Section 4.3.1 with the additional knowledge of the data rate and latency requirement for each vertical.

- Data statistics from data lake are sent to Non-RT RIC residing in the SMO. rApps, running in the Non-RT RIC, uses this information to make intelligent decisions on the choice of COPs (scheduled UEs per *S-cluster*, S-zone size of verticals) to maximize the combination of KPIs in an online manner. It is important to note that rApps proposed in this chapter are assumed to be



(a) SHAP values for predicting latency satisfaction.

(b) SHAP values for predicting reliability satisfaction.

(c) SHAP values for predicting area spectral efficiency.

(d) SHAP values for predicting energy efficiency.

Fig. 4.5: Feature importance analysis using SHAP values of average RSRP, SINR and UE scheduling ratio on latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency.

capable of DRL model training, inference, and updates.

- Some rApps may require additional information, such as risk-aware rApp (discussed in Section 4.4.2), which requires access to an offline trained model from UC-RAN's DT.

- The action policy obtained from risk-oblivious or risk-aware rApps is sent to near-RT RIC. With the proposed deployment framework, the action policy from rApp will change at a timescale of more than one second.

- xApps, hosted at near-RT RIC, will trigger the change in scheduled UEs per *S-cluster* and S-zone size for verticals via the E2 interface. Note that xApps in the proposed deployment framework only apply the policy recommended by rApps (hosted at Non-RT RIC) at O-CU (consisting of both the control and user planes), which further applies this policy to O-DUs via the F1 interface.

4.4.2 Main Components of the Framework

Markov Decision Process Formulation

As depicted in Fig. 4.4, the proposed risk-oblivious and risk-aware rApps use DRL for optimization. Hence the selected COPs-KPIs of UC-RAN are formulated as a Markov decision process in terms of system state space, action space, and reward, which will be described in the following:

System State Space: On closer inspection of the KPIs discussed in Section 4.3.1, it can be noticed that the UC-RAN COPs primarily influence three main network features (RSRP, SINR, and UE scheduling ratio), which in turn impacts latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency KPIs. To quantify each features' impact, SHAP (SHapley Additive exPlanations) [81], which is a game theoretic approach to interpret the output of machine/deep learning algorithms, is utilized on the brute force simulated data. The horizontal axis in Fig. 4.5 show SHAP values which shows the impact of these three features (average RSRP, average SINR, UE scheduling ratio) in predicting the values of KPIs. In addition, each data point's color represents the value of a feature from high to low (for instance, yellow colored data point of SINR features represents a high value of SINR) and the vertical thickness of data points indicates the density of data points.

From domain knowledge, we know that high values of these features impacts latency satisfaction positively, where in particular, UE scheduling ratio and SINR has the most impact on latency satisfaction. The same is evident from the SHAP values plotted in Fig. 4.5a. Similarly, reliability satisfaction is a metric highly dependent on the BLER, which has an inverse relationship with SINR, hence the observed high feature importance of SINR on reliability satisfaction is noticeable in Fig. 4.5b. Both area spectral efficiency and energy efficiency have an interdependent relation with UE scheduling ratio, average RSRP, and SINR, which can be observed in

Fig. 4.5c and Fig. 4.5d. These plots illustrate that higher values of SINR impacts area spectral efficiency and energy efficiency positively, while the high values of scheduling ratio impacts these KPIs inversely, due to the low SINR and high power consumption of large number of activated O-RUs.

In light of the above observations, since these features notably impact the considered KPIs, they will enable the learning algorithm to determine the state of the network accurately. Therefore, we combine these three features, for each vertical, to define the system state.

State Vector: Directly using these instantaneous feature values as the system state may not reflect the KPIs discussed in Section 4.3.1, which are measured across several timestamps or TTIs. For this reason, the system state is defined by stacking the values of RSRP, SINR, and UE scheduling ratio for a predefined number of TTIs, L. Due to its ability to capture latent time flow encoded information effectively, we employ FLARE (Flow of Latents for Reinforcement Learning) method which uses the difference of feature values between current and subsequent timestamps, as a state variable [82]. To that end, the system state $\mathbf{s}_e$ at epoch e can be defined as:

$$\mathbf{s}_e = \{\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, ..., \mathbf{x}_N, \mathbf{y}_N, \mathbf{z}_N\}, \tag{4.9}$$

where $\mathbf{x}_i = \{x_i^{t-L}, x_i^{t-L-1} - x_i^{t-L}, ..., x_i^t - x_i^{t-1}\}$, $\mathbf{y}_i = \{y_i^{t-L}, y_i^{t-L-1} - y_i^{t-L}, ..., y_i^t - y_i^{t-1}\}$, and $\mathbf{z}_i = \{z_i^{t-L}, z_i^{t-L-1} - z_i^{t-L}, ..., z_i^t - z_i^{t-1}\}$ represents the stacked values of the i^{th} vertical, for the previous L TTIs, for RSRP, SINR, and UE scheduling ratio features, respectively, and $t = (e + 1)L$.

Cardinality of the State Space: These features values (RSRP, SINR, and scheduling ratio) are calculated for all UEs belonging to the same vertical; therefore, the size of system state space S is independent of the number of UEs with a cardinality of $3 \times L \times N$.

System Action Space: The optimization variables in Eq. 4.8 are; number of scheduled UEs per *S-cluster* (M) and S-zone sizes ($C_i \; \forall C_i \in \mathbf{C} = \{C_1, C_2, ..., C_N\}$) for

$$r_e = \omega \left(1 - \left|\sqrt{\kappa_1(\zeta - \zeta_{target})^2 + \kappa_2(\varsigma - \varsigma_{target})^2 + \kappa_3(\rho - \rho_{target})^2 + \kappa_4(\xi - \xi_{target})^2}\right|\right) \tag{4.10}$$

the N verticals. Therefore, the system action $\mathbf{a}_e$ at each epoch e is defined as:

$$\mathbf{a}_e = \{M, C_1, C_2, ..., C_N\}.$$

Cardinality of the Action Space: The size of the system action space A is defined by the number of verticals (N), range of scheduled UEs per *S-cluster* ($M_{max} - M_{min} + 1$) and S-zone size ($C_{max} - C_{min} + 1$), such that, $|A| = (M_{max} - M_{min} + 1) \times (C_{max} - C_{min} + 1)^N$.

System Reward: The system reward should capture the objective function value formulated in Eq. 4.8 corresponding to an action $\mathbf{a}_e$ taken at state $\mathbf{s}_e$. Due to the normalization of KPIs and constraints on the values of network operator-defined weights, the objective function values are in the range, $[0, 1]$. Thus, by subtracting the objective function value from 1, the minimization objective can be transformed into a maximization objective, resulting in system reward which is given in Eq. 4.10. Here the ω parameter scales the reward to higher values since the soft actor-critic algorithm is shown to work better with larger rewards magnitude [77].

Risk-oblivious rApp Framework

This section introduces the risk-oblivious rApp, which tends to explore the environment unreliably and can degrade network performance to intolerable levels, as demonstrated in Section 4.5. As a result, the risk-oblivious rApp will serve as a baseline for comparing the performance of the risk-aware rApp, which will be discussed in Section 4.4.2. With the above derivation of system state, action, and reward, the risk-oblivious rApp executes the action $\mathbf{a}_e$ via A1 interface at the current state $\mathbf{s}_e$,

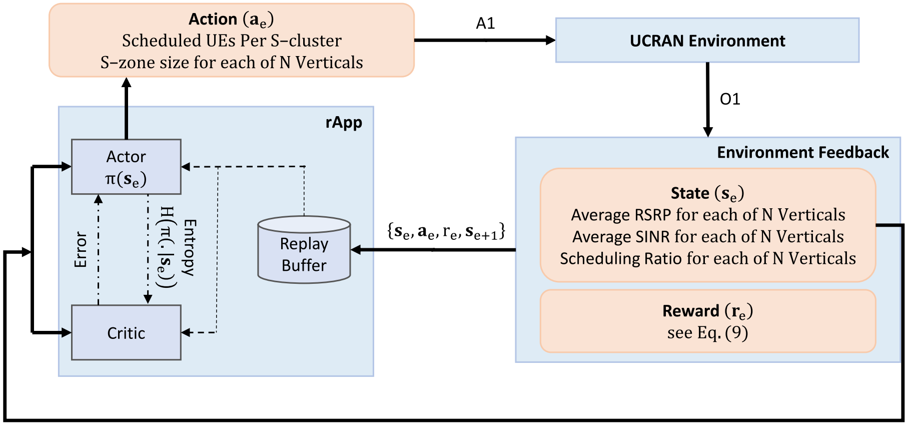

Action ($\mathbf{a}_e$)
Scheduled UEs Per S–cluster
S–zone size for each of N Verticals
A1
UCRAN Environment
O1
rApp
Actor
$\pi(\mathbf{s}_e)$
Error
$H(\pi(\cdot|\mathbf{s}_e))$
Entropy
Replay Buffer
Critic
$\{\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1}\}$
Environment Feedback
State ($\mathbf{s}_e$)
Average RSRP for each of N Verticals
Average SINR for each of N Verticals
Scheduling Ratio for each of N Verticals
Reward ($\mathbf{r}_e$)
see Eq. (9)

Fig. 4.6: Risk-oblivious rApp framework for UC-RAN COPs optimization.

which changes the environment to the next state $\mathbf{s}_{e+1}$, returning a reward signal r_e characterizing the utility for taking action $\mathbf{a}_e$ on UC-RAN. Note that the system and reward are observed using observations sent via the O1 interface to Non-RT RIC.

Fig. 4.6 shows the framework for risk-oblivious rApp for UC-RAN COPs optimization where the system state, action, reward, and next state in combination form an experience tuple $(\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1})$ sent forward to the replay buffer at each epoch e. This tuple is utilized when training function approximators using stochastic gradient.

The actor and critic networks are defined as:

- The actor network $\pi_\psi(\mathbf{s}_e)$, with parameter ψ trained using Eq. 6.17, estimates the mean and standard deviation of the conditional Gaussian probability distribution for each continuous action $\mathbf{a}_e$ when in state $\mathbf{s}_e$.

- The critic network includes: (i) two soft Q-functions $Q_{\theta_1}(\mathbf{s}_e, \mathbf{a}_e)$ and $Q_{\theta_2}(\mathbf{s}_e, \mathbf{a}_e)$, with parameters θ_1 and θ_2 trained using Eq. 6.15, takes state $\mathbf{s}_e$ and action $\mathbf{a}_e$ as input to return the corresponding expectation of value function; and (ii) two target state value functions $V_\vartheta(\mathbf{s}_e)$ and $V_{\vartheta^-}(\mathbf{s}_e)$, with parameters ϑ and ϑ^- trained using Eq. 6.16, to improve the stability of the optimization. Each soft Q-function and corresponding target state value function have the same structure and parameterization.

The actor and critic network parameters update come from randomly sampling mini-batch of experiences from the replay buffer. Recall that the actor in the soft actor-critic method generates the mean and standard deviation of a Gaussian probability distribution for each action dimension, where an action is randomly chosen based on this distribution. These generated actions are unbounded since the mean and standard deviation are estimations from the actor network. To bound the action space, the network applies hyperbolic tangent function, such that, the continuous action for each dimension can be bounded within the range, $[-1, 1]$ [77].

Note that the action space for UC-RAN COPs is not continuous but a large discrete multi-dimensional space that can be mapped to continuous action space by means of quantization. To exemplify the continuous action to discrete action mapping, assume that there are three possible discrete actions for S-zone size such that, $A = \{20, 21, 22\}$ meters then the bounded continuous actions are mapped to discrete actions in the following manner: $[-1, -0.33) \rightarrow 20$ meters, $[-0.33, 0.33) \rightarrow 21$ meters, and $[0.33, 1) \rightarrow 22$ meters.[1]

The pseudo-code of risk-oblivious rApp is presented in Algorithm 5. The soft actor-critic network parameters are initialized, such that, each critic and its target (soft Q-function and subsequent state value function) is initialized with same values. The environment is randomly explored in the warm start period by random selection of system actions. The number of warm start epochs is a hyperparameter that will require optimization depending on the type of problem at hand. Once the warm start period is completed, the system actions are selected using the policy $\pi_\psi(.|\mathbf{s}_e)$. System actions are executed in the UC-RAN environment, followed by observing the system next state and reward as feedback from the UC-RAN environment.

[1] In Section 4.5, the S-zone size used for experimental evaluation is discretized into intervals of 1 meter, ranging from 10 to 70 meters.

Algorithm 5: Psuedo-code for risk-oblivious rApp.

```
initialize network parameters $\theta_1, \theta_2, \psi$;
$\vartheta = \theta_1$, $\vartheta^- = \theta_2$;
for each epoch do
    initialize system state $\mathbf{s}_e$;
    if epoch < warm start epochs then
        select system action $\mathbf{a}_e \in A$ randomly
    else
        select system action $\mathbf{a}_e \sim \pi_\psi(.|\mathbf{s}_e)$
    execute system action in the UC-RAN environemnt;
    observe system reward $r_e$ using Eq. 4.10 and obtain next state $\mathbf{s}_{e+1}$
      feedback from UC-RAN environment;
    store experience $(\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1})$ in replay buffer $D$;
    for each gradient step do
        sample experience mini-batches from replay buffer $D$;
        update the soft Q-functions according to Eq. 6.15;
        update the state value functions according to Eq. 6.16;
        update the policy network according to Eq. 6.17;
```

Experience tuple in the form of $(\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1})$ are stored in the replay buffer D. A mini-batch of experiences are sampled randomly to train actor and critic network parameters using stochastic gradient by minimizing the corresponding loss equations derived in Eq. 6.15, Eq. 6.16, and Eq. 6.17.

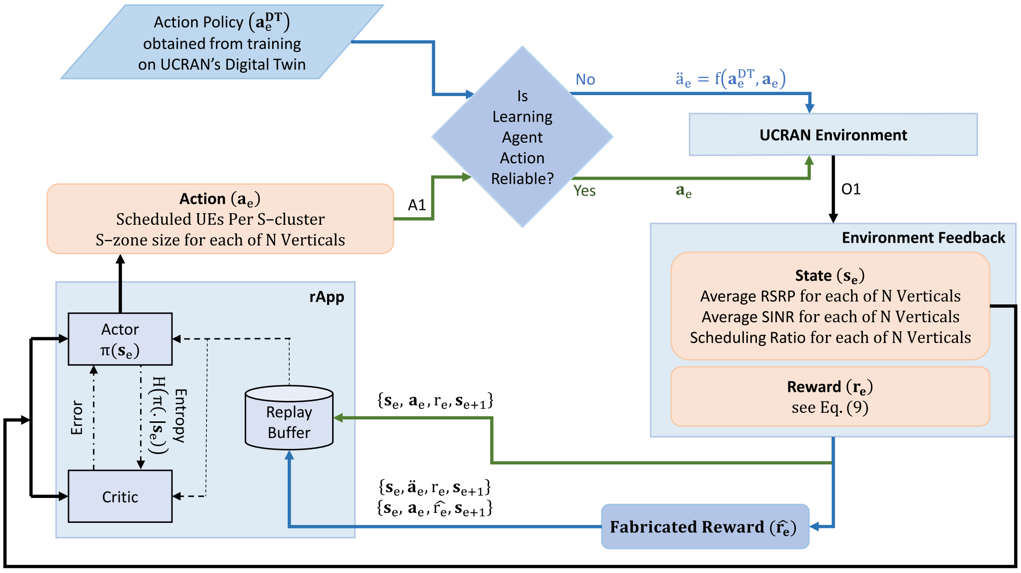
Risk-aware rApp Framework

The proposed risk-oblivious rApp framework operates online to optimize UC-RAN COPs, starting with a warm start during which it explores the solution space. Even though such a data-driven exploratory optimization technique performs the much needed optimization, it does so at the risk of choosing unreliable exploratory actions (COPs) that may harm the system. While optimizing cellular networks in an online fashion, the proposed risk-aware rApp utilizes an offline action policy trained on the DT to foster an improved and risk-aware action policy. Fig. 4.7 shows the risk-aware rApp framework for UC-RAN COPs optimization where the system state, action, reward, and learning algorithm's training procedure is similar to risk-oblivious rApp framework except for three notable differences:

- Under the risk-aware framework, if the learning agent's action is deemed risky, we use the offline-trained DT policy; otherwise, the online learning policy is used.
- A learning agent's action policy is determined to be risky based on the uncertainty in the choice of its actions (measured as high entropy policy) and the divergence of its action policy from that of a DT action policy (measured using Manhattan distance).
- A penalized reward paired with the learning agent's action policy is included in the experience tuple $(\mathbf{s}_e, \mathbf{a}_e, \hat{r}_e, \mathbf{s}_{e+1})$ to encourage the learning agent to avoid risky/unreliable actions.

The pseudo-code of risk-aware rApp is presented in Algorithm 6. The soft actor-critic network parameters initialization, warm start period, and gradient step updates procedure remain the same as the risk-oblivious rApp. Below discussed are a few novel characteristics of the risk-aware rApp.

Offline Policy from Digital Twin: At each epoch e, the risk-aware rApp has access to offline trained action policy obtained from training on UC-RAN's DT. The offline trained policy comprises of system state/action mapping. Note that the cellular network operators have the ability to design and control digital twins, meaning that RL convergence can be tested multiple times on the DT giving the network operator confidence about the superiority of the DT action policy. It is also pertinent to note that Non-RT RIC will have access to the offline trained model since O-RAN specifications define the Non-RT RIC as an enabler to utilize external and contextual information for optimization, which is often unavailable at RANs or near-RT RIC.



