Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

5G NEW RADIO ACCESS AND CORE NETWORK SLICING FOR NEXT-GENERATION NETWORK SERVICES AND MANAGEMENT

by
Abdullah Ridwan Hossain

In recent years, fifth-generation New Radio (5G NR) has attracted much attention owing to its potential in enhancing mobile access networks and enabling better support for heterogeneous services and applications. Network slicing has garnered substantial focus as it promises to offer a higher degree of isolation between subscribers with diverse quality-of-service requirements. Integrating 5G NR technologies, specifically the mmWave waveform and numerology schemes, with network slicing can unlock unparalleled performance so crucial to meeting the demands of high throughput and sub-millisecond latency constraints.

While conceding that optimizing next-generation access network performance is extremely important, it needs to be acknowledged that doing so for the core network is equally as significant. This is majorly due to the numerous core network functions that execute control tasks to establish end-to-end user sessions and route access network traffic. Consequently, the core network has a significant impact on the quality-of-experience of the radio access network customers. Currently, the core network lacks true end-to-end slicing isolation and reliability, and thus there is a dire need to examine more stringent configurations that offer the required levels of slicing isolation for the envisioned networking landscape.

Considering the factors mentioned above, a sequential approach is adopted starting with the radio access network and progressing to the core network. First, to maximize the downlink average spectral efficiency of an enhanced mobile broadband slice in a time division duplex radio access network while meeting the quality-of-service requirements, an optimization problem is formulated to determine the

duplex ratio, numerology scheme, power, and bandwidth allocation. Subsequently, to minimize the uplink transmission power of an ultra-reliable low latency communications slice while satisfying the quality-of-service constraints, a second optimization problem is formulated to determine the above-mentioned parameters and allocations. Because 5G NR supports dual-band transmissions, it also facilitates the usage of different numerology schemes and duplex ratios across bands simultaneously. Both problems, being mixed-integer non-linear programming problems, are relaxed into their respective convex equivalents and subsequently solved.

Next, shifting attention to aerial networks, a priority-based 5G NR unmanned aerial vehicle network (UAV) is considered where the enhanced mobile broadband and ultra-reliable low latency communications services are considered as best-effort and high-priority slices, correspondingly. Following the application of a band access policy, an optimization problem is formulated. The goal is to minimize the downlink quality-of-service gap for the best-effort service, while still meeting the quality-of-service constraints of the high-priority service. This involves the allocation of transmission power and assignment of resource blocks. Given that this problem is a mixed-integer nonlinear programming problem, a low-complexity algorithm, PREDICT, i.e., PRiority BasED Resource AllocatIon in Adaptive SliCed NeTwork, which considers the channel quality on each individual resource block over both bands, is designed to solve the problem with a more accurate accounting for high-frequency channel conditions.

Transitioning to minimizing the operational latency of the core network, an integer linear programming problem is formulated to instantiate network function instances, assign them to core network servers, assign slices and users to network function instances, and allocate computational resources while maintaining virtual network function isolation and physical separation of the core network control and user planes. The actor-critic method is employed to solve this problem for three

proposed core network operation configurations, each offering an added degree of reliability and isolation over the default configuration that is currently standardized by the 3GPP.

Looking ahead to potential future research directions, optimizing carrier aggregation-based resource allocation across triple-band sliced access networks emerges as a promising avenue. Additionally, the integration of coordinated multi-point techniques with carrier aggregation in multi-UAV NR aerial networks is especially challenging. The introduction of added carrier frequencies and channel bandwidths, while enhancing flexibility and robustness, complicates band-slice assignments and user-UAV associations. Another layer of intriguing yet complex research involves optimizing handovers in high-mobility UAV networks, where both users and UAVs are mobile. UAV trajectory planning, which is already NP-hard even in static-user scenarios, becomes even more intricate to obtain optimal solutions in high-mobility user cases.

5G NEW RADIO ACCESS AND CORE NETWORK SLICING FOR NEXT-GENERATION NETWORK SERVICES AND MANAGEMENT

by
Abdullah Ridwan Hossain

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Helen and John C. Hartmann Department of
Electrical and Computer Engineering

December 2023

APPROVAL PAGE

5G NEW RADIO ACCESS AND CORE NETWORK SLICING FOR NEXT-GENERATION NETWORK SERVICES AND MANAGEMENT

Abdullah Ridwan Hossain

Dr. Nirwan Ansari, Dissertation Advisor — Date
Distinguished Professor, Department of Electrical and Computer Engineering, NJIT

Dr. Ali N. Akansu, Committee Member — Date
Professor, Department of Electrical and Computer Engineering, NJIT

Dr. Abdallah Khreishah, Committee Member — Date
Professor, Department of Electrical and Computer Engineering, NJIT

Dr. Roberto Rojas-Cessa, Committee Member — Date
Professor, Department of Electrical and Computer Engineering, NJIT

Dr. Cristian Borcea, Committee Member — Date
Professor, Department of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Abdullah Ridwan Hossain

Degree: Doctor of Philosophy

Date: December 2023

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2023
- Master of Science in Electrical Engineering,
  The City College of New York, New York, NY, 2019
- Bachelor of Engineering in Electrical Engineering,
  The City College of New York, New York, NY, 2017

Major: Electrical Engineering

Presentations and Publications:

A. R. Hossain, A. Kiani, T. Saboorian, A. Xiang, J. Kaippallimalil, and N. Ansari, "AI/ML-Based Sensing-Assisted Edge Computing in Next-Generation Mobile Networks," *IEEE 2023 Conference on Standards for Communications and Networking*, accepted.

A. R. Hossain, W. Liu, N. Ansari, A. Kiani, and T. Saboorian, "AI-Native for 6G Core Network Configuration," *IEEE Networking Letters*, DOI: 10.1109/LNET.2023.3302833, early access.

A. R. Hossain, M. A. Hossain and N. Ansari, "Dual-Band Aerial Networks for Priority-Based Traffic," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 7, pp. 9500-9510, Mar. 2023.

M. A. Hossain, **A. R. Hossain**, W. Liu, N. Ansari, A. Kiani and T. Saboorian, "A Distributed Collaborative Learning Approach in 5G+ Core Networks," *IEEE Network*, DOI: 10.1109/MNET.133.2200527, early access.

A. R. Hossain and N. Ansari, "5G Multi-Band Numerology-Based TDD RAN Slicing for Throughput and Latency Sensitive Services," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1263-1274, Mar. 2023.

M. A. Hossain, **A. R. Hossain** and N. Ansari, "5G NR Numerology in UAV-Based Mobile Edge Computing for Massive IoT Networks," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23860-23868, Dec. 2022.

M. A. Hossain, **A. R. Hossain** and N. Ansari, "AI in 6G: Energy-Efficient Distributed Machine Learning for Multilayer Heterogeneous Networks," *IEEE Network*, vol. 36, no. 6, pp. 84-91, Nov/Dec. 2022.

A. R. Hossain and N. Ansari, "Priority-Based Downlink Wireless Resource Provisioning for Radio Access Network Slicing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9273-9281, Sept. 2021.

Acquire knowledge and impart it to the people.

Prophet Muhammad

ACKNOWLEDGMENTS

My utmost gratitude is to God for his infinite grace and mercy upon me.

My deepest appreciation is directed to my mentor, Dr. Nirwan Ansari, who nominated me for the Provost Assistantship that financed my research early in my journey. His relentless pursuit of perfection has impacted me beyond expression. I was blessed to join his lab, carve my own niche, and impart a, hopefully, lasting contribution to his lab. Without his genuine concern for my growth and success, my research journey would not have materialized, let alone progress this far.