Action Policy ($\mathbf{a}_e^{DT}$) obtained from training on UCRAN's Digital Twin
Is Learning Agent Action Reliable?
No
$\ddot{a}_e = f(\mathbf{a}_e^{DT}, \mathbf{a}_e)$
UCRAN Environment
Yes
$\mathbf{a}_e$
O1
A1
Action ($\mathbf{a}_e$)
Scheduled UEs Per S-cluster
S-zone size for each of N Verticals
Environment Feedback
State ($\mathbf{s}_e$)
Average RSRP for each of N Verticals
Average SINR for each of N Verticals
Scheduling Ratio for each of N Verticals
Reward (r_e)
see Eq. (9)
rApp
Actor
$\pi(\mathbf{s}_e)$
Error
Entropy
$H(\pi(\cdot|\mathbf{s}_e))$
Replay Buffer
Critic
$\{\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1}\}$
$\{\mathbf{s}_e, \ddot{\mathbf{a}}_e, r_e, \mathbf{s}_{e+1}\}$
$\{\mathbf{s}_e, \mathbf{a}_e, \hat{r}_e, \mathbf{s}_{e+1}\}$
Fabricated Reward ($\hat{r}_e$)

Fig. 4.7: Risk-aware rApp framework leveraging DT for UC-RAN COPs optimization.

Assessing Risk and Exploration/Exploitation Trade-off Hyperparameter (α): The learning agent's action risk is determined by the action policy's: (i) entropy which determines the uncertainty; and (ii) its distance from the DT's action policy which determines its divergence from the DT action policy. Recall that in soft actor-critic, the output of the policy network is the mean and standard deviation of Gaussian distribution for each action dimension. In our rApps, each action dimension is an individual COP. For instance, a UC-RAN environment with three verticals will have four action dimensions (COPs), where one dimension is for the scheduled UEs per *S-cluster*, and the other three dimensions are for each vertical's S-zone size. In other words, each of these four action dimensions will have a separate mean and standard deviation.

The entropy of a Gaussian policy is defined as $H_{a_i} = 0.5 + 0.5\log(2\pi) + \log(\sigma_{a_i})$, where σ_{a_i} is the standard deviation of the action dimension $a_i \in \mathbf{a}_e$. Therefore, the standard deviation of a Gaussian policy is solely determined by its entropy. To that end, based on the standard deviation of each action dimension, we measure the uncertainty in the learning agent's action and compute the distance between

Algorithm 6: Psuedo-code for risk-aware rApp.

```
initialize network parameters $\theta_1, \theta_2, \psi$;
$\vartheta = \theta_1, \vartheta^{-} = \theta_2$;
for each epoch do
    initialize system state, $\mathbf{s}_e$;
    if epoch $<$ warm start epochs then
        select system action, $\mathbf{a}_e \in A$, randomly
    else
        select system action, $\mathbf{a}_e \sim \pi_\psi(.|\mathbf{s}_e)$
    for each action dimension $a_i \in \mathbf{a}_e$ do
        Reliable Action := $\mathbb{1}_{\big\{\exp\big(-\sigma_{a_i}\big(\big|a_i^{\mathrm{DT}}-\mu_{a_i}\big|\big)>\alpha\big\}}$;
        if Reliable Action $= 1$ then
            $\ddot{a}_i := a_i$;
        else
            $\ddot{a}_i := a_i^{\mathrm{DT}}$;
    execute risk-aware system action, $\ddot{\mathbf{a}}_e$, in the UC-RAN environment;
    observe system reward, $r_e$, using Eq. 4.10 and obtain next state, $\mathbf{s}_{e+1}$,
      feedback from UC-RAN environment;
    if $\ddot{\mathbf{a}}_e = \mathbf{a}_e$ then
        store experience, $(\mathbf{s}_e, \ddot{\mathbf{a}}_e, r_e, \mathbf{s}_{e+1})$, in replay buffer, $D$;
    else
        calculate entropy-dependent hyperparameter $\rho := \sum_{a_i \in \mathbf{a}_e} \sigma_{a_i}$;
        calculate penalized reward $\hat{r}_e :=$ reward $\times$ exp $\Bigg( - \rho \times \|\ddot{\mathbf{a}}_e - \mathbf{a}_e\|_2 \Bigg)$;
        store experience $(\mathbf{s}_e, \ddot{\mathbf{a}}_e, r_e, \mathbf{s}_{e+1})$ and $(\mathbf{s}_e, \mathbf{a}_e, \hat{r}_e, \mathbf{s}_{e+1})$ in replay buffer $D$;
    for each gradient step do
        sample experience mini-batches from replay buffer $D$;
        update the soft Q-functions according to Eq. 6.15;
        update the state value functions according to Eq. 6.16;
        update the policy network according to Eq. 6.17;
```

the learning agent's mean action and the DT's action. Since both the standard deviation and distance cannot be a negative value, these two terms are combined in the power of a negative exponential function to limit its values in the range of 0 and 1. The value of this exponential term is compared against hyperparameter $0 \leq \alpha \leq 1$, which controls the exploration/exploitation tradeoff in a risk-aware rApp framework. The higher the value of α, the more "risk aware" the combined action policy will be. Nevertheless, if the DT action policy is not of high or medium fidelity, doing so would reduce the exploration of the learning algorithm, resulting in convergence to local optima.

It is important to notice that a high-entropy policy does not necessarily mean that a learning agent is taking actions haphazardly, as a high-entropy policy also encourages stochastic behavior in order to train policies that achieve optimal (or near-optimal) with more diverse strategies/trajectories compared to deterministic (or low entropy) policies that always follow the same trajectory. However, performing extreme exploration in a real cellular network with a high-entropy policy during online optimization can lead to deteriorating performance. For this reason, the proposed algorithm allows for the adjustment of policy entropy based on the hyperparameter α. A high value of α will encourage a more stochastic policy, while a low value of α will encourage a more deterministic policy. The optimal value of α will depend on the specific environment, fidelity of DT and the desired level of exploration. In addition to the entropy of action policy, the risk of a policy is also quantified by the divergence of the policy from the DT's action policy. A high divergence indicates that the policy is taking actions significantly different from those taken in the DT by the learning agent. This can be a sign of risk, as it suggests that the policy may not be able to achieve the same level of performance as it achieved in the DT.

Penalized Reward: Even though the action policy from the DT and the learning algorithm is combined to avert risk, the learning algorithm's unreliable action needs to be reprimanded. The penalty is essential to allow the learning algorithm to eventually be certain of its actions without doing unreliable (extreme) exploration of the environment. To do so, a fabricated reward phenomenon is included to give a penalized reward to the learning algorithm's action if it is deemed unreliable. The penalized reward is a scaled version of the original reward. Several factors are taken into account when scaling the reward, such as the uncertainty of the algorithm's action policy and the distance between the risk-aware action and the DT policy's action. Similar to the expression of determining action reliability, we use a negative exponential term to combine these two terms (standard deviation signifying entropy/uncertainty of action policy and Euclidean distance between risk-averse and learning algorithm's action policy). Note that the higher the uncertainty of action policy (standard deviation) and divergence from DT action policy, the lower will be the penalized reward.

Extent of Trust on the Fidelity of Digital Twin: The proposed risk-aware rApp framework does not assume access to a high-fidelity DT of a UC-RAN network. A high-fidelity DT model of a cellular network must encompass the complete system, including the realistic wireless environment, antenna patterns, traffic, mobility, varying user quality of service requirements, handovers, and multi-connections, to name a few. However, with rapid advancements and innovations in cellular technology, the assumption of having access to a high-fidelity DT for all types of cellular networks may not be applicable. This poses a challenge in a scenario when the DT used to avert risk in risk-aware rApp framework is not of a high-fidelity. The above challenge can be addressed by the right choice of exploration/exploitation hyperparameter α, whose natural interpretation as the inverse of the DT's fidelity provides good intuition for adjusting it.

4.5 Performance Evaluation

The performance of the proposed risk-oblivious and risk-aware rApps is evaluated and compared against pseudo brute force maximum using a 3GPP-compliant event-driven system-level DT with advanced features such as heterogeneous traffic generation, user-centric virtual cells, O-RU activation/deactivation, and UE/O-RU association. A pseudo brute force algorithm approximates the maximum for each KPI and multi-objective function formulated in Eq. 4.8 by exhaustively traversing through all possible combinations of COPs with a predefined step size. Given the step size is not too large, the pseudo brute force method is expected to approximate the actual maximum. In addition to comparing with the pseudo brute force maximum, the proposed rApps are evaluated based on risk associated with the optimization.

Risk in this chapter is defined to reflect the idea of reaching to near-optimal in the most efficient way, that is, without harming the system performance. We consider the risk score to be representative of the multi-objective function, such that, the multi-objective function defined in Eq. 4.8 is transformed to a maximization function by subtracting it from 1, a value hereafter referred to as utility equivalent to unscaled system reward. To this end, the risk score is defined as the sum of indicator function values of the utility falling below a certain percentage (q) of maximum achievable utility which is obtained from pseudo brute force. Mathematically,

$$\text{Risk Score} = \sum_{e}^{\text{num epochs}} \mathbb{1}_{\{\text{Utility}_e < \text{q\% of maximum utility}\}}. \tag{4.11}$$

As, no real cellular network based on UC-RAN conforming to O-RAN specifications is yet developed, we resort to our 3GPP-compliant event-driven system-level DT with different UE/O-RU deployments to differentiate the real cellular network from DT used in this chapter. To show the performance of the proposed risk-aware rApp with different fidelity DTs, we use different cellular network parameters such as shadowing, path loss, etc, than those used for real cellular network simulation.

We consider three verticals representing different data/latency requirements. All UEs belong to either of these three verticals. The path loss model and UE/O-RU deployment are adopted in accordance with the system model presented in Section 4.2 with 500 UEs/O-RU per the simulations region of 1 square kilometer. We consider the bandwidths of all sub-channels to be 100MHz and TTI duration to be 1 millisecond. The minimum and maximum values for COPs are: (i) 1 to 3 for the number of scheduled UEs per *S-cluster*; and (ii) 10 meters to 70 meters for the S-zone size of each vertical with step size of 1 meter. Note that the considered number of verticals, their requirements and the range of COPs values are representative use-cases without loss of generality. The target value for each KPI in Eq. 4.10 is set to 1.

The neural network architectures for the actor and critic networks consist of one hidden layer with 256 neurons and a rectified linear unit activation function. The learning weight is set to 0.003, discount factor γ is set to 0.99, target network update frequency is set to 20 epochs, mini-batch size is set to 256, replay buffer size is set to 10000, and warm start period is set to 300 epochs. We use PyTorch to implement the proposed frameworks in Python and average the simulation results for a number of random seed numbers.

The performance evaluation section first analyzes the performance of risk-oblivious rApp on different KPIs optimization. We then compare the performance of risk-aware and risk-tolerant rApps to show the applicability of risk-aware rApp in live cellular network optimization. Later, we analyze the impact of exploration/exploitation hyperparameter α on the convergence and risk-awareness capability of the proposed risk-aware rApp with different fidelity DTs. Finally, the proposed rApps are evaluated with different QoS requirements.

4.5.1 Learning Efficiency Analysis of Risk-oblivious rApp

In the learning efficiency analysis, we evaluate the performance of the proposed risk-oblivious rApp in terms of convergence to pseudo brute force maximum and risk score associated with the optimization. To prove the applicability of the proposed risk-oblivious rApp, we optimize individual KPIs (such as latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency) and a combination of KPIs in the form of utility with $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = 0.25$.

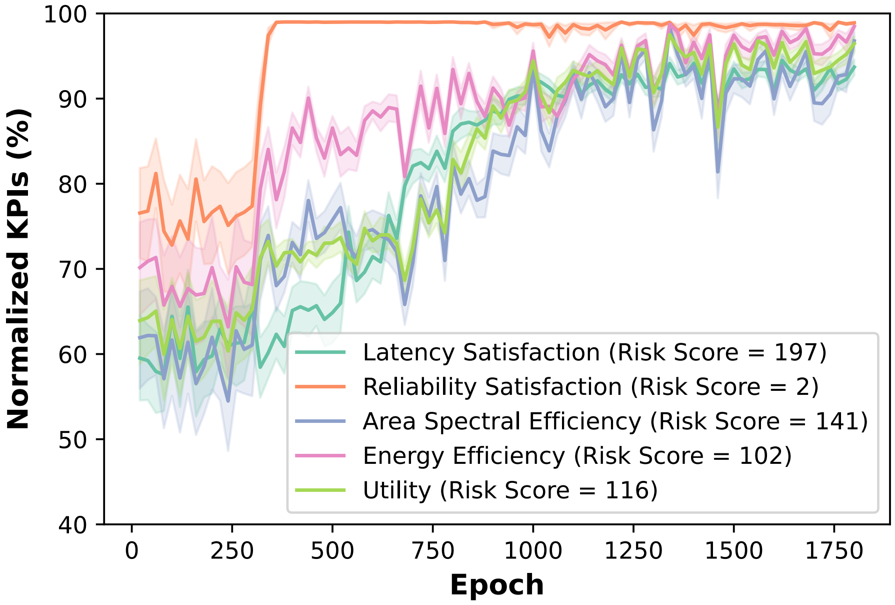


Fig. 4.8: Evaluating the risk-oblivious rApp convergence for normalized key performance indicators as a function of number of epochs.

In Fig. 4.8, we show the normalized KPIs score of the risk-oblivious rApp across a number of training epochs with different KPIs. As shown in the figure, due to random exploration during the warm-up period (0-300 epochs), the normalized KPIs start with rather sub-optimal values. As the training progresses, the normalized KPIs score improves until it converges to the maximum achievable KPIs score. Note that the maximum achievable KPIs score are obtained from the pseudo brute force method. This indicates that the proposed risk-oblivious rApp can converge to near-optimal on an average of 1200 epochs (except for reliability satisfaction which converges much quicker due to the large optimal solution space as observed in Fig. 4.2b). Even though the proposed risk-oblivious rApp can converge to near-optimal, it does so at some risk (defined in Eq. 4.11 for $q = 60$). The risk score associated with optimizing these KPIs varies; the more complex the solution space, the more the risk score.

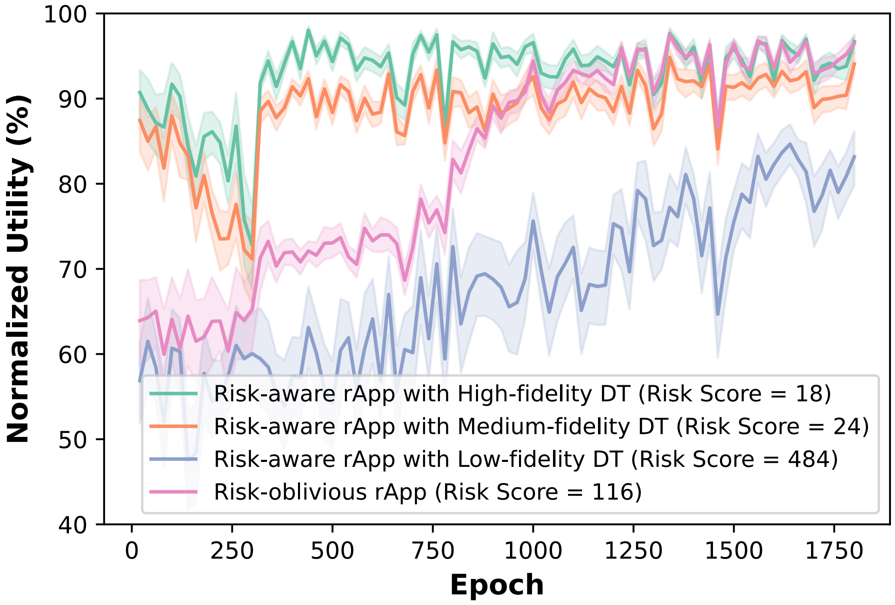
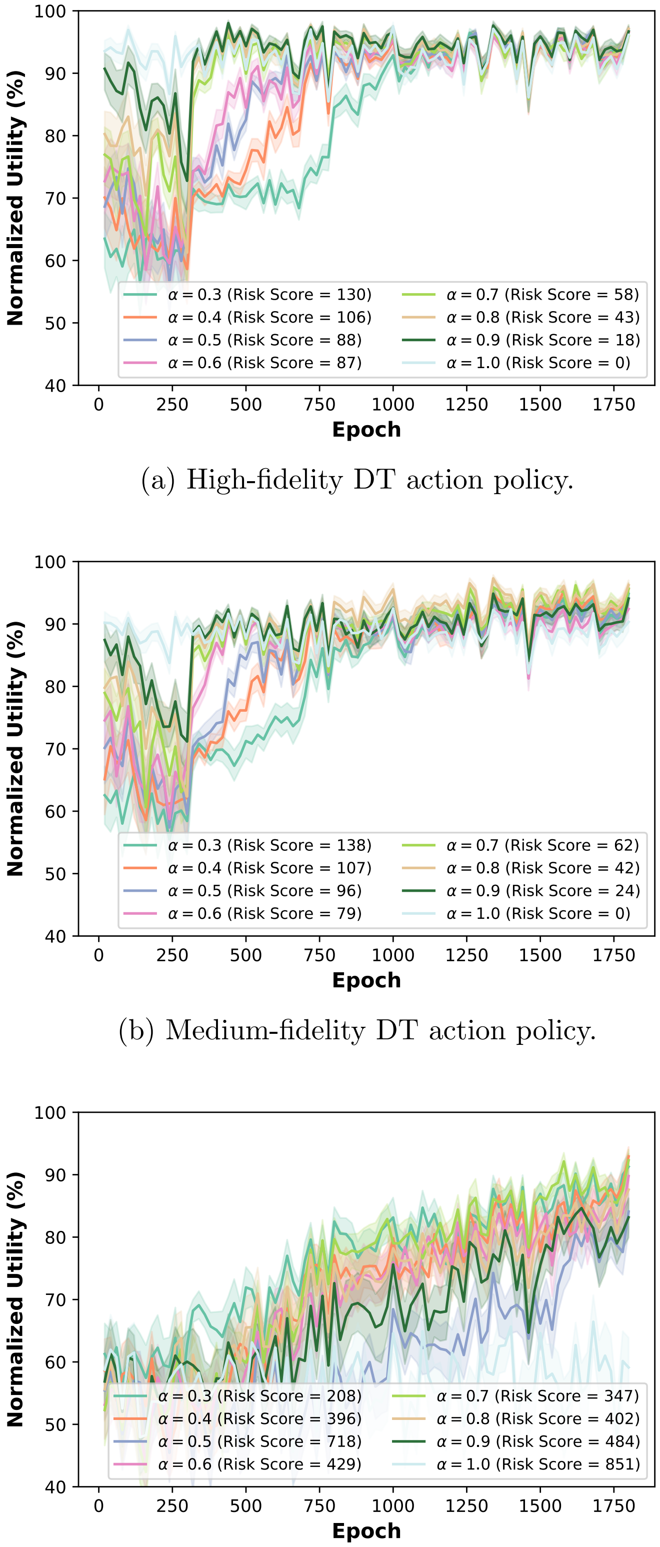


Fig. 4.9: Different fidelity DT impact on risk-aware rApp convergence.