I would like to express much gratitude to my dissertation committee members, Dr. Ali N. Akansu, Dr. Cristian Borcea, Dr. Abdallah Khreishah, and Dr. Roberto Rojas-Cessa, for their thoughtful insights and time.

I am immensely appreciative to the Department of Electrical and Computer Engineering for supporting me throughout my doctoral tenure; special thanks are due to the National Science Foundation for supporting my research (via Grant No. CNS-1814748).

It gives me great pleasure to mention by nearest and dearest colleagues: Mohammad Arif Hossain, Weiqi Liu, Shuai Zhang, Di Wu, Jingjing Yao, Liang Zhang, and many others who have always lent me their time and offered advice when I needed it most.

I would like to express my love and thanks to all my family and friends. Finally but most importantly, I cannot ever do justice to my dear parents, Dr. ASM Delowar Hossain and Mrs. Saifunessa Begum, whose immeasurable sacrifices and undying love made me who I am today. May God Almighty envelop them both with His blessings in this life and in the next.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The recent explosive trend in the volume, diversity, and stringency of mobile traffic has made clear that the current networking infrastructure has overstayed its welcome [1]. Furthermore, the envisioned fifth generation (5G) use cases, categorized as enhanced mobile broadband (EMBB), massive machine type communications (MMTC), and ultra-reliable low latency communications (URLLC), are expected to stress the radio access networks (RANs) far beyond their original design and capabilities [2]. In a bid to better support broad-spectrum services, countless avenues have been visited to remedy the one-size-fits-all approach that current networks are notorious for. Among such remedies, network slicing, which entails partitioning a physical network into multiple virtual networks tailored for specific quality-of-service (QoS) constraints, is considered to be among the more significant paradigm shifts [3]. The aim of network slicing is broad in that it seeks to guarantee QoS requirements and the highest degrees of isolation while achieving seamless and optimal end-to-end (E2E) slicing [4]. As such, its focus is primarily three-fold: the access [5], transport [6, 7, 8], and core networks [9].

Starting at RAN slicing, a significant chunk of the spotlight has been placed on slicing costs [10, 11, 12], orchestration, administration, and management (OAM) [13, 14, 15] while others have placed a high importance on user admission control (UAC) and energy efficiency [16, 17, 18]. While being praiseworthy initiatives, they have been very narrowly limited to frequency division duplex (FDD) Long Term Evolution (LTE)-based networks known for their rigid sub-carrier spacing (SCS) and time slot duration. In anticipation of the next leap, the 3GPP in Release 17 proposed 5G New Radio (NR) standards allowing for flexible time division duplex (TDD) time

slot configuration, joint sub-6 GHz and mmWave band transmission, numerology (scalable SCS), and other modifications.

While such motions are commendable, they are relatively narrow as they do not consider the unmanned aerial vehicle (UAV) networks that, in many practical scenarios, serve as intermediaries between RANs and users. UAV networks are an attractive option due to their ease of deployment, size, and mobility; they step in where RAN coverage is limited or line-of-sight (LoS) channels are minimal. Typically, the primary concern in aerial network engineering tends to be UAV placement and trajectory planning but with legacy radio protocols, *i.e.*, LTE [19, 20]. Nevertheless, advancing the state-of-the-art would require combining 5G NR with network slicing in UAV networks. Needless to say, such an integration brings about another set of challenges not typically shared with RANs including LoS decay resulting from blockage and obstruction. Aerial nodes are very prone to such due to their 3D mobility. Therefore, sliced 5G NR aerial networks require holistic resource allocation approaches and band admission policies that account for dual-band transmission, LoS, and service types. If these considerations are important for RANs, they are far more crucial for aerial networks.

Although far removed from the access end, the core network (CN) cannot be overlooked; it consists of numerous vital network functions (NFs) which execute control tasks to ensure that the E2E QoS requirements are met. Failure to do so in a timely fashion would hamper QoS flows, user sessions, and degrade traffic routing from the RAN to external destinations. Consequently, it can even be argued that the CN performance is perhaps the most important of all since it serves as an intermediary between access and third-party networks. Unfortunately, the CN has yet to be touched by much of the innovation and advancement that has taken place at other segments of the network. Currently, the CN lacks true E2E slicing and isolation because although its user plane (UP) is sliced, the control plane (CP) is

not. Moreover, the CN does not offer physical partitioning between its planes, thus degrading its reliability. In order to shore up its reliability, isolation, and security, both physical planar separation and CP slicing should be considered. Of course, resource allocation and NF assignment strategies need to be optimized since the added restrictions do come at a slight performance cost.

1.1 Numerology and Multiplexing in Time Division Duplex Networks

A basic understanding of the 5G NR numerology schemes is key to forming the basis of the outstanding questions with respect to access network optimization. Furthermore, a working description of FDM within a dual-band TDD network will properly motivate our subsequent discussions. The 5G NR numerology concept essentially does away with the static SCS of LTE networks; instead, it defines five new SCSs that can be used alongside the base SCS (15 kHz). Since a resource block (RB) always consists of 12 sub-carriers and 14 symbols regardless of the scheme, it expands or narrows accordingly on the frequency-time resource grid as depicted in Table 1.1 and visually illustrated in Figure1.1. The first two schemes are restricted to the sub-6 GHz band while the last three are confined to the mmWave band; the third scheme, however, is common to both. Assuming a scheme were to be designated by μ, the SCS spacing, RB width, and time slot duration would be $15 * 2^{\mu}$ kHz, $180 * 2^{\mu}$ kHz, and $\frac{1}{2^{\mu}}$ ms long, accordingly.

Table 1.1 5G NR Numerology Schemes

Numerology scheme (μ)	0	1	2	3	4	5
Sub-carrier spacing (kHz)	15	30	60	120	240	480
Resource block width (kHz)	180	360	730	1440	2880	5760
Cyclic prefix length (µs)	4.8	2.4	1.2	0.6	0.3	0.15
Symbol length (µs)	66.67	33.33	16.67	8.33	4.17	2.08
Time slot duration (ms)	1	0.5	0.25	0.125	0.0625	0.0313

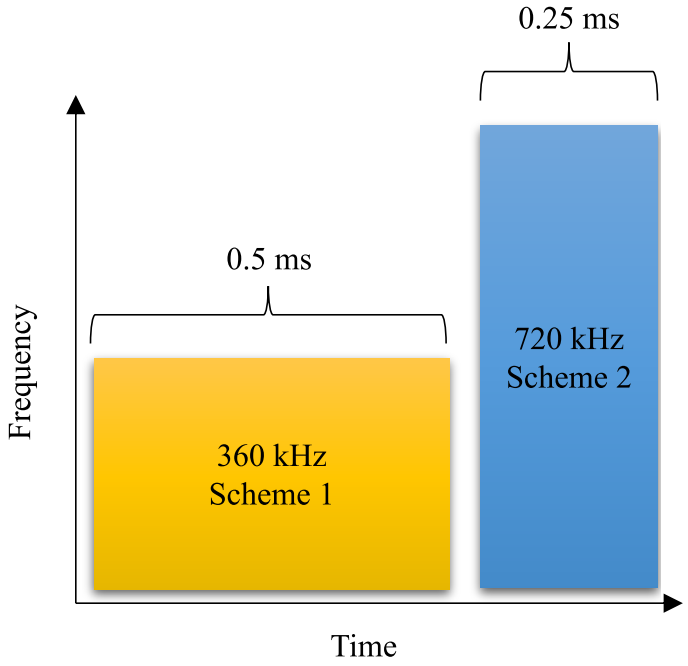

Figure 1.1 Frequency-time grid comparison of numerology schemes.

As can be seen in Figure 1.1, higher schemes enable larger RB widths which in turn allow for shorter time slots resulting in a lower latency conducive for URLLC services. They also facilitate higher throughput for EMBB use cases even under poorer channel conditions [21] primarily because the signal to noise ratio (SNR) can be lower for a given throughput requirement owing to the increased RB width which compensates for the throughput loss otherwise incurred with the standard SCS; this undoubtedly enhances channel reliability and data resiliency [22].

The rationale as to why a TDD system, as shown in Figure 1.2, is being considered in this work despite that the majority of networks utilize FDD, is that the use of higher numerology schemes requires significantly much higher channel bandwidths in FDD systems. If higher channel bandwidths are not available at higher schemes, there will be less RBs available for allocation to the user equipment (UEs), potentially resulting in QoS violations. It can result in significant queuing delays if the network does not have a sufficient amount of RBs to provision to a UE since

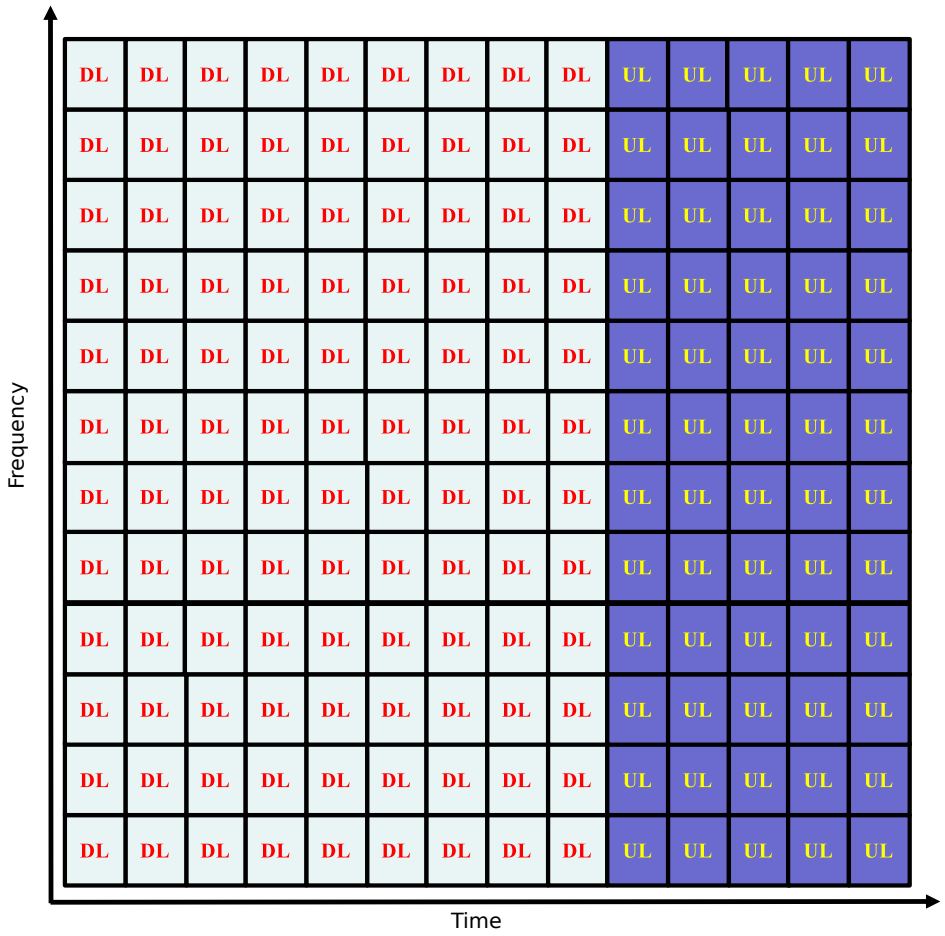

Figure 1.2 Time division duplex in a resource block; each column is a symbol.

the total channel bandwidth does not scale accordingly with the numerology scheme. The UE must wait for the next time slot to be scheduled (best case) or even several time slots (worst case); this surely does not bode well for URLLC services. TDD systems, on the other hand, overcome this limitation by consolidating both directions of communication over a single transmission band. Furthermore, TDD systems do not require frequency guard-bands between uplink and downlink communications; they can also better adapt to and balance between asymmetrical uplink and downlink loads, thus further enhancing spectral efficiency and network adaptability [23, 24].

From a physical layer standpoint, antenna designs for TDD devices are less complex; this is especially important for massive multiple-input multiple-output (MIMO) implementation. TDD channel estimation tends to be faster and simpler

since the same spectrum is utilized for both directions of communications. Along with other motivations, TDD is the preferred choice for massive MIMO which is a major next-gen access technology; and hence, we consider a TDD RAN in this dissertation [25, 26, 27].

The FDM of numerology schemes entails dedicating channel bandwidths of different carrier frequencies to different numerology schemes. As a result, multi-band RANs can support multiple schemes unlike single-band RANs. Network slices can be assigned to schemes and/or bands. A pictorial depiction of such is provided in Figure 1.3.

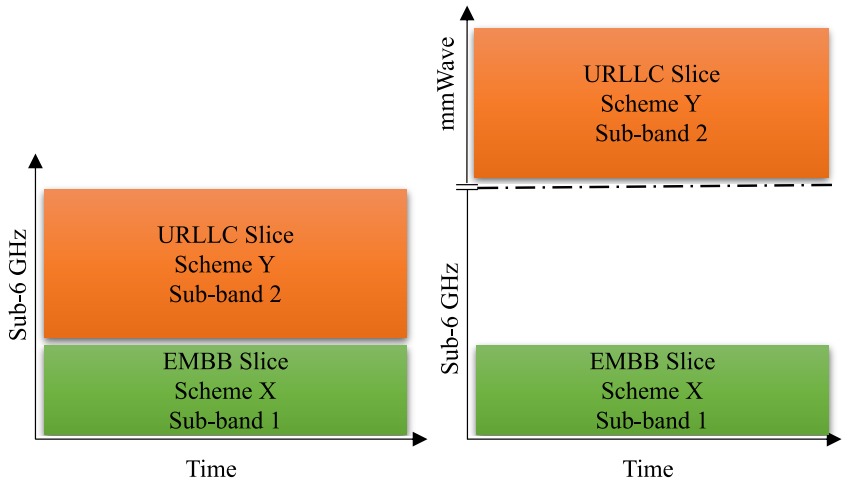

Figure 1.3 Frequency division multiplexing of numerology schemes.

Additionally, although TDD RANs have their benefits, they do make it more difficult to satisfy directional constraints since a time slot is utilized for bi-directional transmission. Hence, we can now pose the following questions:

1. What is the impact of TDD (duplex ratio) on network performance?
2. Which numerology scheme is most optimal and for which aim?
3. What effect does the mmWave schemes have on ground and aerial networks?
4. How can we regulate band access to the RAN services and users?