4.5.2 Optimization in an Efficient Way Using Risk-aware rApp

In practical cellular networks, where network operators can afford such exploratory optimization as indicated by risk-oblivious approach, it is highly desirable to reach the optimal in a reliable way; hence the risk-aware rApp comes into play. In the following experiment, soft actor-critic is used as the underlying optimization algorithm for obtaining an offline action policy trained on DT. The DT action policy can be categorized into three levels: high-fidelity, medium-fidelity, and low-fidelity. The DRL agent is trained on these different fidelity DTs for 1500 epochs and the state/action mapping obtained at the last epoch is used as the offline action policy for risk-aware rApp.

From the training curves shown in Fig. 4.9 with utility to optimize, it can be observed that for high and medium fidelity DTs; risk-aware rApp performs optimization with much lesser exploration and quick convergence to maximum derived from pseudo brute force. It can also be observed that risk-aware rApp reduces the risk



(a) High-fidelity DT action policy.

(b) Medium-fidelity DT action policy.

(c) Low-fidelity DT action policy.

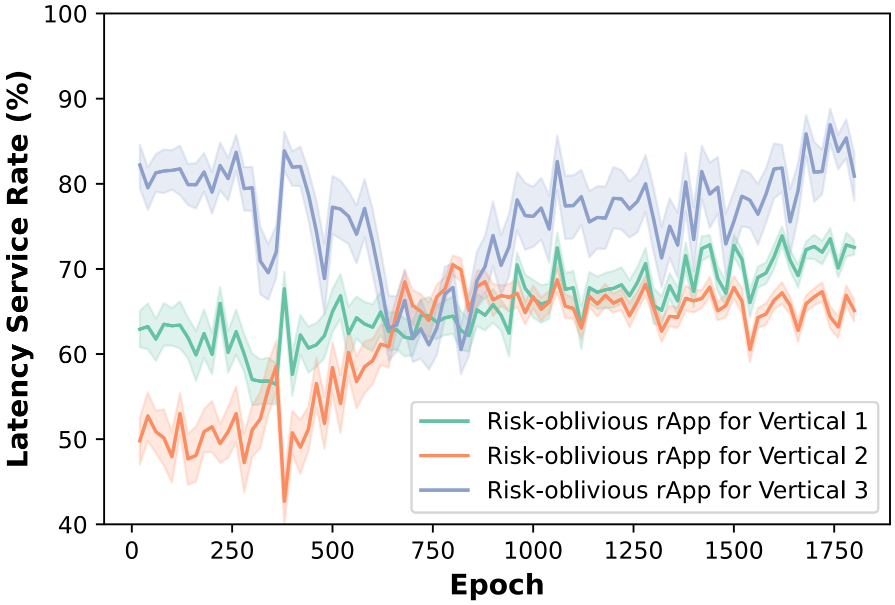
Fig. 4.10: Impact of exploration/exploitation hyperparameter in risk-aware rApp.

score over risk-oblivious rApp by a factor of ten. These results confirm that leveraging DT (with high and medium fidelity) for obtaining offline action policy on top of a state-of-the-art learning algorithm (such as soft actor-critic) significantly reduces the risk score without compromising convergence. However, when using a low-fidelity DT, the risk-aware rApp approach fails to reduce risk score or converge to maximum. This is an expected outcome since the risk-aware rApp leverages the DT action policy to circumvent the unreliable exploration during the warm start period by imitating the DT's action policy. The results show that by integrating DT technology, the environment can be explored with reliability and the warm-up time for optimizing real cellular networks can be minimized.

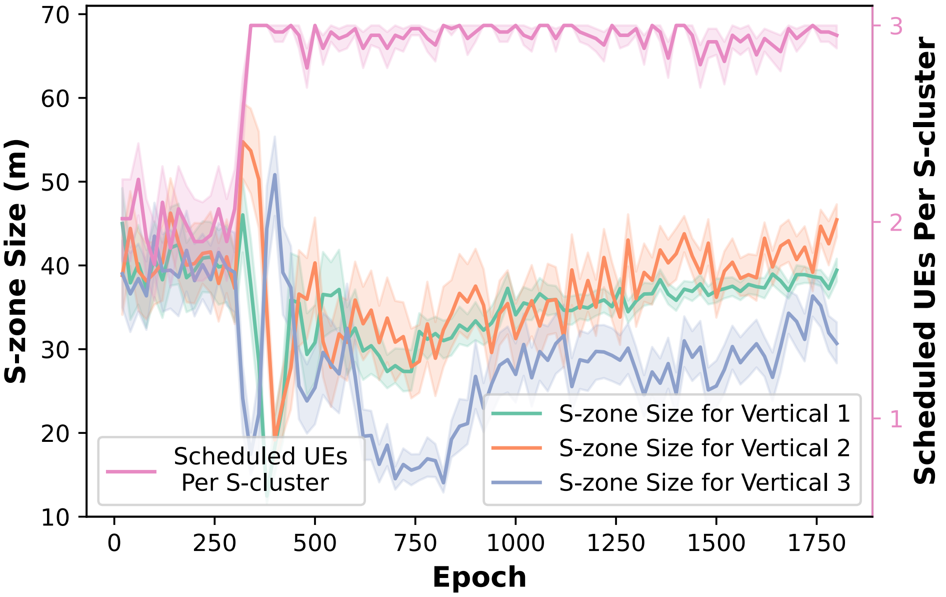
4.5.3 Balancing the Exploration/Exploitation Tradeoff

We use the exploration/exploitation tradeoff hyperparameter α used in the risk-aware rApp to control the exploration depending on DT fidelity level. In Fig. 4.10, we use the different values of α for each of high, medium, and low fidelity DT action policies to analyze its impact on convergence to its utility function's maximum.

Intuitively, the higher the value of α, the more trusted the DT action policy, hence less exploration in the risk-aware rApp. Consequently, this will lead to poor convergence if the DT action policy is of low-fidelity. Contrarily, the lower the value of α, the lower will be the confidence upon the DT action, and the learning algorithm will rely on its exploration to optimize the solution space at the cost of choosing unreliable actions during exploration. This impact of α can be observed in Fig. 4.10, where the higher value of α (i.e. $\alpha = 1$) works best for high-fidelity DT action policy. The medium range value of α (i.e. $\alpha = 0.8$) works for medium-fidelity DT action policy. Similarly, the low values of α (i.e. $\alpha = 0.3$) work best for low-fidelity DT action policy. Notice that even though when a low-fidelity DT is used, α parameter can recover the convergence capability of the optimization framework albeit at a



(a) Latency service rate of verticals.



(b) S-zone size of verticals and scheduled UEs per *S-cluster*.

Fig. 4.11: Evaluating priority access capabilities of UC-RAN for different verticals.

high risk score. The above results support our earlier claim of the choice of α being the inverse of the quality of DT.

4.5.4 Priority Access Capabilities

Despite the obvious utility of serving vertical use cases from a common core and RAN architecture, the diverse requirements of these verticals are often addressed from the perspective of the physical layer technologies. To show the efficacy of UC-RAN in enabling priority access to different verticals, we show the latency service rate for each vertical in Fig. 4.11. Mathematically,

$$\text{Latency Service Rate} = \frac{\sum_{j=1}^{|\Phi_i|} \mathbb{1}_{\left\{\left(\sum_{\tau=T_{ij}}^{T_{ij}+l_i} \Gamma_{ij\tau}\right) \geq \gamma_i\right\}}}{|\Phi_i|}, \tag{4.12}$$

where $|\Phi_i|$ represent the number of UEs belonging to each vertical $i \in N$, $\tau = T_{ij}$ is the TTI at which UE j belonging to vertical i requests service, l_i is the latency requirement for each vertical, γ_i is the data rate requirement for each vertical, and $\Gamma_{ij\tau}$ represents the measured data rate at UE j belonging to vertical i during each TTI.

The three considered verticals have varied latency requirements, with vertical 1 having a medium latency requirement, vertical 3 having a low latency requirement, and vertical 2 having a tolerant latency requirement. These latency requirements can be incorporated in the network operator-defined weights in Eq. 5.10 with weights natural interpretation as the inverse of the latency requirement of verticals; that is, the lower the latency requirement, higher the weight. From the curves shown in Fig. 4.11a, we can observe that the network operator-defined weights can enable prioritized access to UEs belonging to vertical 3 with the lowest latency requirement. The respective convergence of S-zone size of verticals and scheduled UEs per *S-cluster* is shown in Fig. 4.11b, where the vertical 1 is allocated the S-zone size of approximately 40 meters, vertical 2 is allocated the S-zone size of approximately 35 meters and vertical 3 is allocated the S-zone size of approximately 30 meters. Recall that the smaller S-zone size and larger scheduled number of UEs per *S-cluster* lead to higher latency satisfaction which can be observed in these plots. Also, the scheduled UEs per *S-cluster* is increased to maximum allowed limit of 3 which as observed in Fig. 4.3a maximizes the latency satisfaction metric. With prioritized access capabilities, the proposed rApps can adjust the S-zone size of respective verticals and scheduled UEs per *S-cluster* to prioritize specific vertical(s), ensuring reliable communication without competing with other verticals.

4.6 Conclusion

To cope with the performance deterioration risk associated with online network reconfiguration, which has hindered the industry uptake of online learning-based solutions, we propose offline learning leveraging a DT instilling risk-awareness in the DRL optimization framework. The proposed DT-assisted DRL framework's convergence and accumulated risk are compared against brute force results, showing an impressive performance in reaching the near-optimal in a few hundred iterations. Furthermore, the risk-aware optimization framework indicates the viability of online learning techniques in live cellular networks with controlled and reliable exploration. On the other hand, the multifaceted requirements of 5G and beyond applications demand for a system capable of supporting multiple combinations of prioritized access, reliability, throughput, and energy efficiency. In this chapter, we present and evaluate a UC-RAN based on O-RAN architecture to serve the QoS requirements of various verticals. We introduce and investigate two key UC-RAN configuration and optimization parameters; 1) size of user centric virtual cells (S-zones), and 2) number of UEs scheduled per S-zone, which are leveraged through an rApp to control latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency. Overall, this chapter demonstrates a highly flexible O-RAN-based user-centric architecture coupled with a DT-empowered risk-aware DRL optimization framework that can address the fundamental tradeoff between latency, reliability, and throughput with accelerated and reliable optimization in live emerging cellular networks.

CHAPTER 5

User-centric Communication with Aerial Network for 6G: A Reinforcement Learning Approach

5.1 Introduction

5.1.1 Motivation

UC-RAN, introduced in recent works [2–4, 7, 8] and previous chapters, utilize virtual Lean, Elastic, Agile, and Proactive Service exclusion Zones (LEAP S-Zones) to provide very high-throughput within ultra-low latency bounds. While the virtual LEAP S-Zones-based architecture has significantly improved the quality of service for high-priority verticals, it often negatively impacts low-priority verticals due to the preemptive scheduling of high-priority verticals with large service exclusion zones, which delays the scheduling of low-priority verticals [8]. This problem is further exacerbated in hotspot areas where the delay in scheduling will further increase due to the service exclusion zones.

Aerial deployment can provide a potential cost-efficient solution to this problem, with an overlay network deployment to provide service to all verticals and improve user satisfaction [83]. This approach of integrating virtual LEAP S-Zones-based terrestrial architecture with an aerial network can offer dual advantages: (i) seamless coverage for diverse verticals with varying needs, where the terrestrial network caters to high-priority verticals for next-generation use cases, and the aerial network serves low-priority verticals that experience scheduling delays from the terrestrial network; and (ii) cost-effective provision of rapid on-demand communications through the agile deployment of aerial base stations (ABSs), allowing for greater flexibility in movement between locations, making it an ideal solution for next-generation wireless systems [84].

To fully exploit the benefits of aerial network-assisted communication, it is imperative to intelligently control the configuration and optimization parameters (COPs) of ABSs, including their locations, transmit powers, altitudes, and beamwidths. However, the increase in the number of ABSs in the aerial network leads to an exponential increase in the combinations of these COPs, which can result in scalability issues for heuristic and brute force optimization approaches. Furthermore, the dynamicity of the cellular network due to user mobility and traffic demands requires these COPs to be adjusted dynamically. To overcome these challenges, we propose a comprehensive analysis of the impact of aerial network COPs on key performance indicators (KPIs), including coverage, latency satisfaction, average spectral efficiency, and energy efficiency. We propose a DRL optimization framework, henceforth referred to as AIR-DRL: AerIal netwoRk Optimization through DRL, that enables intelligent control of these COPs to optimize multiple KPIs jointly in a dynamic cellular environment. Our ultimate objective is to introduce a novel two-tier cellular network architecture that synergistically combines an artificial intelligence-empowered aerial network with a user-centric terrestrial network, with the goal of fulfilling the evolving requirements of emerging verticals.

While the existing studies mentioned in related work section in 1.3 have addressed the optimization of specific KPIs for ABS-aided wireless communications, they do not consider multi-objective optimization of system-level KPIs that are crucial for 6G networks. Such KPIs include, but are not limited to, coverage, latency satisfaction, average spectral efficiency, and energy efficiency. Additionally, to the best of our knowledge, no previous work has investigated the performance of ABSs in conjunction with a LEAP S-Zones-based terrestrial architecture which has been shown in [2–4, 7, 8] to provide services to various verticals based on their diverse service requirements.

5.1.2 Contributions

The key contributions of this chapter can be summarized as below:

- We integrate an aerial network with the LEAP S-Zones-based terrestrial architecture to fully leverage its advantages. We analyze the impact of this integration on a comprehensive set of KPIs, including coverage, latency satisfaction, average spectral efficiency, and energy efficiency. Through this analysis, we demonstrate that our integrated architecture can provide services to all verticals for 6G networks, unlike traditional heterogeneous networks and standalone LEAP S-Zones-based UC-RAN.
- To balance the tradeoff between key aerial network design parameters, such as ABSs locations, transmit powers, altitudes, and beamwidths, we formulate a multi-objective optimization problem. Moreover, we discuss two methods of combining multiple KPIs (linear optimization and targeted optimization) to ensure each KPI has an equal contribution in the multi-objective formulation, rather than biased optimization towards an individual or a few KPIs.
- Analytical approaches and offline learning solutions are complex and inefficient for performing real-time online optimization. Therefore, we formulate the system multi-objective optimization problem as a Markov Decision Process and propose the AIR-DRL framework based on the state-of-the-art soft actor-critic algorithm to solve it.
- To evaluate the convergence of our proposed AIR-DRL framework against brute force results, we conduct numerous experiments. Our results show that the proposed framework converges system-level KPIs in just a few thousand iterations without biasing towards any individual or a set of KPIs. Further,

we evaluate the performance of the proposed AIR-DRL framework in a much larger solution space, showcasing the utility of our approach beyond brute force and heuristic methods.

5.1.3 Chapter Organization

The chapter is structured as follows. In Section 5.2, we introduce the system model, followed by the formulation of a multi-objective optimization problem with ABS COPs and system-level KPIs in Section 5.3. In Section 5.4, we present a Markov decision process formulation of the optimization problem and propose an AIR-DRL framework to control the ABS COPs intelligently to optimize system-level KPIs jointly. Section 5.5 presents the results of experiments that demonstrate the effectiveness of the proposed AIR-DRL framework. Finally, we conclude the chapter in Section 5.6.

5.2 System Model

The system model, as shown in Fig. 5.1, considers three separate verticals to cater to the diverse needs of users. The categorization of UEs into these verticals is based on their specific throughput and latency requirements. The model defines M hotspot regions located at $(x_m^{\text{hotspot}}, y_m^{\text{hotspot}})$, where $m \leq M$. Since the hotspots can change spatiotemporally due to user mobility, their number and locations are considered variables. With a densely deployed terrestrial network, the size of LEAP S-Zones, defined by the radius of the virtual circle, determines the minimum region where no other UE can be scheduled. While this approach can provide high throughput with ultra-low latency to high-priority verticals, as shown in [8], it can also create coverage holes in the network for low-priority verticals.

To provide coverage to unscheduled UEs, a set $\mathcal{N}$ of N ABSs, indexed by $n =$

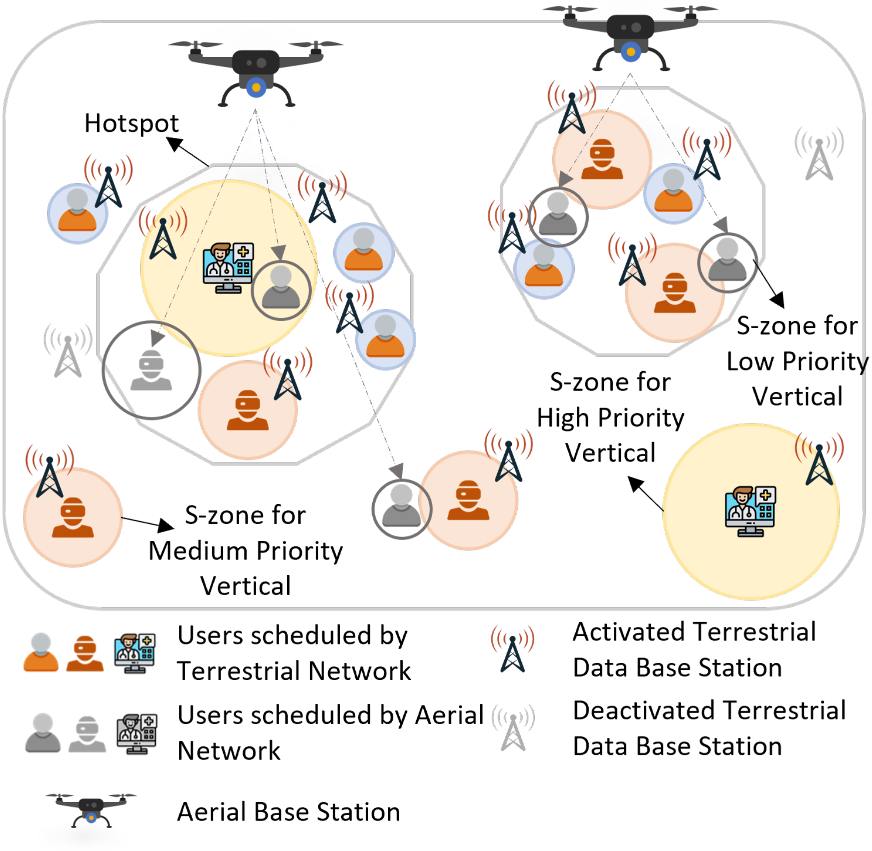


Fig. 5.1: Integrated aerial network with terrestrial user-centric RAN. Three different verticals are defined for normal, meta-verse, and telemedicine terrestrial users.