5. How does the wider SCS affect SNR and RB availability for a fixed bandwidth?

6. How can FDD NR aerial networks minimize the QoS gaps of best-effort users?

7. What admission policies are required to manage band access in aerial networks?

As for the CN, it too has a unique set of issues that need to be resolved. As it stands, the CN does not offer the highest degree of E2E reliability, slicing isolation, or security. Additionally, there are numerous vital NFs that need to support the RAN since any of them can result in a bottleneck for access network traffic. A unique set of questions in this regard can be presented as follows:

1. How can we optimally allocate resources for the NFs to execute their tasks?

2. How can we maintain slicing with control and user plane separation (CUPS)?

3. Is the performance enhancement of physical planar separation justifiable?

4. For which services are each of the proposed CN configurations most optimal?

We seek to address these outstanding issues at the RAN, UAV network, and CN by developing extensive problem formulations, solving them, and assessing their corresponding simulation results in this dissertation which is organized as follows: in Chapter 2, we summarize the state-of-the-art advances with respect to the above discussed points. We then outline the vital gaps that this work seeks to bridge to facilitate next-generation networking. In Chapter 3, we present our dual-band TDD numerology-based RAN in which we multiplex throughput-dependent and latency-sensitive slices across the sub-6 GHz and mmWave bands. Depending on the service type, we either maximize the spectral efficiency or minimize the UE transmission power while meeting the QoS constraints. We explore the impact of primarily the duplexing ratio on the time domain, numerology scheme, and transmission band on the spectral efficiency of the throughput-dependent slice and the power minimization

of the latency-sensitive slice. In Chapter 4, we present our dual-band UAV network in which we provision guaranteed and best-effort services across both the sub-6 GHz and mmWave bands. We attempt to minimize the QoS gap of the best-effort service while meeting the QoS constraints of the guaranteed services. In Chapter 5, we propose and implement three CN configurations to enhance the reliability and isolation of slices in an effort to reduce the E2E latency. In Chapter 6, we discuss avenues of future research which are believed to be instrumental in furthering next-generation networking, such as coordinated multi-point (CoMP) with multi-UAV networks, E2E network slicing, and integrated sensing and communications (ISAC). Finally, in Chapter 7, we conclude this dissertation by summarizing the research problems addressed herein and their corresponding results.

CHAPTER 2

RELATED WORK

A reasonable amount of works incorporating the 5G NR numerology schemes into RANs has proliferated the research literature. Lagen *et al.* [28] optimized load balancing across multi-numerology slices for bi-directional communications for a single-band network. Patriciello *et al.* [29] studied the latency experienced by the physical layer in a multi-numerology but non-sliced mmWave network. Ha *et al.* [30] investigated user admission control policies for multi-numerology slices while Zhang *et al.* [31] incorporated machine-learning for time slot scaling with fixed SCS for non-sliced networks. Diez *et al.* [32] studied the utilization of numerology over multiple radio access technology (multi-RAT) networks from the perspective of multi-connectivity user scheduling.

While these works are significant, there have not been ample studies on the effect of multiplexing different services on different bands within a TDD network. Additionally, the impact of the numerology schemes on the spectral efficiency of dual-band networks is crucial; it stands to reason that widening the RB bandwidth would have a noticable impact, for better or for worse, on the resource utilization and efficiency given the fixed channel bandwidth of a network. Numerology also has a sizable impact on the transmission power of both the base stations (BSs) and UEs and should be studied accordingly.

With respect to UAV communications, there has been a plethora of works especially related to optimal placement, trajectory, and resource allocation. Sun *et al.* [33] investigated UAV placement and resource allocation strategies to minimize the latency experienced by users within a hotspot. Al-Hourani *et al.* [34] derived the optimal UAV elevation which is dependent on the users' maximum pathloss thresholds and the radial coverage of the UAV. Alzenad *et al.* [35] maximized the number

of served users by the UAV while minimizing downlink transmission power. More recently, Yang *et al.* [36] utilized machine learning for user location prediction and channel estimation. Gui *et al.* [37] investigated the use of mmWave bands in UAV networks for wireless recharging and connectivity in disaster scenarios. Wu *et al.* [38] explored the numerous challenges with UAV communications and the provisioning of resources for different purposed slices for 5G. They specifically elaborated upon the difficulties of incorporating massive-input-massive-output antennae, mmWave communications, and NOMA schemes in UAV networks. Hsu *et al.* [39] studied an IoT network which offloads tasks to the cloud servers utilizing both licensed and unlicensed 5G radio spectra. Their proposed algorithms successfully minimized the blocking probability and enhanced power savings and increased user throughputs. Weerasinghe *et al.* [40] studied grant-free resource allocations for mMTC traffic with dynamic time slot formats. Ansari *et al.* [41] proposed the use of free-space optics to provide both charging and backhaul functionalities for UAVs in order to alleviate the burden on the RF fronthaul while elongating the UAV's total flight time. Hossain *et al.* [42] proposed the allocation of numerology schemes at a highly granular level, on a per-device level in a mobile edge computing Internet-of-Things (IoT) network to maximize the flexibility in resource block tiling and network spectral efficiency while minimizing the energy consumption of the network and maintaining the deadlines of the offloaded tasks of the devices. Yin *et al.* [43] investigated user clustering, transmission power allocation, and content caching in a NOMA-based multi-UAV network via the ρ-K clustering and cross layer allocation methods; they took into consideration both the instantaneous and statistical QoS constraints to maximize the contents' hit probability while minimizing their outage probability. While the integration of optical communications into aerial networks is not a new phenomenon, Tadayyoni *et al.* [44] exploited the ultraviolet (UV) spectrum instead of the conventional optical wavelengths to enable the collection of data from devices

situated within an IoT farm; they studied the bit error rate (BER) and derived its closed-form expression which was then verified by simulations. Furthermore, they demonstrated that despite the leakage of UV transmission between one node to another, with the proper displacement choice between said nodes, the effect of interference is minimal and the performance of the system remains identical to that of the no-interference case.

While recent works are quite impactful and have advanced the state of the art, there remains much to be explored. mmWave has been deployed experimentally on a growing scale across different settings. For a UAV to truly indeed function as a transparent relay node between a BS and UE, it must use the same transmissions bands of the BSs. Since BSs are expected to utilize numerology schemes and dual-band transmissions, UAVs should support the same to truly function as a transparent arm of RANs (ground networks). In ultra-dense UAV networks, it is improbable to provision all UEs without any QoS degradation thus necessitating the triaging of UEs when allocating resources. Within the UAV context, in this dissertation, we consider the channel condition of each individual RB as opposed to assuming an average channel condition across all RBs for a carrier frequency. This gives us a much more realistic depiction of practical networks which must consider the frequency-dependent channel gains on each RB of a wireless network for optimal wireless resource allocation. Although the effect of the per-RB dependency is negligible in the sub-6 GHz region, it is extremely pronounced in the mmWave band and greatly impacts the overall resource allocation. This impact is considered in our resource allocation scheme.

Shifting from the RANs to CNs, there recently have been appreciable steps made in an effort to bridge the RAN and CN research gap. Manias *et al.* [45] implemented a virtualized CN to service a RAN user to investigate the control packet contents and traffic characteristics for inter-NF communications. Chouman *et al.*

[46] designed a virtual Network Data Analytics Function (NWDAF) and observed the traffic patterns between the NFs and NWDAF. Du *et al.* [47] implemented a virtual Access and Mobility Function (AMF) of the CN to dissect the signalling between a RAN user and CN when executing device registration. Alawi *et al.* [48] proposed load balancing via a control theory-based dynamic scaling algorithm for the AMF while minimizing the CN response time. Sattar and Matrawy [49] optimized the route selection to minimize the latency for inter-NF communication within a slice. Salhab *et al.* [50] investigated flexible offloading of CN computing onto the cloud to minimize CN operation costs while maximizing the resource utilization efficiency depending on the time-varying RAN loads.