$1, 2, ..., N$, with altitude H_n^{ABS} with horizontal and vertical location $(x_n^{\mathrm{ABS}}, y_n^{\mathrm{ABS}})$ are deployed and connected to a centralized baseband processing unit to obtain power supply. A set $\mathcal{U}$ of U UEs, indexed by $u = 1, 2, ..., U$, with fixed height H_u^{UE} are distributed using a Poisson point process Φ_{UE_1} with density λ_{UE_1} and a Poisson cluster process Φ_{UE_2} with density λ_{UE_2}. The horizontal and vertical locations of the UEs can be denoted as $(x_u^{\mathrm{UE}}, y_u^{\mathrm{UE}})$. Similarly, a set $\mathcal{O}$ of O data base stations (DBSs), indexed by $o = 1, 2, ..., O$, with fixed height H_o^{DBS} are distributed using a Poisson point process Φ_{DBS_1} with density λ_{DBS_1} and a Poisson cluster process Φ_{DBS_2} with density λ_{DBS_2}. The horizontal and vertical locations of the DBSs can be denoted as $(x_o^{\mathrm{DBS}}, y_o^{\mathrm{DBS}})$. Note that hotspots' centers are set as the parent location for Poisson cluster processes Φ_{UE_2} and Φ_{DBS_2}. In the following, we discuss the aerial channel path loss model and terrestrial channel path loss model where the downlink scenario is considered in sub-6 GHz frequency bands, and the centralized baseband processing unit shares the backhaul spectrum equally between aerial network and LEAP S-Zones-based terrestrial network to minimize cross-tier interference.

5.2.1 Aerial Channel Path Loss Model

ABSs are equipped with directional transmit antennas with gains expressed as:

$$G_{nu} = 2(\kappa_n + 1)\cos^{\kappa_n}(\theta_{nu}), \tag{5.1}$$

where $\kappa_n = \frac{-\log(2)}{\log(\cos(B_n/2))}$ defines the maximum directivity of the antenna of ABS n with beamwidth B_n, $\theta_{nu} = \cos^{-1}\left(\frac{H_n^{\text{ABS}} - H_u^{\text{UE}}}{d_{nu}^{\text{2D}}}\right)$ defines the radiation angle between ABS n and UE u with $\theta_{nu} \in [-\pi/2, \pi/2]$, $d_{nu}^{\text{2D}} = \sqrt{(x_n^{\text{ABS}} - x_u^{\text{UE}})^2 + (y_n^{\text{ABS}} - y_u^{\text{UE}})^2}$ is two-dimensional Euclidean distance between ABS n and UE u, H_n^{ABS} and H_u^{UE} are the heights of ABS n and UE u, respectively. It should be noted that the antenna gain between ABS n and UE u is symmetric along the vertical direction, meaning it is independent of the azimuth angle. This symmetry is commonly observed in antenna designs such as horn or uniform linear array antennas [85, 86].

Further, the ABS-to-ground channel can be characterized in terms of probabilities of line-of-sight (LoS) and non-line-of-sight (NLoS) scenarios between ABS n and UE u. These LoS probabilities can be estimated as follows [84]:

$$\text{LoS}_{nu} = 0.01a - \frac{0.01(a-b)}{1 + \left(\frac{\theta_{nu} - c}{d}\right)^e}, \tag{5.2}$$

where (a, b, c, d, e) are the set of environment-dependent empirical parameters which are given for high-rise urban scenario as $a = 352, b, = -1.37, c = -53, d = 173.8$, and $e = 4.67$ [84]. Similarly, the probability of NLoS scenario between ABS n and UE u can be expressed as $\text{NLoS}_{nu} = 1 - \text{LoS}_{nu}$.

The signal transmitted from the ABS n to UE u is modeled to be affected not only by free space path loss but also by radiation angle-dependent shadowing, whose mean and standard deviation can be modeled as follows [84].

$$\mu_{nu}^{\text{shadowing}} = \frac{q_\mu + \theta_{nu}}{r_\mu + s_\mu \theta_{nu}}, \tag{5.3}$$

$$\sigma_{nu}^{\text{shadowing}} = \frac{q_\sigma + \theta_{nu}}{r_\sigma + s_\sigma \theta_{nu}}, \tag{5.4}$$

where $(q_\mu, r_\mu, s_\mu, q_\sigma, r_\sigma, s_\sigma)$ are frequency-dependent empirical parameters which are given for 3.5 GHz frequency band as $q_\mu = -92.90, r_\mu = -3.14, s_\mu = 0.0302, q_\sigma = -89.06, r_\sigma = -8.63, s_\sigma = 0.0921$ [84].

The received signal strength at UE u from ABS n in LoS and NLoS scenarios, as a function of path loss and antenna gain, can be expressed as:

$$R_{nu}^{\text{LoS}} = T_n - 20 \log \left(\frac{4\pi f_n d_{nu}^{\text{3D}}}{c} \right) + G_{nu} - X_{nu}^{\text{LoS}}, \tag{5.5}$$

$$R_{nu}^{\text{NLoS}} = T_n - 20 \log \left(\frac{4\pi f_n d_{nu}^{\text{3D}}}{c} \right) + G_{nu} - X_{nu}^{\text{NLoS}} - X_{nu}^{\text{Shadowing}}, \tag{5.6}$$

where R_{nu}^{LoS} and R_{nu}^{NLoS} are the received signal strength (in dBm) at UE u from ABS n for LoS and NLoS communication paths, T_n is the transmit power (in dBm) of ABS n, c is the speed of light, f_n denotes the carrier frequency of ABS n, $d_{nu}^{\text{3D}} = \sqrt{(x_n^{\text{ABS}} - x_u^{\text{UE}})^2 + (y_n^{\text{ABS}} - y_u^{\text{UE}})^2 + (H_n^{\text{ABS}} - H_u^{\text{UE}})^2}$ is the three-dimensional Euclidean distance between ABS n and UE u, $X_{nu}^{\text{Shadowing}}$ is shadow fading represented as Gaussian random variable with mean $\mu_{nu}^{\text{shadowing}}$ and standard deviation $\sigma_{nu}^{\text{shadowing}}$, X_{nu}^{LoS} and X_{nu}^{NLoS} are location dependent randomness in the received signal represented as log normal distribution with mean zero and standard deviation σ_{nu}^{LoS} and $\sigma_{nu}^{\text{NLoS}}$ (in decibels), respectively. The variable $X_{nu}^{\text{Shadowing}}$ is only included in NLoS scenario because shadowing is a phenomenon that occurs exclusively in NLoS scenarios due to the presence of obstacles that affect wave propagation.

5.2.2 Terrestrial Channel Path Loss Model

The DBS o, which provides best cell coverage, serves the UE u by utilizing the full bandwidth allocated to the terrestrial network. The centralized baseband processing unit deactivates any DBSs that are not associated with any UE. The DBS schedules UEs in each transmission time interval (TTI) according to its scheduling priorities only if: (i) there are no other scheduled users within a UE's LEAP S-Zone; and (ii) there is at least one available DBS with which it can associate. The UE priorities are determined by their latency requirements, meaning that the lower the latency requirement, the higher the scheduling priority. It is important to note that the vertical-specific LEAP S-Zones are adjusted by the network operator to meet the needs of each corresponding vertical. These values must be set appropriately as they serve as proxy parameters that control interference between scheduled UEs.

The received signal strength between DBS o and UE u can be modeled as:

$$R_{ou} = T_o - PL_{ou} + G_{ou} - X_{ou}^{\text{Shadowing}}, \tag{5.7}$$

where T_o is the transmit power (in dBm) of DBS o, G_{ou} is the antenna gain between DBS o and UE u, $X_{ou}^{\text{Shadowing}}$ is shadow fading represented as Gaussian random variable with zero mean and standard deviation of 4 dB (see [78]), and PL_{ou} is the linear dual slope path loss model derived from the Third Generation Partnership Project (3GPP) Technical Report 38.901 UMi Street Canyon line-of-sight model [78]. The non-linear dual slope path loss (in dB) is expressed as follows:

$$PL_{ou} = \begin{cases} PL_1; & 10m \leq d_{ou} \leq d_{\text{breakpoint}} \\ PL_2; & d_{\text{breakpoint}} < d_{ou} \leq 5km \end{cases}, \tag{5.8}$$

where $PL_1 = 32.4 + 21\log_{10}(d_{ou}) + 20\log_{10}(f_o)$, $PL_2 = 32.4 + 40\log_{10}(d_{ou}) + 20\log_{10}(f_o) - 9.5\log_{10}(d_{\text{breakpoint}}^2 + (H_o^{\text{DBS}} - H_u^{\text{UE}})^2)$, $d_{\text{breakpoint}}$ is the breakpoint distance, f_o is the carrier frequency of DBS o, H_o^{DBS} is the height of serving DBS o, and $d_{ou}^{\text{3D}} = \sqrt{(x_o^{\text{DBS}} - x_u^{\text{UE}})^2 + (y_o^{\text{DBS}} - y_u^{\text{UE}})^2 + (H_o^{\text{DBS}} - H_u^{\text{UE}})^2}$ is the three-dimensional Euclidean distance between DBS o and UE u.

5.3 Problem Formulation

In this section, we discuss the KPIs used to assess system performance, followed by the formulation of multi-objective optimization problems.

5.3.1 Key Performance Indicators

The system performance in this chapter is assessed in terms of coverage, latency satisfaction, average spectral efficiency, and network energy efficiency.

Coverage

Coverage is measured as the percentage of UEs receiving signal-to-interference-noise ratio (SINR) beyond a minimum threshold. Achieving the desired level of coverage in terms of SINR is a critical aspect of network planning and optimization and requires careful selection of ABS COPs such as placement, transmit power, height, and beamwidth. Mathematically,

$$\text{Coverage} = \mathbb{E}_\tau \left[\sum_{u \in \mathcal{U}}^{U} \mathbb{1}_{\left\{\Gamma_{u\tau} \geq \gamma\right\}} \right] \tag{5.9}$$

where $\Gamma_{u\tau} = \frac{R_{iu}}{N_0 + \sum_{\substack{i' \in \mathcal{I}, \\ i' \neq i}} R_{i'u}}$ represents the measured SINR at UE u served by terrestrial network ($i = o$ and $\mathcal{I} = \mathcal{O}$) or aerial network ($i = n$ and $\mathcal{I} = \mathcal{N}$) during each TTI τ, $\mathbb{1}_{\{.\}}$ is the characteristic function, U is the number of users, γ is the minimum SINR requirement for each user, N_0 denotes the noise power, and $\mathbb{E}_\tau[.]$ represents averaging over several TTIs.

Latency Satisfaction

The definition of user plane latency, according to 3GPP, refers to the time required for unidirectional data transfer from the access point's radio protocol layer to the UE's radio protocol ingress point, assuming the UE is in an active state [79]. To determine whether a given network meets the 3GPP's definition of latency, we quantify latency satisfaction as the weighted sum of the percentage of UEs from each vertical, which are served with the required data rate within their latency constraints. Mathematically,

$$\text{Latency Satisfaction} = \sum_{v=1}^{V} \dot{w}_v \left(\frac{\sum_{u_v=1}^{U_v} \mathbb{1}_{\left\{ \left(\sum_{\tau=T_{u_v}}^{T_{u_v}+l_v} \Omega_{u_v\tau} \right) \geq \omega_v \right\}}}{U} \right), \tag{5.10}$$

where V is the number of verticals, U_v represents the number of UEs belonging to vertical v, T_{u_v} is the TTI at which UE u belonging to vertical v requests service, l_v is the latency requirement for each vertical, ω_v is the data rate requirement for each vertical during each TTI, $\Omega_{u_v\tau}$ represents the measured data rate at UE u belonging to vertical v during each TTI, and $\dot{w}_v \geq 0, \forall v$ and $\sum_{v=1}^{V} \dot{w}_v \leq 1$ are network operator-defined weights assigned to prioritize latency requirements of specific verticals. Note that latency satisfaction metric is calculated for every Z TTIs where $Z = \max(l_v),\ \forall v \leq V$.

Average Spectral Efficiency

The average spectral efficiency is the time-averaged spectral efficiency per bandwidth channel. Mathematically,

$$\text{Average Spectral Efficiency} = \frac{\mathbb{E}_{\tau} \left[\sum_{u=1}^{U} \Omega_{u\tau} \right]}{B}, \tag{5.11}$$

where $\Omega_{u\tau}$ represents the measured data rate at UE u during each TTI and B is the channel bandwidth.

Network Energy Efficiency

The energy consumption of a network relies on two crucial factors: area spectral efficiency and power consumption. In this chapter, we adopted an ABS mechanical power consumption model inspired by [87]. This model provides an estimation of the energy needed for ABS to stay in air and propel itself forward, effectively counteracting the forces of gravity, wind, and air density. Expanding upon the model proposed in [87], we define the minimum mechanical power required for ABS n to achieve forward motion as follows:

$$P_n^{mech_{min}} = (vel_n^i + vel_n \sin \alpha_n) D_n, \tag{5.12}$$

where vel_n^i is the induced velocity required for a given thrust D_n, vel_n is the average ground speed of the ABS and α_n is the pitch angle of n^{th} ABS. The required thrust D_n of n^{th} ABS is given as:

$$D_n = (mass_n^{body} + mass_n^{batt})g + F_n^{drag}, \tag{5.13}$$

where $mass_n^{body}$ is the mass of ABS body, $mass_n^{batt}$ is the mass of battery, g is the gravitational constant and F_n^{drag} is the total drag force of n^{th} ABS. The drag force is estimated as follows:

$$F_n^{drag} = 1/2 \rho vel_a^2 (C_n^{body} A_n^{body} + C_n^{batt} A_n^{batt}), \tag{5.14}$$

where ρ is the density of air, vel_a is the velocity in air, C_n^{body}, C_n^{batt} and A_n^{body}, A_n^{batt} are the drag coefficients and projected area of n^{th} ABS body and battery, respectively. Given the drag force, the pitch angle α_n of n^{th} can be expressed as follows:

$$\alpha_n = \tan^{-1}\left(\frac{F_n^{drag}}{(mass_n^{body} + mass_n^{batt})g}\right). \tag{5.15}$$

With the derivation of the above terms, the induced velocity can be calculated by solving the following non-linear equation:

$$vel_n^i = \left(\frac{2D_n}{\pi by\rho\sqrt{(vel_n \cos\alpha_n)^2 + (vel_n \sin\alpha_n + vel_n^i)^2}}\right), \tag{5.16}$$

where y is the number of rotors with diameter b. Therefore, the theoretical mechanical power consumption in Eq. 5.12 can be used to find the expanded power by dividing the minimum power consumption with power efficiency ε of ABS, such that $P_n^{mech} = P_n^{mech_{min}}/\varepsilon$. The total flying energy required for n^{th} ABS to travel distance d_n can be given as $P_n^{mech} = P_n^{mech} d_n v_n$. Note that the above formulation depends on the altitude of ABS as air density changes with altitude. With this said, the total power conumsption of n^{th} ABS can be given as:

$$P_n = P_n^{comm} + P_n^{mech}, \tag{5.17}$$

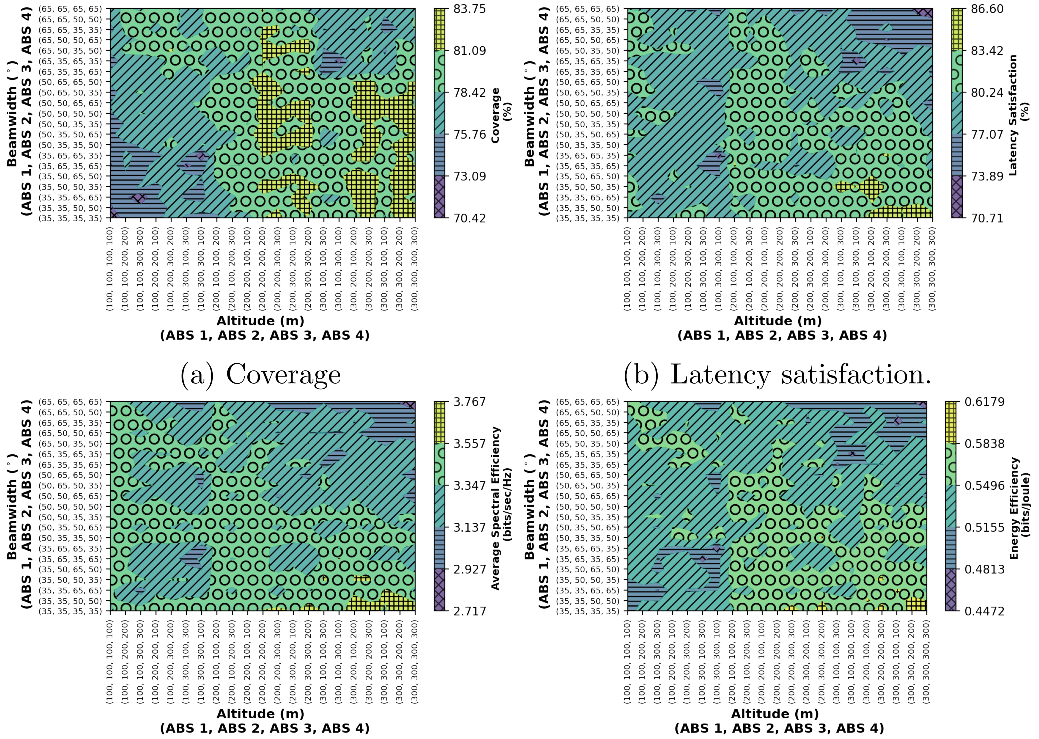
where P_{comm} is the power consumed during communication which depends on the transmit power of ABSs.

Therefore, energy efficiency can be defined as:

$$\text{Energy Efficiency} = \frac{\text{Area Spectral Efficiency}}{\sum_n^N P_n}. \tag{5.18}$$

5.3.2 Impact of ABS COPs on KPIs

In this section, we explore the impact of altitude and beamwidth of ABS COPs on KPIs such as coverage, latency satisfaction, average spectral efficiency, and energy



(a) Coverage

(b) Latency satisfaction.

(c) Average spectral efficiency.

(d) Energy efficiency.

Fig. 5.2: Impact of ABS COPs (altitude, beamwidth) on coverage, latency satisfaction, average spectral efficiency, and energy efficiency.

efficiency. To investigate the relation between COPs and KPIs, we conduct simulations of four hotspots with four ABSs positioned at the center of each hotspot, where the transmit power of each ABS was fixed at 23 dBm. We varied the altitude values within the range of 100 and 400 meters, and the beamwidth values within the range of 35 and 65 degrees in a network spanning 1 square kilometer.

First, we examine the influence of ABS altitude on KPIs. As the altitude of ABS directly affects the signal strength received by UEs, it has a significant impact on coverage, latency satisfaction, average spectral efficiency, and energy efficiency. Higher altitudes result in larger coverage areas, but they also lead to a reduction in the received signal strength at UE and increased power consumption, which can decrease energy efficiency. We show the impact of altitude of each ABS on the considered KPIs in Fig. 5.2 which demonstrates that higher altitudes have a positive impact on all considered KPIs. While higher altitudes can lead to increased power

$$
\begin{aligned}
\underset{\mathbf{L},\mathbf{T},\mathbf{H},\mathbf{B}}{\text{minimize}} \quad & f(\varsigma,\zeta,\rho,\xi) \\
\text{subject to} \quad & L_{\min} \leq L_n \leq L_{\max}; \forall L_n \in \mathbf{L} = \{L_1, L_2, ..., L_N\}, \\
& T_{\min} \leq T_n \leq T_{\max}; \forall T_n \in \mathbf{T} = \{T_1, T_2, ..., T_N\}, \\
& H_{\min} \leq H_n \leq H_{\max}; \forall H_n \in \mathbf{H} = \{H_1, H_2, ..., H_N\}, \\
& B_{\min} \leq B_n \leq B_{\max}; \forall B_n \in \mathbf{B} = \{B_1, B_2, ..., B_N\}, \\
& 0 \leq \eta_1, \eta_2, \eta_3, \eta_4 \leq 1, \\
& 0 \leq \eta_1 + \eta_2 + \eta_3 + \eta_4 \leq 1
\end{aligned}
\tag{5.19}
$$

$$
f_{target} = \sqrt{\eta_1(\varsigma - \varsigma_{target})^2 + \eta_2(\zeta - \zeta_{target})^2 + \eta_3(\rho - \rho_{target})^2 + \eta_4(\xi - \xi_{target})^2} \tag{5.20}
$$

$$
f_{linear} = \eta_1\varsigma + \eta_2\zeta + \eta_3\rho + \eta_4\xi \tag{5.21}
$$

consumption, the effect on the energy efficiency KPI is relatively minor. This is attributed to the advantages of improved area spectral efficiency at higher altitudes, which outweigh the slight increase in power consumption. It is worth noting that the power consumption of ABS does increase with altitude, as air density varies, but this change is insignificant for altitude differences smaller than a thousand meters.

Next, we analyze the impact of ABSs beamwidth on KPIs. As the beamwidth of ABSs antenna controls the directionality and spreading of signals, it has a significant impact on coverage, latency satisfaction, average spectral efficiency, and energy efficiency. Narrower beamwidth translates to more directional beam with stronger signal strength over a smaller area, while wider beamwidths have less directional beam and spread out over a larger area. We observe a similar pattern in Fig. 5.2, where wider beamwidths improve coverage but negatively impact latency satisfaction, average spectral efficiency, and energy efficiency. Relatively moderate beamwidths are more effective in improving these KPIs. Considering the impact of ABSs altitude and beamwidth, we hypothesize that these COPs, including ABSs location and transmit power, will require intelligent control, and the contrasting trends of KPIs with respect to COPs motivate multi-objective optimization.

5.3.3 Multi-Objective Optimization Problem

We formulate a multi-objective optimization problem that considers several optimization variables, including ABSs locations, transmit power, altitude, and beamwidth while aiming to achieve specific targets for coverage, latency satisfaction, average spectral efficiency, and energy efficiency. Since these KPIs are measured on different scales, we normalize their values using the min-max normalization technique. To approximate the minimum and maximum values of each KPI, we use a pseudo brute force method that sweeps the solution space for reduced combinations of the optimization variables [61].

In the objective function of Eq. 5.19 with $f(\varsigma, \zeta, \rho, \xi) = f_{target}$ in Eq. 5.20, we aim to minimize the difference between the normalized value of each KPI and its target value, which is set by the network operator. We define the optimization variables as follows: the location of N ABSs denoted as $\mathbf{L}$, the transmit power of N ABSs denoted as $\mathbf{T}$, the altitude of N ABSs denoted as $\mathbf{H}$, and the beamwidth of N ABSs denoted as $\mathbf{B}$. We also use ς to denote the normalized values of coverage, ζ to denote the normalized values of latency satisfaction, ρ to denote the normalized values of average spectral efficiency, and ξ to denote the normalized values of energy efficiency. Moreover, we use ζ_{target} to represent the target value for latency satisfaction, ς_{target} to represent the target value for coverage, ρ_{target} to represent the target value for average spectral efficiency, and ξ_{target} to represent the target value for energy efficiency. Finally, we use η_1, η_2, η_3, and η_4 to denote the network operator-defined weights that can be used to adjust the relative importance of the KPIs.

Our objective with the optimization problem presented in Eq. 5.19 is to effectively balance a set of KPIs to approach their respective target values set by the network operator, while minimizing the inherent conflicts between them. An alternative approach to combining these KPIs is through a linear combination using operator-defined weights and normalized KPIs, as demonstrated in Eq. 5.21 where $f(\varsigma, \zeta, \rho, \xi) = f_{linear}$. In Section 5.5, we will explore the advantages of formulating optimization problems as target minimization, which allows for a non-biased convergence of multiple KPIs, as opposed to the biased impact that some KPIs may have in linear minimization.

It is important to note that the multi-objective problem presented in Eq. 5.19 is a mixed-integer nonlinear programming problem with a complexity of $\mathcal{O}((L_{max} - L_{min} + 1)^N \times (T_{max} - T_{min} + 1)^N \times (H_{max} - H_{min} + 1)^N \times (B_{max} - B_{min} + 1)^N)$. Given the complexity of this problem, conventional methods such as analytical or heuristic-based models will fail to provide real-time solutions that can accommodate varying UE mobility and other network dynamics, as they tend to make simplistic assumptions about the environment and often fail to capture the complexity of network dynamics.

5.4 Proposed AIR-DRL Optimization Framework

Compared to conventional methods, the AIR-DRL approach enables network operators to optimize COPs online by learning from the real-time responses of actual networks, reflecting the dynamic and complex nature of cellular networks. To implement the AIR-DRL optimization framework, we leverage the state-of-the-art soft actor-critic algorithm [77]. This choice is driven by its ability to efficiently explore large action spaces, which is crucial given that COPs described in Eq. 5.19 can involve tens of thousands of combinations. Furthermore, the soft actor-critic method offers high sample efficiency and avoids brittleness to hyperparameters. In the following subsections, we outline the details of our proposed AIR-DRL optimization framework.

Markov Decision Process Formulation

We formulate the ABS COPs and system KPIs as a Markov decision process (MDP) based on the environment state space, action space, and reward.

State Space: The KPIs described in Section 5.3.1 are predominantly impacted by three critical network features: received signal strength, SINR, and UE scheduling ratio. To define the system state, we combine these three features by considering the number of UEs with received signal strength above a threshold of -90 dBm, number of UEs with SINR above a threshold 2 dB, and UE scheduling ratio for each ABS and stack them for a specified number of TTIs, denoted by Z. To accomplish this, we employ the FLARE (Flow of Latents for Reinforcement Learning) approach, which leverages the difference in feature values between the current and subsequent timestamps as a state variable [82]. Thus, the system state $\mathbf{s}_e$ at epoch e can be defined as:

$$\mathbf{s}_e = \{\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, ..., \mathbf{x}_N, \mathbf{y}_N, \mathbf{z}_N\}, \tag{5.22}$$

where $\mathbf{x}_i = \{x_i^{\tau-Z}, x_i^{\tau-Z-1} - x_i^{\tau-Z}, ..., x_i^{\tau} - x_i^{\tau-1}\}$, $\mathbf{y}_i = \{y_i^{\tau-Z}, y_i^{\tau-Z-1} - y_i^{\tau-Z}, ..., y_i^{\tau} - y_i^{\tau-1}\}$, and $\mathbf{z}_i = \{z_i^{\tau-Z}, z_i^{\tau-Z-1} - z_i^{\tau-Z}, ..., x_i^{\tau} - x_i^{\tau-1}\}$ represents the stacked values of the n^{th} ABS, for the previous Z TTIs, for received signal strength, SINR, and UE scheduling ratio features, respectively, and $\tau = (e + 1)Z$. These feature values are calculated for each ABS; therefore, the size of system state space S will be independent of the number of UEs with a cardinality of $3 \times Z \times N$.

Action Space: The optimization variables in Eq. 5.19 are; position, transmit power, altitude and beamwidth of N ABSs, therefore, the action $\mathbf{a}_e$ at each epoch e is defined as:

$$\mathbf{a}_e \quad = \quad \{L_1, L_2, ..., L_N, T_1, T_2, ..., T_N, H_1, H_2, ..., H_N, B_1, B_2, ..., B_N\}. \tag{5.23}$$

The size of the system action space A is defined by the number of ABS (N), range of values for position, transmit power, altitude and beamwidth of each ABS, such that,

$|A| = (L_{max} - L_{min} + 1) \times (T_{max} - T_{min} + 1)^N \times (H_{max} - H_{min} + 1)^N \times (B_{max} - B_{min} + 1)^N.$

Reward: The reward is formulated to capture the objective function value defined in Eq. 5.19 for a specific action $\mathbf{a}_e$ taken at the system state $\mathbf{s}_e$. Due to the normalization of the KPIs and constraints on the network operator-defined weights, the objective function values fall in the range of $[0, 1]$. To convert the minimization objective into a maximization objective, the objective function value is subtracted from 1, resulting in the system reward. The reward will be scaled using ω parameter to higher values since it has been demonstrated that the soft actor-critic algorithm works better with larger reward magnitudes [77].

AIR-DRL Framework

The proposed AIR-DRL framework executes the action $\mathbf{a}_e$ at the current state $\mathbf{s}_e$, which transitions the environment to the next state $\mathbf{s}_{e+1}$ and generates a reward signal r_e characterizing the utility of the action $\mathbf{a}_e$ on the environment, based on the state-action-reward tuple $(\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1})$ This tuple is sent to the replay buffer at each epoch e, and is utilized when training the function approximators using stochastic gradient.

As shown in Fig. 5.3, the framework for ABS COPs optimization consists of state, action, reward, and next state, which, in combination, form an experience tuple that is used for training. The actor and critic networks play an important role in the optimization process. The actor network $\pi_\psi(\mathbf{s}_e)$, with parameter ψ, estimates the mean and standard deviation of the conditional Gaussian probability distribution for each continuous action $\mathbf{a}_e$ in state $\mathbf{s}_e$, using Eq. 6.17. The critic network includes two soft Q-functions $Q_{\theta_1}(\mathbf{s}_e, \mathbf{a}_e)$ and $Q_{\theta_2}(\mathbf{s}_e, \mathbf{a}_e)$, with parameters θ_1 and θ_2, which take state $\mathbf{s}_e$ and action $\mathbf{a}_e$ as input to return the corresponding expectation of value function, and two target state value functions $V_\vartheta(\mathbf{s}_e)$ and $V_{\vartheta^-}(\mathbf{s}_e)$, with parameters ϑ and ϑ^-, to improve the stability of the optimization. Each soft

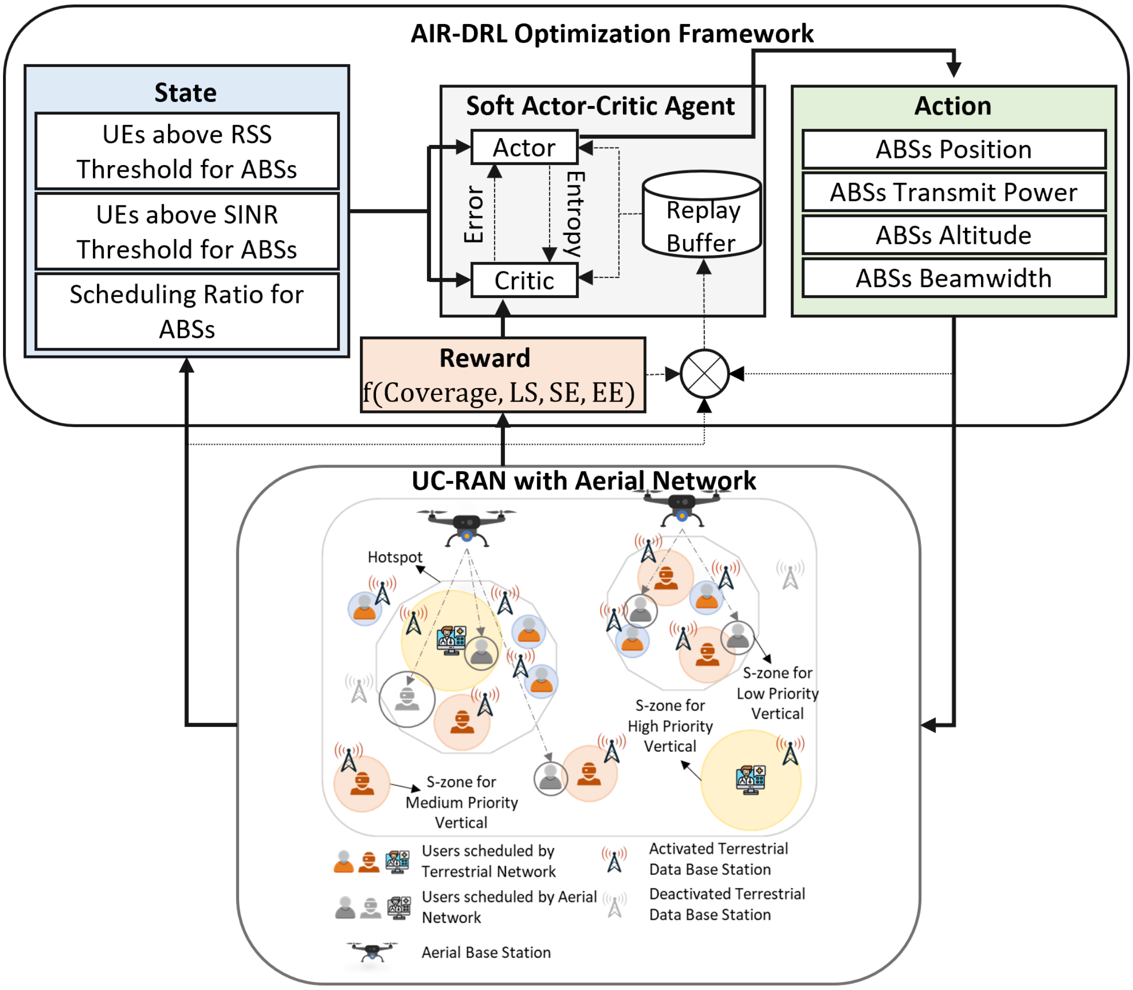

AIR-DRL Optimization Framework
State
UEs above RSS Threshold for ABSs
UEs above SINR Threshold for ABSs
Scheduling Ratio for ABSs
Soft Actor-Critic Agent
Actor
Critic
Error
Entropy
Replay Buffer
Action
ABSs Position
ABSs Transmit Power
ABSs Altitude
ABSs Beamwidth
Reward
f(Coverage, LS, SE, EE)
UC-RAN with Aerial Network
Hotspot
S-zone for Low Priority Vertical
S-zone for High Priority Vertical
S-zone for Medium Priority Vertical
Users scheduled by Terrestrial Network
Users scheduled by Aerial Network
Aerial Base Station
Activated Terrestrial Data Base Station
Deactivated Terrestrial Data Base Station

Fig. 5.3: Proposed AIR-DRL optimization framework for aerial network COPs control.

Q-function and corresponding target state value function have the same structure and parameterization, and are trained using Eq. 6.15 and Eq. 6.16, respectively.

The actor and critic network parameters are updated by randomly sampling mini-batches of experiences from the replay buffer. In the soft actor-critic method, the actor generates the mean and standard deviation of a Gaussian probability distribution for each action dimension, and an action is randomly chosen based on this distribution. These generated actions are unbounded, so the network applies the hyperbolic tangent function to bound the continuous action for each dimension within the range, $[-1, 1]$ [77].