A significant limitation of a majority of these works is that they treat the CN as a mere computational data center, neglecting the qualitative distinctions between different NFs. The NFs are modeled as generic virtual functions, lacking a genuine qualitative or quantitative reflection of the workload specific to each NF. More importantly, these works often fail to consider CUPS of the CN as required by 3GPP. Consequently, this study focuses on investigating the current 3GPP CN slicing standard and operating configuration, proposing three additional operating configurations that account for true inter-NF traffic characteristics affecting the NFs latency. Furthermore, we emphasize the importance of CUPS in NF placement, resource allocation, and optimization.

CHAPTER 3

5G MULTI-BAND NUMEROLOGY-BASED TDD RAN SLICING

In this chapter, we investigate the impact of numerology over the sub-6 GHz and mmWave bands in a TDD RAN. As the maximum channel bandwidths vary with respect to the transmission band, the incorporation of numerology over multiple bands has major consequences on spectral efficiency, throughput, and latency. In this light, we formulate two optimization problems to maximize the average spectral efficiency (ASE) of an EMBB slice over the sub-6 GHz band and minimize the UE's transmission powers in a URLLC slice over the mmWave band, respectively. Since these problems are mixed-integer nonlinear programming (MINLP) problems, we relax them into convex approximations and develop a low-complexity algorithm to solve them. Finally, through our simulations, we assess the impact of multi-band TDD numerology networks on downlink spectral efficiency, uplink transmission power, and latency.

We now outline the organization of this chapter: first, we present our system model in Section 3.1. Next, we formulate the first optimization problem in Section 3.2 and outline our proposed low-complexity algorithm to solve it in Section 3.3. Then, we formulate our second optimization problem in Section 3.4. We discuss the results of both optimization problems in Section 3.5. Finally, we summarize the problems addressed herein, the adopted approach, and obtained results in Section 3.6.

3.1 System Model

We begin with our system model for our TDD RAN. Consider a single orthogonal frequency division multiple access (OFDMA) cell with a BS capable of multi-band transmission, specifically over both the sub-6 GHz and mmWave bands. The UEs are randomly located within the coverage area of the BS and are assumed to operate

on a single-band transmission mode only. They are also anchored to a band and consequently, cannot switch between or aggregate transmission bands. We assume that there are two services that need to be provisioned for: EMBB and URLLC. The EMBB slice is limited to the sub-6 GHz band while the URLLC slice is limited to the mmWave band. As mentioned in Section 1.1, the mmWave-specific numerology schemes offer much lower latency due to the shortened time slot configurations; and hence, the URLLC slice is assigned to the mmWave band. The EMBB and URLLC users are denoted as $\mathcal{U}_{\mathcal{E}}$ and $\mathcal{U}_{\mathcal{L}}$, respectively.

The numerology schemes utilized by the EMBB and URLLC slices are represented by μ_E and μ_L, correspondingly. The total channel bandwidths to which the EMBB and URLLC slices have access are denoted by B_E and B_L, respectively; the channel bandwidth is dependent on the transmission band. Hence, the corresponding RBs of the slices are $|\mathcal{N}_{\mathcal{E}}| = \frac{B_E}{B_{RB}^E}$ and $|\mathcal{N}_{\mathcal{L}}| = \frac{B_L}{B_{RB}^L}$, where B_{RB}^E and B_{RB}^L are the RB bandwidths resulting from μ_E and μ_L, respectively (refer to Table 1.1). The total RB symbol count is denoted by S_T and is assumed to be identical for both slices. We define the downlink symbol count for the EMBB slice as S_{dl}^E, the uplink symbol count as S_{ul}^E, and the downlink duplex ratio as $\frac{S_{dl}^E}{S_T}$. For the URLLC slice, the uplink symbol count is S_{ul}^L, the downlink symbol count is S_{dl}^L, and the duplex ratio is $\frac{S_{ul}^L}{S_T}$. We assume that the BS has a maximum transmission power of P_E and P_L over the sub-6 GHz and mmWave bands, respectively. On the other hand, the UEs have a maximum uplink transmission power of p. The notations utilized in this chapter are tabulated in Table 3.1.

Note that for EMBB use cases, the emphasis is on massive downlink throughput and spectral efficiency. The associated uplink throughput requirements are often quite lax. As a result, we primarily concern ourselves with the downlink communications for the EMBB slice when we formulate our problem. In contrast, uplink traffic is given much more importance in the URLLC use cases and its throughput requirements

Table 3.1 Summary of Notations for RAN

Notation	Definition
a_n^u	Binary variable if RB n is allocated to UE u of the EMBB slice.
B_{RB}^E	RB bandwidth of the EMBB slice.
B_{RB}^L	RB bandwidth of the URLLC slice.
D_E	downlink throughput requirement of the EMBB slice.
D_L	downlink throughput requirement of the URLLC slice.
g_u	Channel gain of UE u of the EMBB slice.
g_v	Channel gain of UE v of the URLLC slice.
g_n^u	Channel gain of RB n for UE u of the EMBB slice.
N_0	Noise spectral density.
N_u	Amount of RBs allocated to UE u of the EMBB slice.
N_v	Amount of RBs allocated to UE v of the URLLC slice.
p	Maximum user transmission power.
p_u	Transmission power of UE u of the EMBB slice.
p_n^u	Transmission power of UE u on RB n of the EMBB slice.
p_v	Transmission power of UE v of the URLLC slice.
P_E	Maximum BS transmission power for the EMBB slice.
P_L	Maximum BS transmission power for the URLLC slice.
P_n^u	BS transmission power on RB n for UE u of the EMBB slice.
P_u	BS transmission power for UE u of the EMBB slice.
P_v	BS transmission power on RB v of the URLLC slice.
S_{dl}^E	Symbol count for downlink transmission for the EMBB slice.
S_{ul}^E	Symbol count for uplink transmission for the EMBB slice.
S_{ul}^L	Symbol count for uplink transmission for the URLLC slice.
S_{dl}^L	Symbol count for downlink transmission for the URLLC slice.
S_T	Total symbol count of an RB.
T_{CP}	Cyclic prefix duration.
T_{sym}	RB symbol duration.
T	Uplink transmission latency requirement of the URLLC slice.
U_E	Uplink throughput requirement of the EMBB slice.
U_L	Uplink throughput requirement of the URLLC slice.
μ_E	Numerology scheme of the EMBB slice.
μ_L	Numerology scheme of the URLLC slice.
γ_u	ACGNR of UE u of the EMBB slice.
γ_v	ACGNR of UE v of the URLLC slice.
τ_{prop}^v	Propagation delay of UE v of the URLLC slice.

are negligible relative to the EMBB use case. URLLC use cases place a very high premium on latency and uplink transmission power [51, 52, 53].