It's worth noting that the action space for ABS COPs is a large discrete multi-

Algorithm 7: Psuedo-code for proposed AIR-DRL Optimization Framework.

```
initialize network parameters $\theta_1, \theta_2, \psi$;
$\vartheta = \theta_1$, $\vartheta^- = \theta_2$;
for each epoch do
    initialize system state $\mathbf{s}_e$;
    if epoch < warm start epochs then
        select system action $\mathbf{a}_e \in A$ randomly
    else
        select system action $\mathbf{a}_e \sim \pi_\psi(.|\mathbf{s}_e)$
    execute system action in the environemnt;
    observe system reward $r_e$ and obtain next state $\mathbf{s}_{e+1}$ feedback from
      environment;
    store experience $(\mathbf{s}_e, \mathbf{a}_e, r_e, \mathbf{s}_{e+1})$ in replay buffer $D$;
    for each gradient step do
        sample experience mini-batches from replay buffer $D$;
        update the soft Q-functions according to Eq. 6.15;
        update the state value functions according to Eq. 6.16;
        update the policy network according to Eq. 6.17;
```

dimensional space that can be mapped to a continuous action space by means of quantization. The AIR-DRL alogirthm is presented in Algorithm 7. The soft actor-critic network parameters are initialized such that each critic and its target (soft Q-function and subsequent state value function) is initialized with the same values. During the warm start period, the environment is randomly explored by selecting actions randomly. The number of warm start epochs is a hyperparameter that requires optimization depending on the type of problem at hand. After the warm start period is completed, the system actions are selected using the policy $\pi_\psi(.|\mathbf{s}_e)$. The system actions are executed in the environment, followed by observing the next state and reward as feedback from the environment.

5.5 Experimental Evaluation

The experimental evaluation section analyzes several aspects of the proposed integrated aerial network and LEAP S-Zones-based terrestrial network coupled with AIR-DRL optimization framework. First, we compare the integrated network's performance against standalone LEAP S-Zones-based terrestrial network and traditional terrestrial heterogeneous network. Next, we evaluate the performance of the proposed AIR-DRL optimization framework for each KPI against the pseudo brute force results. The pseudo brute force method approximates the maximum for each KPI and multi-objective function formulated in Eq. 5.19 by exhaustively traversing through all possible combinations of ABS COPs with a predefined step size. If the step size is not too large, the pseudo brute force method is expected to approximate the actual maximum. Then, we analyze the performance of the AIR-DRL optimization framework by combining KPIs as targeted optimization in Eq. 5.20 and linear optimization as shown in Eq. 5.21, emphasizing the importance of multi-objective problem formulation in AIR-DRL framework. Finally, we discuss the utility of using AIR-DRL optimization framework as compared to brute force, as it can optimize larger action spaces.

We use the network model depicted in Fig. 5.2 with UE/DBS densities of 300 per simulation region of 1 square kilometer with four hotspots and four ABSs. Each UE belongs to one of three verticals with varied throughput and latency requirements. The bandwidth of each terrestrial channel and aerial channel is set to 50 MHz. The LEAP S-Zone size for each vertical is set to 30 meters inspired from optimal performance results shown in [8]. The chosen values for optimizing ABS transmit power, altitude, and beamwidth are (15, 19, 23) dBm, (100, 250, 400) meters amd (35, 50, 65) degrees, respectively. The target value for each KPI in Eq. 5.20 is set to 1.

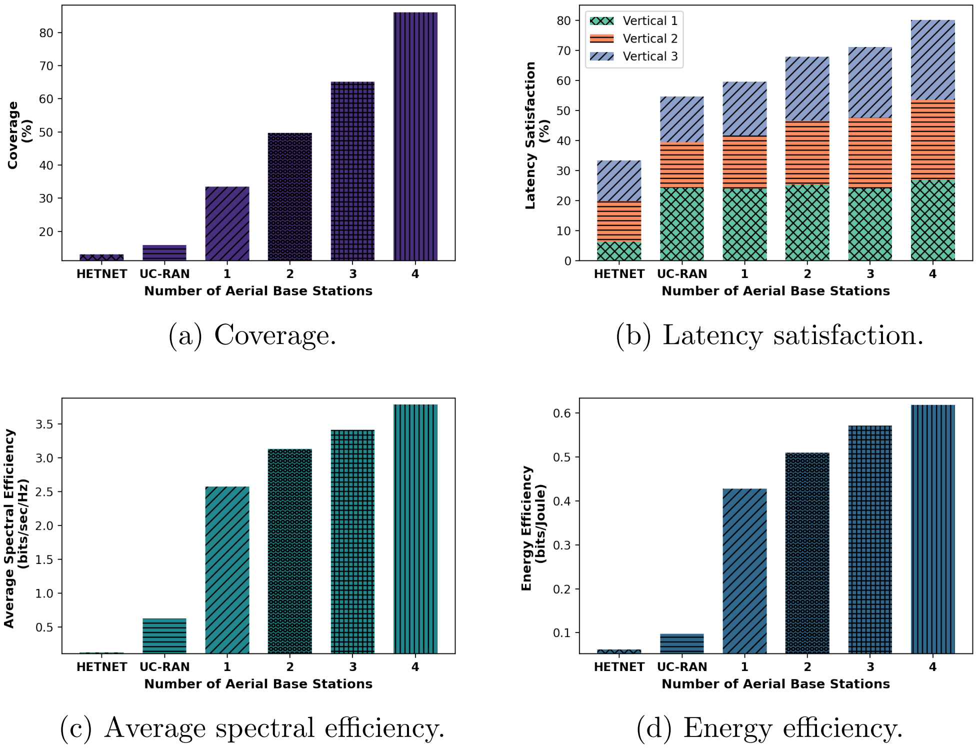
The network model of 1 square kilometer is discretized into 16 squares of 0.0625 square kilometers, where ABS will be deployed at the center of the square. We define four zones in these 16 discretized locations, with the condition that each zone will only be served by one ABS. These zones are defined to avoid any chances of tangling among the ABSs, which is an unwanted scenario. The choice of discretization is also motivated by reducing the number of possible options for the deployment of ABS, which will reduce the complexity of finding the optimal position. Note that the considered number of discretized deployment options, corresponding zones, and the range of COPs values are representative use cases without loss of generality.

The actor and critic networks in the proposed AIR-DRL framework use a neural network architecture with one hidden layer comprising 256 neurons and a rectified linear unit activation function. A learning weight of 0.003, a discount factor of 0.99, a target network update frequency of 20 epochs, a mini-batch size of 256, a replay buffer size of 10000, and a warm start period of 300 epochs are set for the networks. The frameworks are implemented in Python using PyTorch, and simulation results are averaged over multiple random seed numbers.

5.5.1 Comparison with Standalone LEAP S-Zone Terrestrial Network and Heterogeneous Network

When comparing different network architectures, it's essential to use a similar underlying architecture to ensure a fair comparison. This means that the density of DBS will be equal to the density of UE for all three variations of networks considered, and UE requirements will be set according to the vertical it belongs to. One notable difference between the three considered architectures is their association with terrestrial and aerial base stations. In heterogeneous networks, UEs are associated with DBSs providing the strongest signal. LEAP S-Zone terrestrial network, on the other hand, have service exclusion zones that limit the number of scheduled UEs and its association with DBSs. The integrated aerial network and LEAP S-Zones-based terrestrial network provides coverage to UEs via ABSs that are left unscheduled from the LEAP S-Zones-based terrestrial network.

To ensure fairness in comparison, the integrated aerial network and LEAP S-Zones-based terrestrial network will be allocated a 50 MHz bandwidth, and the standalone



(a) Coverage.

(b) Latency satisfaction.

(c) Average spectral efficiency.

(d) Energy efficiency.

Fig. 5.4: Comparison of integrated aerial network and LEAP S-Zone terrestrial network with the standalone LEAP S-Zone terrestrial network and heterogeneous network.

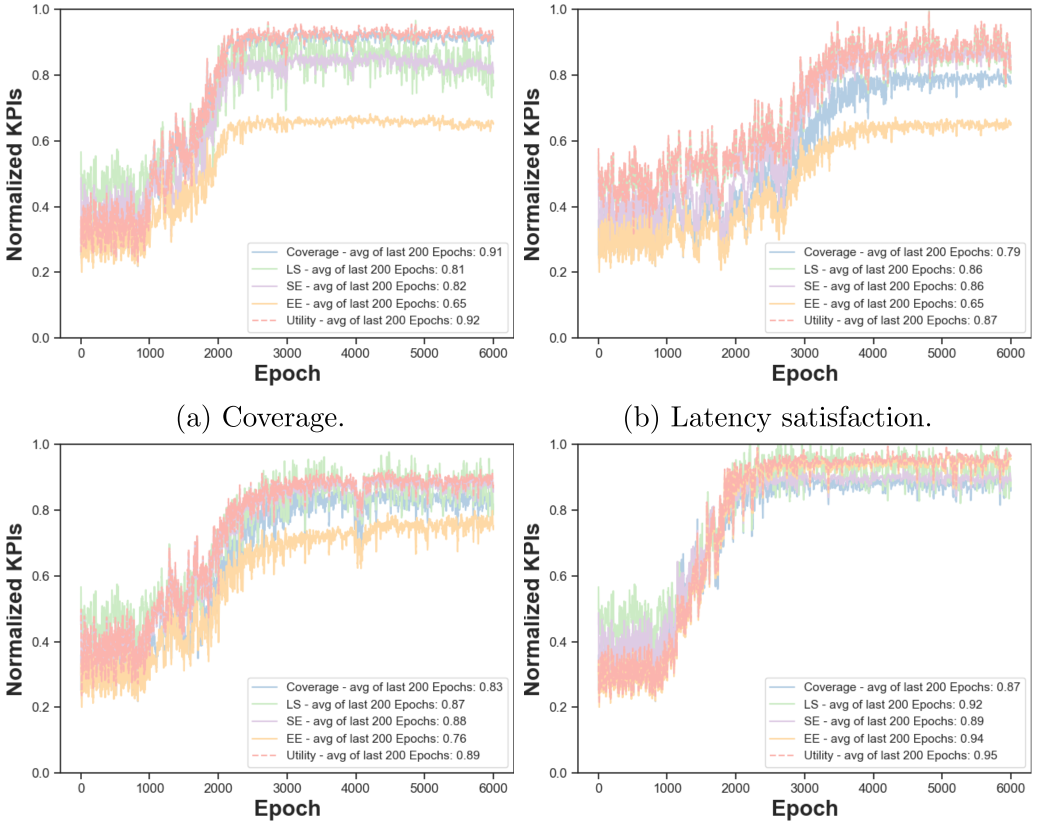
LEAP S-Zones-based terrestrial network and traditional terrestrial heterogeneous network will be allocated a 100 MHz bandwidth. In Fig. 5.4, we present a comparison of the performance of each architecture for four KPIs: coverage, latency satisfaction, average spectral efficiency, and energy efficiency. The results show that LEAP S-Zones-based terrestrial network significantly outperform heterogeneous networks in all four KPIs. This is mainly due to the increase in the density of DBSs in the network, which causes traditional heterogeneous network designs to become interference-limited.

However, despite the significant improvements achieved by LEAP S-Zones-based terrestrial network architecture, latency satisfaction remains an issue, as low-priority verticals (vertical 2 and 3) are impacted in this architecture. To address this, the integrated aerial network and LEAP S-Zones-based terrestrial network provides coverage to UEs that are not scheduled by the LEAP S-Zones-based-terrestrial network. The performance of the ABS-assisted architecture improves considerably as the number of ABSs in the network increases. It's worth noting that the results presented are for the optimal deployment of ABSs, which are considered at the center of hotspots. The transit power, altitude, and beamwidth, among the possible options discussed earlier, are optimized for this comparison. These plots demonstrate the utility of integrating an aerial network (with optimized COPs and deployment) with terrestrial user-centric architecture, which has the ability to serve UEs with optimized coverage, latency satisfaction, average spectral efficiency, and energy efficiency, making it a suitable choice for 6G networks.

5.5.2 Optimizing Individual System-level KPIs

In this analysis, we evaluate the performance of the proposed AIR-DRL optimization framework in terms of the convergence of individual KPIs (coverage, latency satisfaction, average spectral efficiency, and energy efficiency) to near-optimal solutions obtained from pseudo brute force. We run the proposed AIR-DRL framework with a reward function to optimize and show the results in terms of normalized values of each KPI and the unscaled reward function (utility). For each of the results shown in Fig. 5.5, we set the weight of 1 in Eq. 5.19 for the corresponding KPI and 0 for the rest of the KPIs.

The results in Fig. 5.5 demonstrate that the proposed AIR-DRL optimization framework can converge each KPI to near-optimal solutions in around 3000 epochs. When individual KPIs are maximized, the AIR-DRL optimization framework is indifferent to the rest of the system KPIs. This highlights the importance of multi-objective optimization problem formulation in AIR-DRL, as the convergence of AIR-DRL is highly dependent on the formulation of the problem.



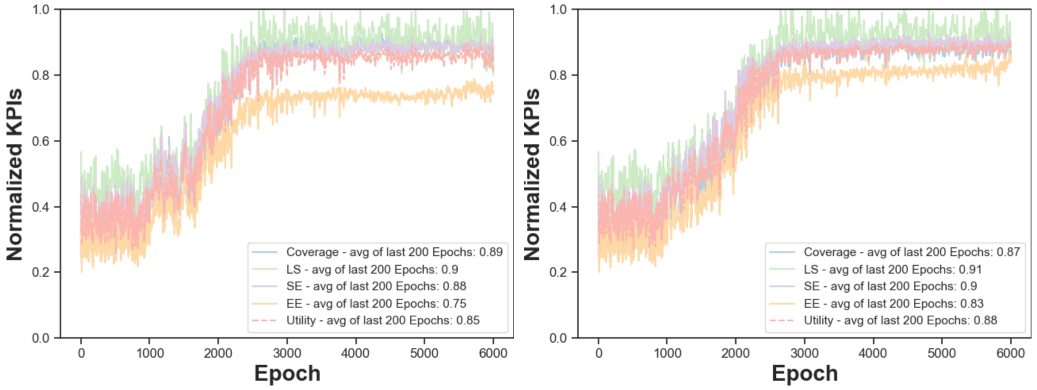
(a) Coverage.

(b) Latency satisfaction.

(c) Average spectral efficiency.

(d) Energy efficiency.

Fig. 5.5: Proposed AIR-DRL optimization framework performance in optimizing individual system-level KPIs.



(a) Linear optimization.

(b) Targeted optimization.

Fig. 5.6: Proposed AIR-DRL optimization framework performance in optimizing multiple system-level KPIs jointly.

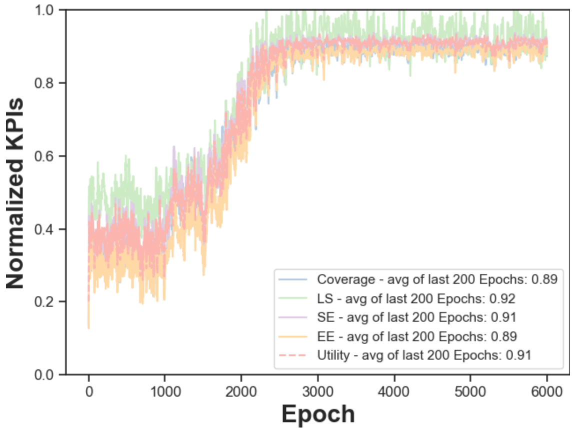


Fig. 5.7: Proposed AIR-DRL optimization framework performance in optimizing large action spaces.

5.5.3 Optimizing Multiple System-level KPIs

In practical scenarios, network operators aim to optimize multiple KPIs simultaneously because maximizing only one KPI will not satisfy the requirements of either the users or the network operators. In this analysis, we perform optimization using the proposed AIR-DRL optimization framework with reward to maximize by assigning equal weights of 0.25 to all KPIs in Eq. 5.20 and Eq. 5.21.

From the results shown in Fig. 5.6, we observe that the proposed AIR-DRL optimization framework has powerful capabilities in optimizing multiple KPIs simultaneously in a few thousand epochs. Additionally, as mentioned earlier, the formulation of the targeted optimization problem gives better capabilities of assigning near-equal priorities to all KPIs, and this unbiased optimization of all KPIs to their near-optimal values can be observed in Fig. 5.6. In contrast, the linear optimization problem is biased towards optimizing three KPIs at the expense of energy efficiency.

5.5.4 Optimizing Larger Action Spaces

In previous experimental evaluations, we demonstrated the effectiveness of the proposed AIR-DRL framework in optimizing ABS COPs with a moderate number of options for each COP. However, the limited number of COP options was due to the need to compare performance with benchmarks such as pseudo brute force, which is computationally expensive even with a smaller number of COP options. Nevertheless, the advantage of using proposed AIR-DRL optimization framework is that it can optimize much larger action spaces than pseudo brute force. To evaluate the capability of the proposed AIR-DRL optimization framework, we increase the range of ABS COPs to include choices of transmit power within (15, 17, 19, 21, 23) dBm, altitude within (100, 150, 200, 250, 300, 350, 400) meters, and beamwidth within (35, 40, 45, 50, 55, 60, 65) degrees.

To compare the performance with the approximate minimum and maximum of KPIs, we initialize with the minimum/maximum of KPIs observed for a small set of COP combinations (pseudo brute force) and update the values if the AIR-DRL observes any higher value than the maximum or lower value than the minimum during the exploration/exploitation phase. By using this method, the proposed AIR-DRL framework can perform optimization on an increased range of COPs beyond those that can be tested in pseudo brute force.

Fig. 5.7 shows the results of the proposed framework optimization with reward to optimize and weights set to equal in Eq. 5.20 in terms of the normalized values of each KPI and utility, where the normalized values are initiated with pseudo brute force and then updated as AIR-DRL explores the environment further. From the results shown in Fig. 5.7, it can be observed that the proposed AIR-DRL optimization framework can optimize multiple KPIs, even when the optimization problem space is increased.

5.6 Conclusion

In this chapter, we proposed integrating an aerial network and a LEAP S-Zones-based terrestrial network to meet emerging vertical requirements. We analyzed how ABS COPs impact system KPIs such as coverage, latency satisfaction, average spectral efficiency, and energy efficiency. We also formulated a multi-objective optimization problem that achieves the target set by the network operator for each KPI. Finally, we proposed AIR-DRL optimization framework to optimize KPIs jointly without biasing towards a few specific KPIs. Our results demonstrate that the aerial network integrated with LEAP S-Zones-based terrestrial network can provide high gains in all four considered KPIs. Furthermore, with intelligent control of ABS COPs through the proposed AIR-DRL optimization framework, the system KPIs can converge to near-optimal solutions obtained from pseudo brute force, showcasing the utility of the proposed approach.

CHAPTER 6

Conclusions and Future Work

6.1 Conclusion

The rise of UDNs is fueled by ambitious objectives of achieving high spectral efficiency, energy efficiency, and enhanced user experience while meeting strict latency and reliability requirements for heterogeneous user applications. However, the increased network densification to accommodate growing user demands comes at the expense of higher inter-cell interference and energy consumption. While traditional cellular architectures struggle to minimize these without compromising service quality or capacity, this dissertation investigates the optimal design and operation of a user-centric network in mitigating interference, reducing power consumption, ensuring diverse user/vertical service quality, facilitating proactive network operation, risk-aware optimization, adopting an open radio access network, and enabling universal coverage.

We provided an analytical and numerical analysis of the impact of enabling CoMP in a UC-RAN architecture. Additionally, we investigated the influence of new degrees of freedom, such as S-zone size and density of DBSs, on KPIs including mean serviced UEs, area spectral efficiency, and energy efficiency of the network. The numerical results based on the derived analytical model revealed an intriguing interplay between S-zone size and DBS density. It was observed that there exists an optimal size of S-zone and DBS density that maximizes area spectral efficiency and energy efficiency for any number of cooperative DBSs in an S-zone. However, the values of optimal S-zone size and DBSs density for maximizing area spectral efficiency differed significantly from those that maximize network-wide energy efficiency. Therefore, an AI-assisted self-organizing framework is needed to dynamically orchestrate these network design parameters and strike an ideal tradeoff between these KPIs for a network operator.

To this end, a demand-driven elastic user-centric architecture was developed, utilizing a data-driven model based on deep reinforcement learning to cater to diverse user application needs. We proposed D-RAN, a deep reinforcement learning-based user-centric RAN optimization framework capable of adapting to dynamic user application demands and network conditions. Unlike previous cellular network approaches, D-RAN introduced the concept of elasticity within user-centric systems, employing non-uniform virtual cells (S-zones) for different QoS categories such as Augmented/Virtual Reality and E-health applications. To avoid exhaustive search using brute-force or meta-heuristics, we developed a D-RAN framework that dynamically adjusts S-zone sizes based on changing network dynamics like user mobility. D-RAN offers a less complex approach than brute-force or meta-heuristic techniques by accurately learning the mapping of environmental conditions to S-zone sizes corresponding to different QoS categories. The proposed architecture optimizes a multi-objective problem in real-time based on KPIs such as area spectral efficiency, energy efficiency, UE service rate, and throughput satisfaction. Simulated results indicate that the D-RAN framework is nearly as effective as brute-force and outperforms meta-heuristics like simulated annealing, while maintaining lower complexity and adaptability to dynamic network changes.

Furthermore, we proposed a data-driven model empowered by digital twin technology to optimize KPIs in complex cellular networks while ensuring system safety. We introduced offline learning leveraging a digital twin instilling risk-awareness in the DRL optimization framework. The convergence and accumulated risk of the proposed DT-assisted DRL framework were compared against brute force results, demonstrating impressive performance in reaching near-optimal solutions within a few hundred iterations. Moreover, the risk-aware optimization framework showcased the viability of online learning techniques in live cellular networks with controlled and reliable exploration. On the other hand, the multifaceted requirements of 5G and beyond applications demand a system capable of supporting multiple combinations of prioritized access, reliability, throughput, and energy efficiency. We presented and evaluated a UC-RAN based on O-RAN architecture that caters to the QoS requirements of various verticals. We introduced and investigated two key UC-RAN configuration and optimization parameters: the size of user-centric virtual cells (S-zones) and the number of UEs scheduled per S-zone. These parameters are leveraged through an application (rApp) to control latency satisfaction, reliability satisfaction, area spectral efficiency, and energy efficiency. Overall, this work demonstrates a highly flexible O-RAN-based user-centric architecture coupled with a DT-empowered risk-aware DRL optimization framework that can address the fundamental tradeoff between latency, reliability, and throughput with accelerated and reliable optimization in live emerging cellular networks.

Lastly, we established an integrated aerial network, adopting a user-centric approach to enhance the capabilities of the user-centric network and provide ubiquitous services to users across various verticals. We proposed integrating an aerial network with a LEAP S-Zones-based terrestrial network to meet emerging vertical requirements. We analyzed the impact of aerial base stations (ABS) on system KPIs, including coverage, latency satisfaction, average spectral efficiency, and energy efficiency. Additionally, we formulated a multi-objective optimization problem that aims to achieve the target set by the network operator for each KPI. Finally, we introduced the AIR-DRL optimization framework to jointly optimize the KPIs without biasing towards specific ones. Our results demonstrate that the integration of the aerial network with the LEAP S-Zones-based terrestrial network can provide significant improvements in all four considered KPIs. Furthermore, with intelligent control of ABS through the proposed AIR-DRL optimization framework, the system KPIs can converge to near-optimal solutions obtained from pseudo brute force, showcasing the utility of the proposed approach.

To summarize, this dissertation investigated the optimal design and operation of a user-centric network through the analysis of UC-RAN integrated with CoMP technology, the development of a demand-driven elastic user-centric architecture empowered by deep reinforcement learning, the development of digital twin-empowered risk-aware optimization framework, and an integrated aerial network. The proposed solutions highlight the effectiveness, adaptability, and performance improvements brought by these novel approaches, paving the way for the advancement of future wireless communication systems.

6.2 Future Work

The work undertaken in this dissertation can be further enhanced on multiple fronts.

6.2.1 Error-aware Optimization Framework

One significant aspect pertains to the assumption made in the proposed data-driven optimization approaches, which assumes error-free state information obtained from the environment. However, it is crucial to acknowledge that inherent inaccuracies may exist in determining UE and DBS locations, consequently introducing errors into the states observed by the optimization framework. To overcome this limitation, future research efforts should focus on the development of an error-aware optimization framework capable of accommodating and mitigating these potential discrepancies. Such advancement will ensure the robustness and accuracy of the optimization process, thereby facilitating the practical implementation of the proposed solutions.

6.2.2 Implementation in Cellular Testbed

Another valuable addition to this dissertation would be the implementation of a cellular testbed that incorporates the proposed user-centric network coupled with a data-driven optimization framework. This implementation will not only identify and bridge any gaps in the theoretical research but also provide tangible evidence of the achievable gains offered by the proposed architecture. Furthermore, a cellular testbed implementation will establish the reliability and viability of the proposed solutions, enabling rapid adoption and adaptation by collaborating with industry partners and a wider industry audience.

Bibliography

[1] B. Romanous, N. Bitar, A. Imran, and H. Refai, “Network Densification: Challenges and Opportunities in Enabling 5G,” in *2015 IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, 2015, pp. 129–134.

[2] M. Nabeel, U. S. Hashmi, S. Ekin, H. Refai, A. Abu-Dayya, and A. Imran, “SpiderNet: Spectrally Efficient and Energy Efficient Data Aided Demand Driven Elastic Architecture for 6G,” *IEEE Network*, vol. 35, no. 5, pp. 256–263, 2021.

[3] U. S. Hashmi, S. A. R. Zaidi, and A. Imran, “User-Centric Cloud RAN: An Analytical Framework for Optimizing Area Spectral and Energy Efficiency,” *IEEE Access*, vol. 6, pp. 19 859–19 875, 2018.

[4] U. S. Hashmi, S. A. R. Zaidi, A. Imran, and A. Abu-Dayya, “Enhancing Downlink QoS and Energy Efficiency Through a User-Centric Stienen Cell Architecture for mmWave Networks,” *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 2, pp. 387–403, 2020.

[5] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, “A Survey of Self Organisation in Future Cellular Networks,” *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 336–361, 2013.

[6] O-RAN Alliance, “Operator Defined Open and Intelligent Radio Access Networks.” [Online]. Available: https://www.o-ran.org/

[7] S. K. Kasi, U. S. Hashmi, M. Nabeel, S. Ekin, and A. Imran, “Analysis of Area Spectral and Energy Efficiency in a CoMP-Enabled User-Centric Cloud RAN,” *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 4, pp. 1999–2015, 2021.

[8] S. K. Kasi, U. S. Hashmi, S. Ekin, A. Abu-Dayya, and A. Imran, “D-RAN: A DRL-based Demand-Driven Elastic User-Centric RAN Optimization for 6G & Beyond,” *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2022.

[9] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, “Cloud RAN for Mobile Networks - A Technology Overview,”

IEEE Communications surveys & tutorials, vol. 17, no. 1, pp. 405–426, 2014.

[10] K. Humadi, I. Trigui, W.-P. Zhu, and W. Ajib, "Dynamic Base Station Clustering in User-Centric mmWave Networks: Performance Analysis and Optimization," *IEEE Transactions on Communications*, 2021.

[11] N. Guo, M.-L. Jin, and N. Deng, "Coverage Analysis for Heterogeneous Network with User-centric Cooperation," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2724–2727, 2018.

[12] S. K. Kasi, U. Sajid Hashmi, M. Nabeel, S. Ekin, and A. Imran, "Is CoMP Beneficial In User-Centered Wireless Networks?" in *2022 1st International Conference on 6G Networking (6GNet)*, 2022, pp. 1–5.

[13] J. Shi, X. Chen, N. Huang, H. Jiang, Z. Yang, and M. Chen, "Power-efficient transmission for user-centric networks with limited fronthaul capacity and computation resource," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5649–5660, 2020.

[14] S. Zaidi, O. B. Smida, S. Affes, U. Vilaipornsawai, L. Zhang, and P. Zhu, "User-centric Base-station Wireless Access Virtualization for Future 5G Networks," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 5190–5202, 2019.

[15] J. Shi, C. Pan, W. Zhang, and M. Chen, "Performance Analysis for User-Centric Dense Networks With mmWave," *IEEE Access*, vol. 7, pp. 14 537–14 548, 2019.

[16] M. Nabeel, U. S. Hashmi, S. Ekin, H. Refai, A. Abu-Dayya, and A. Imran, "SpiderNet: Spectrally Efficient and Energy Efficient Data Aided Demand Driven Elastic Architecture for 6G," *IEEE Network*, vol. 35, no. 5, pp. 256–263, 2021.

[17] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *CoRR*, vol. abs/1606.06565, 2016. [Online]. Available: http://arxiv.org/abs/1606.06565

[18] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, "Trial without Error: Towards Safe Reinforcement Learning via Human Intervention," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 2067–2069.

[19] P. Abbeel, A. Coates, and A. Y. Ng, "Autonomous Helicopter Aerobatics through Apprenticeship Learning," *The International Journal of Robotics Research*, vol. 29, no. 13, pp. 1608–1639, 2010.

[20] J. Tang, A. Singh, N. Goehausen, and P. Abbeel, "Parameterized Maneuver Learning for Autonomous Helicopter Flight," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 1142–1148.

[21] C. Gaskett, "Reinforcement Learning under Circumstances Beyond its Control," 2003.

[22] T. M. Moldovan and P. Abbeel, "Safe Exploration in Markov Decision Processes," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, 2012, p. 1451–1458.

[23] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained Policy Optimization," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 22–31.

[24] A. Wachi, Y. Sui, Y. Yue, and M. Ono, "Safe Exploration and Optimization of Constrained MDPs Using Gaussian Processes," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[25] M. Samir, D. Ebrahimi, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Leveraging UAVs for Coverage in Cell-Free Vehicular Networks: A Deep Reinforcement Learning Approach," *IEEE Transactions on Mobile Computing*, vol. 20, no. 9, pp. 2835–2847, 2021.

[26] W. Huang, J. Peng, and H. Zhang, "User-Centric Intelligent UAV Swarm Networks: Performance Analysis and Design Insight," *IEEE Access*, vol. 7, pp. 181 469–181 478, 2019.

[27] Y. Guo, S. Yin, and J. Hao, "Joint Placement and Resources Optimization for Multi-User UAV-Relaying Systems With Underlaid Cellular Networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12 374–12 377, 2020.

[28] S. Zhang and N. Ansari, "Latency Aware 3D Placement and User Association in Drone-Assisted Heterogeneous Networks With FSO-Based Backhaul," *IEEE*

Transactions on Vehicular Technology, vol. 70, no. 11, pp. 11 991–12 000, 2021.

[29] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, “Coordinated Multi-Point Clustering Schemes: A Survey,” *IEEE Communications Surveys and Tutorials*, vol. 19, no. 2, pp. 743–764, 2017.

[30] G. T. 36.819, “Coordinated Multi-Point Operation for LTE Physical Layer Aspects,” 2013.

[31] S. Chen, T. Zhao, H. H. Chen, Z. Lu, and W. Meng, “Performance Analysis of Downlink Coordinated Multipoint Joint Transmission in Ultra-dense Networks,” *IEEE Network*, vol. 31, no. 5, pp. 106–114, 2017.

[32] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu, “Future of Ultra-dense Networks Beyond 5G: Harnessing Heterogeneous Moving Cells,” *IEEE Communications Magazine*, vol. 57, no. 6, pp. 86–92, 2019.

[33] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and its Applications.* John Wiley & Sons, 2013.

[34] A. Imran, M. A. Imran, A. Abu-Dayya, and R. Tafazolli, “Self Organization of Tilts in Relay Enhanced Networks: A Distributed Solution,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 764–779, 2014.

[35] M. Haenggi, “Mean Interference in Hard-core Wireless Networks,” *IEEE Communications Letters*, vol. 15, no. 8, pp. 792–794, 2011.

[36] Q. Cui, X. Yu, Y. Wang, and M. Haenggi, “The SIR Meta Distribution in Poisson Cellular Networks with Base Station Cooperation,” *IEEE Transactions on Communications*, vol. 66, no. 3, pp. 1234–1249, 2018.

[37] W. Sun and J. Liu, “2-to-M Coordinated Multipoint-Based Uplink Transmission in Ultra-Dense Cellular Networks,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8342–8356, 2018.

[38] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, “User-centric Ultra-dense Networks for 5G: Challenges, Methodologies, and Directions,” *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78–85, 2016.

[39] A. AlAmmouri, J. G. Andrews, and F. Baccelli, "SINR and Throughput of Dense Cellular Networks with Stretched Exponential Path Loss," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1147–1160, 2017.

[40] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities," *IEEE Access*, vol. 5, pp. 4821–4845, 2017.

[41] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient Power Control in Cell-free and User-centric Massive MIMO at Millimeter Wave," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 651–663, 2019.

[42] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume *et al.*, "How Much Energy is Needed to Run a Wireless Network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, 2011.

[43] R. Gupta, E. C. Strinati, and D. Kténas, "Energy efficient joint DTX and MIMO in cloud radio access networks," in *2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET)*. IEEE, 2012, pp. 191–196.

[44] A. J. Fehske, P. Marsch, and G. P. Fettweis, "Bit Per Joule Efficiency of Cooperating Base Stations in Cellular Networks," *2010 IEEE Globecom Workshops, GC'10*, pp. 1406–1411, 2010.

[45] H. S. Lichte, S. Valentin, and H. Karl, "Expected Interference in Wireless Networks with Geometric Path Loss: A Closed-form Approximation," *IEEE communications letters*, vol. 14, no. 2, pp. 130–132, 2010.

[46] H. P. Keeler, B. Błaszczyszyn, and M. K. Karray, "SINR-based k-coverage Probability in Cellular Networks with Arbitrary Shadowing," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 1167–1171.

[47] Y. Yang, K. W. Sung, J. Park, S.-L. Kim, and K. S. Kim, "Cooperative Transmissions in Ultra-dense Networks Under a Bounded Dual-slope Path Loss Model," in *2017 European Conference on Networks and Communications (EuCNC)*. IEEE, 2017, pp. 1–6.

[48] P. Korrai and D. Sen, "Downlink SINR Coverage and Rate Analysis with Dual

Slope Pathloss Model in mmWave Networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[49] C. Galiotto, N. K. Pratas, N. Marchetti, and L. Doyle, "A Stochastic Geometry Framework for LOS/NLOS Propagation in Dense Small Cell Networks," in *2015 IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 2851–2856.

[50] M. Nabeel, V. K. Singh, and F. Dressler, "Efficient Data Gathering for Decentralized Diversity Combining in Heterogeneous Sensor Networks," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–6.

[51] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.

[52] Z. Cheng, D. Zhu, Y. Zhao, and C. Sun, "Flexible Virtual Cell Design for Ultradense Networks: A Machine Learning Approach," *IEEE Access*, vol. 9, pp. 91 575–91 583, 2021.

[53] U. S. Hashmi, S. A. R. Zaidi, and A. Imran, "User-centric cloud RAN: An analytical framework for optimizing area spectral and energy efficiency," *IEEE Access*, vol. 6, pp. 19 859–19 875, 2018.

[54] Y. Zhang, B. Di, H. Zhang, J. Lin, C. Xu, D. Zhang, Y. Li, and L. Song, "Beyond Cell-Free MIMO: Energy Efficient Reconfigurable Intelligent Surface Aided Cell-Free MIMO Communications," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 2, pp. 412–426, 2021.

[55] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to Empower SON with Big Data for Enabling 5G," *IEEE network*, vol. 28, no. 6, pp. 27–33, 2014.

[56] S.-F. Cheng, L.-C. Wang, C.-H. Hwang, J.-Y. Chen, and L.-Y. Cheng, "On-Device Cognitive Spectrum Allocation for Coexisting URLLC and eMBB Users in 5G Systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 171–183, 2021.

[57] K. Khawam, S. Lahoud, M. E. Helou, S. Martin, and F. Gang, "Coordinated Framework for Spectrum Allocation and User Association in 5G HetNets With

mmWave," *IEEE Transactions on Mobile Computing*, vol. 21, no. 4, pp. 1226–1243, 2022.