3.2 EMBB Downlink Average Spectral Efficiency Maximization

In this section, we focus on the EMBB slice which resides on the sub-6 GHz band. We formulate an ASE maximization problem as follows:

$$\textbf{P1:} \max_{S_{dl}^E, a_n^u, P_n^u, p_n^u, \mu_E} \frac{S_{dl}^E}{S_T |\mathcal{U}_\mathcal{E}|} \sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} \sum_{n=1}^{|\mathcal{N}_\mathcal{E}|} a_n^u \log_2 \left(1 + \frac{g_n^u P_n^u}{B_{RB}^E N_0}\right) \tag{3.1}$$

$$\text{s.t. } \frac{S_{dl}^E}{S_T} \sum_{n=1}^{|\mathcal{N}_\mathcal{E}|} a_n^u B_{RB}^E \log_2 \left(1 + \frac{g_n^u P_n^u}{B_{RB}^E N_0}\right) \geq D_E, \forall u \in \mathcal{U}_\mathcal{E}, \tag{3.2}$$

$$\frac{S_{ul}^E}{S_T} \sum_{n=1}^{|\mathcal{N}_\mathcal{E}|} a_n^u B_{RB}^E \log_2 \left(1 + \frac{g_n^u p_n^u}{B_{RB}^E N_0}\right) \geq U_E, \forall u \in \mathcal{U}_\mathcal{E}, \tag{3.3}$$

$$\sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} \sum_{n=1}^{|\mathcal{N}_\mathcal{E}|} a_n^u P_n^u \leq P_E, \tag{3.4}$$

$$\sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} \sum_{n=1}^{|\mathcal{N}_\mathcal{E}|} a_n^u p_n^u \leq p, \tag{3.5}$$

$$\sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} \sum_{n=1}^{|\mathcal{N}_\mathcal{E}|} a_n^u \leq |\mathcal{N}_\mathcal{E}|, \tag{3.6}$$

$$\sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} a_n^u \leq 1, \forall n \in \mathcal{N}_\mathcal{E}, \tag{3.7}$$

$$a_n^u \in \{0, 1\}, n \in \mathcal{N}_\mathcal{E}, \forall u \in \mathcal{U}_\mathcal{E}, \tag{3.8}$$

$$\mu_E \in \{0, 1, 2\}, \tag{3.9}$$

$$P_n^u \geq 0, \forall n \in \mathcal{N}_\mathcal{E}, \forall u \in \mathcal{U}_\mathcal{E}, \tag{3.10}$$

$$p_n^u \geq 0, \forall m \in \mathcal{N}_\mathcal{E}, \forall v \in \mathcal{U}_\mathcal{E}, \tag{3.11}$$

$$S_{dl}^E + S_{ul}^E = S_T, S_{dl}^E, S_{ul}^E \in \mathbb{Z}^+. \tag{3.12}$$

The objective of Equation (3.1) is to maximize the downlink ASE of the EMBB slice with the following decision variables: a_n^u is a binary variable to indicate whether RB n is assigned to UE u, P_n^u is the BS transmission power on RB n for UE u, p_n^u is the uplink transmission power on RB n for said UE, μ_E is the numerology scheme for the EMBB slice, and S_{dl}^E determines how many symbols should be allocated for the downlink direction. In Equations (3.2)-(3.3), we formulate the constraints for the minimum throughput requirements for both directions of transmissions where D_E is the minimum throughput requirement of the EMBB slice in the downlink direction, U_E is the minimum throughput requirement of said slice in the uplink direction, g_n^u is the channel gain on RB n for UE u, and N_0 is the noise spectral density. As for Equations (3.4)-(3.5), we limit the maximum possible transmit power available in either direction of transmission. As for Equation (3.6, we enforce the total amount of RBs allocated not to exceed the available bandwidth. In Equation (3.7), we enforce each allocated RB should be utilized by one UE at most. We enforce the integer nature of the binary RB association indicator as well as for the numerology scheme in Equations (3.8)-(3.9), accordingly. Finally, Equations (3.10)-(3.11) allow positive continuous values for the downlink and uplink power allocation, respectively. Finally, in Equation (3.12), we ensure that the duplex ratio does not exceed the total RB symbol count.

As evident, **P1** is an MINLP problem, due to the non-linearity of Equations (3.1)-(3.3), binary and integer constraints of Equations (3.8)-(3.9) and Equation (3.12), which is complex to solve. In the interest of simplification, we transform the problem into a convex approximation and then propose a solution algorithm

that determines the optimal numerology scheme. First, we assume that a UE experiences an average channel gain over all its allocated RBs; and hence, the power and bandwidth allocation is decoupled from each individual RB. Because of this assumption, each UE also experiences an average channel gain to noise ratio (ACGNR) $\gamma_u = \frac{g_u}{B^E_{RB} N_0}$; this is true for both uplink and downlink of both slices. We study the cost of this assumption with respect to accuracy and optimality in a slightly different context in [54]. A new decision variable is introduced with the the relaxation of its integer domain to represent the amount of RBs assigned to each user, N_u. Hence, **P1** becomes

$$\textbf{P2:} \quad \max_{N_u, P_u, p_u, S^E_{dl}} \frac{S^E_{dl}}{S_T |\mathcal{U}_\mathcal{E}|} \sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} N_u \log_2 \left(1 + \frac{\gamma_u P_u}{N_u}\right) \tag{3.13}$$

$$\text{s.t.} \ \frac{S^E_{dl}}{S_T} B^E_{RB} N_u \log_2 \left(1 + \frac{\gamma_u P_u}{N_u}\right) \geq D_E, u \in \mathcal{U}_\mathcal{E}, \tag{3.14}$$

$$\frac{S^E_{ul}}{S_T} B^E_{RB} N_u \log_2 \left(1 + \frac{\gamma_u p_u}{N_u}\right) \geq U_E, u \in \mathcal{U}_\mathcal{E}, \tag{3.15}$$

$$\sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} N_u \leq |\mathcal{N}_\mathcal{E}|, N_u \in \mathbb{R}^+, \tag{3.16}$$

$$\sum_{u=1}^{|\mathcal{U}_\mathcal{E}|} P_u \leq P_E, P_u \in \mathbb{R}^+, \tag{3.17}$$

$$p_u \leq p, p_u \in \mathbb{R}^+, u \in \mathcal{U}_\mathcal{E}, \tag{3.18}$$

$$S^E_{dl} + S^E_{ul} = S_T, S^E_{dl} \in \mathbb{R}^+. \tag{3.19}$$

The objective function in Equation (3.1) is adjusted as per the ACGNR assumption and becomes that of Equation (3.13). Similarly, Equations (3.2)-(3.3)

are transformed to Equations (3.14)-(3.15). Due to the aforementioned assumptions, a_n^u no longer serves any useful function because it is the total bandwidth being allocated to the users. The problem does not decide on a per-RB basis, thus eliminating Equations (3.7)-(3.8). As a result, the transmission power constraints which were originally written for a per-RB basis in Equations (3.4)-(3.5) and Equations (3.10)-(3.11) are consolidated into Equations (3.17)-(3.18). The maximum bandwidth constraint which was written in terms of the binary indicator variable (a_n^u) in Equation (3.6) is transformed to that of Equation (3.16). Note the relaxation of the integer domain to a continuous domain in the above conversion. The duplex ratio constraint of Equation (3.12) which was originally constrained to an integer domain is now relaxed to continuous values in Equation (3.19). Equation (3.9) will be inherently handled by our solution algorithm presented later. Likewise, although Equation (3.19) is a continuous function, we exploit our algorithm to ensure integer values for the RB symbol counts.

We proceed to prove the concavity of **P2**. Equations (3.16)-(3.19) are linear functions with continuous domains. Thus, if we can prove Equation (3.13) to be concave which also proves the concavity of Equations (3.14)-(3.15), then **P2** will be a concave optimization problem. We begin by factoring out S_{dl}^E from Equation (3.13) as it is simply a scalar; let us now prove that a function of the form $E(a,b) = a\log\big(1+\frac{b}{a}\big)$ is concave with respect to a and b. The Hessian matrix, $\nabla^2 E$, is determined as follows:

$$\nabla^2 E = \begin{bmatrix} -\frac{b^2}{a^3(1+\frac{b}{a})^2\ln 2} & \frac{b}{a^2(1+\frac{b}{a})^2\ln 2} \\ \frac{b}{a^2(1+\frac{b}{a})^2\ln 2} & -\frac{1}{a(1+\frac{b}{a})^2\ln 2} \end{bmatrix}. \tag{3.20}$$

As the diagonal terms are negative, $\nabla^2 E$ is negative semi-definite; and hence, $E(a,b)$ is concave. A summation of concave functions is also a concave function. Thus, **P2** is a concave optimization problem which can be optimally solved via CVX or CPLEX.

3.3 Proposed Algorithm

In this section, we present our solution algorithm to determine the optimal duplex ratio and numerology scheme. It is designed to solve the optimization problem by cycling through each numerology scheme, obtaining the resultant objective function's value, and selecting the scheme which results in the maximum objective function value. While MINLP problems in general need to comb through a wide range of possible integer values without *a priori* knowledge, we can narrow down our range prior to executing the algorithm (since we know the restricted integer values μ_E or μ_L can take on). We then ensure the integer nature of the symbol count by utilizing the ceiling function.

Algorithm 1: Numerology and Duplex Selection

Input: Wireless network parameters and QoS requirements

Output: Optimal numerology scheme, resource, power, and duplex allocations

1 ASE^*, N_u^*, P_u^*, p_u^*, S_{dl}^{E*}, μ_E^*, $\mu_E = 0$
2 **while** $\mu_E < 3$ **do**
3 Solve **P3** via CVX
4 Obtain ASE^*, N_u^*, P_u^*, p_u^*, S_{dl}^{E*}
5 Set $S_{dl}^E = \lceil S_{dl}^{E*} \rceil$ in **P3** and solve via CVX
6 **if** *Equation (3.13)* $> ASE^*$ **then**
7 $\mu_E^* = \mu_E, N_u^* = N_u, P_u^* = P_u, p_u^* = p_u, S_{dl}^{E*} = S_{dl}^E$
8 **end**
9 $\mu_E = \mu_E + 1$
10 **end**

We now analyze the complexity of our proposed algorithm. The complexity of Line 3 is $O(|\mathcal{U}_\mathcal{E}|^{N-1})$ where N is the number of variables to be solved for; note that S_{dl}^E is not user dependent, *i.e.*, it is enforced homogeneously over the entire slice. The complexity of Step 5 is identical to that of Step 3. Lines 6-9 are a simple compare and copy operation and as such, the complexity is simply $O(1)$. Lastly, all the preceding

steps have to be repeated for at most X times as per Line 2, where X is the number of possible numerology schemes in a particular band of transmission (*i.e.*, the total number of iterations that the while loop must execute); thus, the overall complexity of our algorithm is $O(X(|\mathcal{U}_\mathcal{E}|^{N-1}+|\mathcal{U}_\mathcal{E}|^{N-1}+1))$ but can be simplified to $O(X(|\mathcal{U}_\mathcal{E}|^{N-1}))$.

3.4 URLLC Uplink Transmission Power Minimization

In the same fashion that we transformed **P1** into **P2**, we directly formulate our URLLC-specific problem, **P3**. As stated earlier, the URLLC slice is assigned to the mmWave band. The goal is to minimize the total uplink transmission power as follows, where the decision variables are N_v which denotes the amount of RBs allocated to UE v, P_v represents the downlink transmission power of the BS for said UE, p_v is the uplink transmission power of a UE, and S_{ul}^L is the symbol count for uplink communications:

$$\textbf{P3:} \quad \min_{N_v,P_v,p_v,S_{ul}^L} \sum_{v=1}^{|\mathcal{U}_\mathcal{L}|} p_v \tag{3.21}$$

$$\text{s.t. } \frac{S_{dl}^L}{S_T} B_{RB}^L N_v \log_2\left(1+\frac{\gamma_v P_v}{N_v}\right) \geq D_L, v \in \mathcal{U}_\mathcal{L}, \tag{3.22}$$

$$\frac{S_{ul}^L}{S_T} B_{RB}^L N_v \log_2\left(1+\frac{\gamma_v p_v}{N_v}\right) \geq U_L, v \in \mathcal{U}_\mathcal{L}, \tag{3.23}$$

$$\sum_{v=1}^{|\mathcal{U}_\mathcal{L}|} N_v \leq |\mathcal{N}_\mathcal{L}|, N_v \in \mathbb{R}^+, \tag{3.24}$$

$$\sum_{v=1}^{|\mathcal{U}_\mathcal{L}|} P_v \leq P_L, P_v \in \mathbb{R}^+, \tag{3.25}$$

$$p_v \leq p, p_v \in \mathbb{R}^+, v \in \mathcal{U}_\mathcal{L}, \tag{3.26}$$

$$S_{dl}^L + S_{ul}^L = S_T, S_{ul}^L \in \mathbb{R}^+, \tag{3.27}$$

$$\frac{1}{2^{\mu_L}}\left(T_{CP}+T_{sym}S_{ul}^{L}\right)+\tau_{prop}^{v}\leq T, v\in\mathcal{U}_{\mathcal{L}}. \tag{3.28}$$

We ensure that the downlink and uplink data rates are met in Equations (3.22)-(3.23), respectively, where γ_v is the channel gain between the BS and UE v. In Equation (3.24). We ensure that the total number of RBs allocated does not exceed the channel bandwidth. Equation (3.25) enforces the maximum power budget of the BS while Equation (3.26) does the same for the UE. Equation (3.27) enforces that the symbols allocated to both directions of transmission equal to the total symbol count. Finally, Equation (3.28) ensures the RAN latency requirement of the slice to be met. This problem is also a concave optimization problem and can be solved via the proposed algorithm in Section 3.3.

3.5 Simulation Results

In this section, we discuss our results on a per-slice basis, starting with the EMBB slice. We assume that the time slot is narrow enough to neglect fading and user mobility. The simulations were conducted through the CVX interface in MATLAB. For each data point, Monte Carlo simulations of a 100 runs were executed.

3.5.1 EMBB slice performance

We set the amount of UEs to 20. The BS has a maximum transmission power of 50 dBm, coverage radius of 500m, and operates at a 2.5 GHz carrier frequency with a total channel bandwidth of 100 MHz. The RB symbol count is assumed to be the default, *i.e.*, which is 14 symbols. We compare the downlink ASE for each of the numerology schemes. Note that scheme 0 is the LTE scheme (baseline) against which all other schemes should be evaluated. The simulation parameters are presented in Table 3.2.