[58] L. Sboui, Z. Rezki, A. Sultan, and M.-S. Alouini, "A New Relation Between Energy Efficiency and Spectral Efficiency in Wireless Communications Systems," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 168–174, 2019.

[59] H. Khaled, I. Ahmad, D. Habibi, and Q. V. Phung, "A Green Traffic Steering Solution for Next Generation Communication Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 222–238, 2021.

[60] A. R. Ramos, B. C. Silva, M. S. Lourenço, E. B. Teixeira, and F. J. Velez, "Mapping between Average SINR and Supported Throughput in 5G New Radio Small Cell Networks," in *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2019, pp. 1–6.

[61] P. Mannion, S. Devlin, J. Duggan, and E. Howley, "Reward Shaping for Knowledge-based Multi-objective Multi-agent Reinforcement Learning," *The Knowledge Engineering Review*, vol. 33, 2018.

[62] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* MIT press, 2018.

[63] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized Experience Replay," *arXiv preprint arXiv:1511.05952*, 2015.

[64] S. K. Kasi, M. K. Kasi, K. Ali, M. Raza, H. Afzal, A. Lasebae, B. Naeem, S. ul Islam, and J. J. Rodrigues, "Heuristic Edge Server Placement in Industrial Internet of Things and Cellular Networks," *IEEE Internet of Things Journal*, 2020.

[65] K. Amine, "Multiobjective Simulated Annealing: Principles and Algorithm Variants," *Advances in Operations Research*, vol. 2019, 2019.

[66] A. Tavakoli, F. Pardo, and P. Kormushev, "Action Branching Architectures for Deep Reinforcement Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[67] F. Wei, G. Feng, Y. Sun, Y. Wang, S. Qin, and Y.-C. Liang, "Network Slice

Reconfiguration by Exploiting Deep Reinforcement Learning with Large Action Space," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2197–2211, 2020.

[68] X. Tao and A. S. Hafid, "Deepsensing: A Novel Mobile Crowdsensing Framework with Double Deep Q-network and Prioritized Experience Replay," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 547–11 558, 2020.

[69] S. K. Kasi, S. Das, and S. Biswas, "TCP Congestion Control with Multiagent Reinforcement and Transfer Learning," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 1507–1513.

[70] T. Zahavy, M. Haroush, N. Merlis, D. J. Mankowitz, and S. Mannor, "Learn What Not To Learn: Action Elimination with Deep Reinforcement Learning," *arXiv preprint arXiv:1809.02121*, 2018.

[71] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. Van Hasselt, and D. Silver, "Distributed Prioritized Experience Replay," *arXiv preprint arXiv:1803.00933*, 2018.

[72] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G Evolution: A View on 5G Cellular Technology Beyond 3GPP Release 15," *IEEE access*, vol. 7, pp. 127 639–127 651, 2019.

[73] J. Wang, Y. Liu, and B. Li, "Reinforcement Learning with Perturbed Rewards," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6202–6209.

[74] S. Jiang and A. Alkhateeb, "Digital Twin Based Beam Prediction: Can we Train in the Digital World and Deploy in Reality?" 2023.

[75] L. Bariah and M. Debbah, "The Interplay of AI and Digital Twin: Bridging the Gap between Data-Driven and Model-Driven Approaches," 2022. [Online]. Available: https://arxiv.org/abs/2209.12423

[76] C.-L. I, S. Han, Z. Xu, S. Wang, Q. Sun, and Y. Chen, "New Paradigm of 5G Wireless Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 474–482, 2016.

[77] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft Actor-Critic Algorithms and Applications," 2018. [Online]. Available: https://arxiv.org/abs/1812.05905

[78] 3GPP TR 38.901, "Study on Channel Model for Frequencies from 0.5 to 100 GHz." [Online]. Available: http://www.3gpp.org/DynaReport/38901.htm

[79] 3GPP TS 22.261, "Service Requirements for the 5G system." [Online]. Available: http://www.3gpp.org/DynaReport/22261.htm

[80] Forsk, "Forsk Atoll." [Online]. Available: https://www.forsk.com/atoll-overview/

[81] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[82] W. Shang, X. Wang, A. Srinivas, A. Rajeswaran, Y. Gao, P. Abbeel, and M. Laskin, "Reinforcement Learning with Latent Flow," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 22 171–22 183.

[83] M. Mozaffari, X. Lin, and S. Hayes, "Toward 6G with Connected Sky: UAVs and Beyond," *IEEE Communications Magazine*, vol. 59, no. 12, pp. 74–80, 2021.

[84] H. N. Qureshi and A. Imran, "On the Tradeoffs Between Coverage Radius, Altitude, and Beamwidth for Practical UAV Deployments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 2805–2821, 2019.

[85] J. Guo, P. Walk, and H. Jafarkhani, "Optimal Deployments of UAVs With Directional Antennas for a Power-Efficient Coverage," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5159–5174, 2020.

[86] C. Diaz-Vilor, A. Lozano, and H. Jafarkhani, "On the Deployment Problem in Cell-Free UAV Networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[87] J. K. Stolaroff, C. Samaras, E. R. O'Neill, A. Lubers, A. S. Mitchell, and D. Ceperley, "Energy Use and Life Cycle Greenhouse Gas Emissions of Drones for Commercial Package Delivery," *Nature communications*, vol. 9, no. 1, p. 409, 2018.

Funding Acknowledgments

The projects undertaken in this dissertation are based upon work supported by the National Science Foundation under Grant Numbers 1923669 and 1730650. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. For more inquiries about these works, please visit www.AI4networks.com.

Appendix

Appendix A: Proof of Theorem 1: The activation probability is computed based on the criterion that no more than M DBSs are activated in an S-zone iff: (i) there is a scheduled UE within a distance R_{szone} to at least 1 DBS, and (ii) no other DBS in an S-zone can provide superior channel gains to a scheduled user. Since both the constraints are independent events, we can compute p_{ACT} as:

$$p_{ACT} = Pr\Big(\Pi^{'}_{UE} \cap b(\mathbf{d}, R_{szone}) \neq \phi \Big| \mathbf{d} \in \Pi^{'}_{DBS}\Big) \boldsymbol{.} \\ Pr\Big(max\ \{h_{d_1} r_{d_1}^{-\alpha}, ..., h_{d_M} r_{d_M}^{-\alpha}\}, 1 \leq m \leq \infty \Big| \mathbf{d} \in \Pi^{'}_{DBS}, M < X\Big). \quad (6.1)$$

Solving in parts for each constraint separately, such that, $p_{ACT} = p^{'}_{ACT} \boldsymbol{.} p^{''}_{ACT}$, where

$$p^{'}_{ACT} = Pr\Big(\Pi^{'}_{UE} \cap b(\mathbf{d}, R_{szone}) \neq \phi \Big| \mathbf{d} \in \Pi^{'}_{DBS}\Big) = 1 - \exp(-\lambda^{'}_{UE} \pi R^2_{szone}), \quad (6.2)$$

and

$$p^{''}_{ACT} = Pr\Big(max\ \{h_{d_1} r_{d_1}^{-\alpha}, ..., h_{d_M} r_{d_M}^{-\alpha}\}, 1 \leq m \leq \infty \Big| \mathbf{d} \in \Pi^{'}_{DBS}, M < X\Big), \quad (6.3)$$

where m is the actual number of DBSs in an S-zone which are distributed through the Poisson point process. The joint probability given above can be divided into two parts, that is, the probability that the actual number of DBSs in an S-zone is less than/equal to M or greater than M. Also, we assume that M is chosen such that $M < \lambda_{DBS} \pi R^2_{szone}$ where M can only be an integer number. Therefore,

$$p^{''}_{ACT} = Pr\Big(max\ \{h_{d_1} r_{d_1}^{-\alpha}, ..., h_{d_M} r_{d_M}^{-\alpha}\}, 1 \leq m \leq M \Big| \mathbf{d} \in \Pi^{'}_{DBS}, M < X\Big) + \\ Pr\Big(max\ \{h_{d_1} r_{d_1}^{-\alpha}, ..., h_{d_M} r_{d_M}^{-\alpha}\}, M+1 \leq m \leq \infty \Big| \mathbf{d} \in \Pi^{'}_{DBS}, M < X\Big). \quad (6.4)$$

Employing the conditional probability formula, we obtain:

$$
\begin{aligned}
p^{''}_{ACT} = Pr(1 \leq m \leq M)& \\
Pr\Big(max\ \{h_{d_1}r_{d_1}^{-\alpha}, ..., h_{d_M}r_{d_M}^{-\alpha}\}\Big|\mathbf{d} \in \Pi^{'}_{DBS}, M < X, 1 \leq m \leq M\Big) +& \\
Pr(M+1 \leq m \leq \infty)& \\
Pr\Big(max\ \{h_{d_1}r_{d_1}^{-\alpha}, ..., h_{d_M}r_{d_M}^{-\alpha}\}\Big|\mathbf{d} \in \Pi^{'}_{DBS}, M < X, M+1 \leq m \leq \infty\Big).& \quad (6.5)
\end{aligned}
$$

If there are less than or equal to M DBSs in an S-zone then the probability of selection will be exactly 1 and if there are more than M DBS in an S-zone then only M out of m DBSs with the strongest channel gains will be selected. Therefore, the above expression can be rewritten as:

$$
p^{''}_{ACT} = \sum_{m=1}^{M} \exp(-X)\frac{(X)^m}{m!}\Big(1\Big) + \sum_{m=M+1}^{\infty} \exp(-X)\frac{(X)^m}{m!}\Big(\frac{M}{m}\Big). \quad (6.6)
$$

After numerical simplification of the above expressions, we get:

$$
p^{''}_{ACT} = \frac{\Gamma(M+1,X)}{\gamma(M+1)} + \exp(-X)\left[\frac{M(X)^{M+1}{}_2F_2(1,M+1;M+2,M+2;X)}{(M+1)\gamma(M+2)} - 1\right]. \quad (6.7)
$$

Therefore, the modified density of activated DBSs can be approximated as:

$$
\begin{aligned}
p^{''}_{ACT} = \Big(1-\exp(-\lambda^{'}_{UE}\pi R^2_{szone})\Big).\Bigg(&\frac{\Gamma(M+1,X)}{\gamma(M+1)} \\
&+ \exp(-X)\left[\frac{M(X)^{M+1}\ {}_2F_2(1,M+1;M+2,M+2;X)}{(M+1)\gamma(M+2)} - 1\right]\Bigg). \quad (6.8)
\end{aligned}
$$

Appendix B: Proof of Theorem 2: The typical user is served successfully by a DBS only if the received SIR is greater than the desired threshold. Using the concepts of thinned marked Poisson processes [33] we derive the relationship between the void probability of modified active DBSs ($\Pi^{'}_{DBS}$) and $Pr(S < \gamma_{th} I)$.

$$P_r\Big(S < \gamma_{th} I\Big) = P_r(\Pi^{'}_{DBS} = \emptyset) = \exp(-\Lambda(B)), \tag{6.9}$$

where the average measure $\Lambda(B)$ can be evaluated by:

$$\Lambda(B) = \int_0^{\infty} \int_B \lambda(r,h) dh dr, \tag{6.10}$$

where B is the area of ball region and $\lambda(r,h)$ is the intensity of the modified process which is given as [33]:

$$\lambda(r,h) = 2\pi\lambda_{DBS} r \mathbb{1}(hr^{-\alpha} \geq \gamma_{th} I) f_H(h). \tag{6.11}$$

Therefore,

$$\Lambda(B) = \int_0^{\infty} \int_B 2\pi\lambda_{DBS} r \mathbb{1}(hr^{-\alpha} \geq \gamma_{th} I) f_H(h) dh dr \overset{\text{(a)}}{=} 2\pi\lambda_{DBS} \int_0^{R_{szone}} r P_r(h \geq \gamma_{th} I r^{\alpha}) dr \overset{\text{(b)}}{=} \frac{\pi\lambda_{DBS}\delta\gamma(\delta, \gamma_{th} I R^{\alpha}_{szone})}{\gamma^{\delta}_{th} I^{\delta}}, \tag{6.12}$$

where (a) is due to the cumulative distribution function of an exponentially distributed random function, and (b) is obtained by defining the integration variable $t = \gamma_{th} I r^{\alpha}$ and integrating over t.

Substituting the value of Eq. (29) and Eq. (32) in Eq. (11), we obtain:

$$P_{cov}(\gamma_{th}, R_{szone}) = 1 - \mathbb{E}_{\mathbf{I}}\Big[exp\Big(- \frac{\pi\lambda_{DBS}\delta\gamma(\delta, \gamma_{th} I R^{\alpha}_{szone})}{\gamma^{\delta}_{th} I^{\delta}}\Big)\Big]. \tag{6.13}$$

Applying Jensen's inequality will give the lower bound for coverage probability as follows:

$$P_{cov}(\gamma_{th}, R_{szone}) \geq 1 - exp\Big(- \frac{\lambda_{DBS}\pi\delta\gamma(\delta, \gamma_{th}\mathbb{E}_{\mathbf{I}}[I] R^{\alpha}_{szone})}{\gamma^{\delta}_{th}\mathbb{E}_{\mathbf{I}}[I]^{\delta}}\Big). \tag{6.14}$$

Employing the value of $\mathbb{E}_{\mathbf{I}}[I]$ from Eq. (7) in the above expression concludes the proof.

Appendix C: Soft Actor-Critic Soft actor-critic maximizes entropy for stability and exploration with an actor-critic architecture consisting of separate policy and value function networks. Soft actor-critic considers a Markov decision process defined by the tuple (S, A, p, r), where the state space S and action space A are continuous, $p : S \times A \times S$ represent the unknown transition probability at epoch e of next state $s_{e+1} \in S$ given the current state $s_e \in S$ and $a_e \in A$, and $r : S \times A \rightarrow [r_{min}, r_{max}]$ is the bounded reward for each state-action transition. General reinforcement learning algorithms aim to maximize the expected sum of rewards $\sum_e \mathbb{E}_{(s_e,a_e)\sim\varrho_\pi}[r(s_e, a_e)]$, where ϱ_π denotes the state-action marginals of the trajectory distribution induced by action policy $\pi(a_e|s_e)$. Whereas soft actor-critic augments this objective with entropy $H(\pi(.|s_e))$ of the action policy $\sum_e \mathbb{E}_{(s_e,a_e)\sim\varrho_\pi}[r(s_e, a_e) + \alpha H(\pi(.|s_e))]$ to maximize both expected reward and policy entropy, that is, to optimize while acting as randomly as possible. The temperature parameter α controls the stochasticity of the optimal policy by determining the relative importance of the entropy term against the reward. With entropy term in the objective function, the policy is incentivized to explore more widely and capture multiple modes of near-optimal behavior.

At each epoch, soft actor-critic evaluates the policy using soft Q-function, which can be defined as $Q(s_e, a_e) = r(s_e, a_e) + \eta\mathbb{E}_{(s_e,a_e)\sim\varrho_\pi}[V(s_{e+1})]$, where $V(s_e) = \mathbb{E}_{a_e\sim\pi}[Q(s_e, a_e)] + \alpha\mathbb{E}_{a_e\sim\pi}[-\log\pi(a_e|s_e)]$ is the state value function, η is the discount factor to ensure that the sum of expected rewards and entropies remains finite, and the term $\mathbb{E}_{a_e\sim\pi}[-\log\pi(a_e|s_e)]$ signifies the expected entropy of action policy. With continuous domains, soft Q-function, state value function, and policy are estimated using neural networks. The stochastic gradient descent method is used to update parameter θ for soft Q-function $Q_\theta(s_e, a_e)$, ϑ for state value function

$V_\vartheta(s_e)$, and ψ for policy $\pi_\psi(a_e|s_e)$. The soft Q-function is trained to minimize the soft Bellman residual error $J_Q(\theta)$ between the prediction of soft Q-function and reward plus the discounted expected state-value function of next epoch:

$$J_Q(\theta) = \mathbb{E}_{(s_e,a_e)\sim D}\left[\frac{1}{2}\bigg(Q_\theta(s_e,a_e) - \hat{Q}(s_e,a_e)\bigg)^2\right], \tag{6.15}$$

where $\hat{Q}(s_e,a_e) = r(s_e,a_e) + \eta \mathbb{E}_{s_e+1\sim\varrho_\pi}[V_{\vartheta^-}(s_{e+1})]$ and D is the replay buffer which contains the distribution of previously sampled states and actions. The soft Q-function update makes use of target value network, $V_{\vartheta^-}(s_e)$, which has been shown to stabilize training [77]. To speed up the training process, the soft actor-critic model uses two soft Q-functions, which are trained independently with parameters θ_1 and θ_2. The minimum of these two parameterized soft Q-functions is then used for state value function training.

The state value function is trained to minimize the squared residual error $J_V(\vartheta)$ between the state value function and the expected prediction of entropy regularized soft Q-function:

$$J_V(\vartheta) \;=\; \mathbb{E}_{s_e\sim D}\left[\frac{1}{2}\bigg(V_\vartheta(s_e) \;-\; \mathbb{E}_{a_e\sim\pi_\psi}\big[Q_\theta(s_e,a_e) \;-\; \alpha\log\pi_\psi(a_e|s_e)\big]\bigg)^2\right]. \tag{6.16}$$

Finally, the policy function is trained to minimize the expected Kullback-Leibler divergence $J_\pi(\psi)$ using a reparameterization trick [77]:

$$J_\pi(\psi) \;=\; \mathbb{E}_{s_e\sim D,\epsilon_e\sim\mathcal{N}}\left[\log\pi_\psi\bigg(f_{psi}\big(\epsilon_e;s_e\big)\bigg|s_e\bigg) \;-\; Q_\theta\bigg(s_e,f_\psi\big(\epsilon_e;s_e\big)\bigg)\right], \tag{6.17}$$

where a_e is evaluated at $f_\psi\big(\epsilon_e;s_e\big)$ to make the action sampling a differentiable process and ϵ_e is a noise vector sampled from a Gaussian distribution at epoch. The policy network outputs the Gaussian mean and standard deviation for each action dimension included in the action space. In summary, the soft actor-critic method collects experience from the environment and updates the neural networks using the stochastic gradient method from batches sampled from a replay buffer.