In Figure 3.1, the ASE and average throughput (ATP) are plotted for each numerology scheme. It can be observed that scheme 0 is not only the most efficient,

Table 3.2 Simulation Parameters for RAN

Parameter	Value
Total sub-6 GHz channel bandwidth	100 MHz
Total mmWave channel bandwidth	[100, 400] MHz
Sub-6 GHz carrier frequency	2.5 GHz
mmWave carrier frequency	28 GHz
Downlink EMBB throughput requirement	15 Mbps
Uplink EMBB throughput requirement	0.5 - 5 Mbps
Uplink URLLC throughput requirement	1 Mbps [55]
Downlink URLLC throughput requirement	0.5 - 15 Mbps
Maximum BS Transmission Power	50 dBm
Maximum UE Transmission Power	24 dBm
BS coverage radius	500 m
Number of UEs	20
Resource block symbol count	14
Noise spectral density	-174 dBm/Hz
Air-interface deadline	0.5 ms [56]
Channel Model	Free Space Pathloss (FSPL)

but it is uniquely efficient, as shown by the absence of any overlap of the ASE of the schemes. The average ASE of scheme 0 is approximately 276 bps/Hz; for scheme 1, it is 155 bps/Hz while for scheme 2, it is 87 bps/Hz. Interestingly, the gap between the lowest achieved ASE of each scheme and the highest ASE of the adjacent scheme decreases as the network progresses to using higher schemes. Furthermore, the range of ASE with each progressing scheme narrows greatly, meaning that the UEs are experiencing more similar spectral efficiencies at the highest scheme. The decrease in ASE with each progressive scheme can be attributed to several factors:

1. Since higher schemes utilize wider SCSs, assuming a fixed channel bandwidth, the amount of RBs available for allocation to the UEs decrease. Each progressive scheme decreases the RB amount by at least half; therefore, a UE will be provisioned fewer RBs in each progressive scheme.
2. At higher schemes, the BS need not utilize higher transmission powers since the same throughput of lower schemes can be achieved at poorer channel conditions at higher schemes.

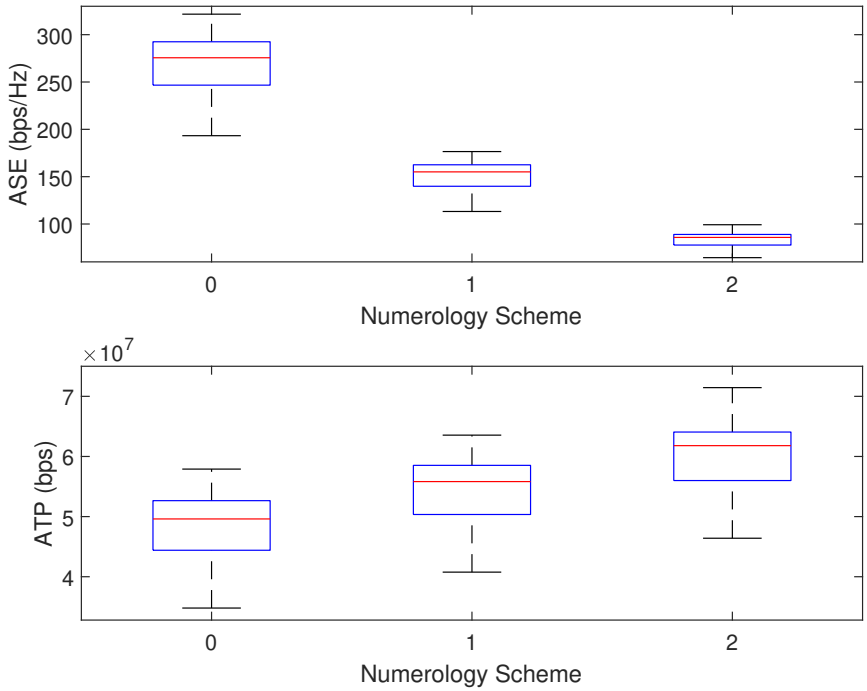

Figure 3.1 Downlink ASE and ATP of EMBB slice vs numerology scheme.

3. At the sub-6 GHz band, the network is limited to a maximum channel bandwidth of 100 MHz which does not allow for high throughputs as that of the mmWave band schemes (mmWave band allows for up to 400 MHz channel bandwidth without carrier aggregation or 800 MHz channel bandwidth with carrier aggregation).
4. The asymptotic behavior of throughput with respect to bandwidth at a given channel condition contributes to poor ASE and ATP.
5. There is more noise in a wider RB as dictated by the denominator of the SNR term in the Shannon Capacity theorem.

While spectral efficiency is generally considered to be an important metric, throughput is of a higher concern from the perspective of the UEs; thus, we also examine the downlink ATP of the EMBB slice in Figure 3.1. The corresponding average ATP of schemes 0, 1, and 2 were around 50 Mbps, 56 Mbps, and 62 Mbps, respectively. The maximum ATPs achieved were 58 Mbps, 63.5 Mbps, and 71.3 Mbps, correspondingly. The minimum ATPs of the respective schemes were 34.9 Mbps, 40.9 Mbps, and 46.5 Mbps. The ATP does not improve proportionally to the degradation

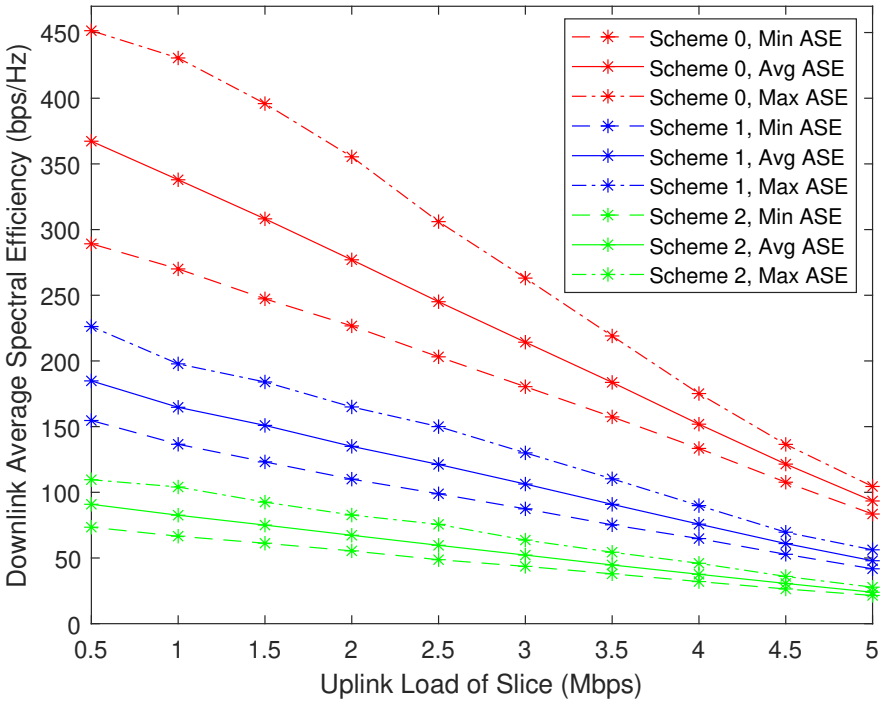

Figure 3.2 Downlink ASE of EMBB slice vs uplink load.

of the ASE, thus affording some credibility to the argument that the unimpressive ATP improvement accompanied by the massive ASE degradation does not justify the use of higher schemes (principle of diminishing returns).

Moreover, the ATP of each scheme has much intersection with that of the adjacent schemes, *i.e.*, they are not mutually exclusive. It is more practical to utilize lower numerology schemes since the lower schemes can achieve ATPs of the higher schemes for the most part. In this fashion, a reasonable ASE can be maintained while maintaining a *sweet spot* ATP. The highest numerology scheme achieves 75 percent of scheme 1 ATP; the same is true for scheme 1 ATP with respect to scheme 0 ATP. Impressively, scheme 0 ATP is 60 percent of scheme 2 ATP. Naturally, if networks can achieve high ASEs of lower schemes alongside high ATPs of higher schemes, this would be optimal for both UEs and network operators.

Next, we assess the impact of the EMBB uplink load on the ASE in Figure 3.2. Clearly, the ASE improvement with higher uplink loads is much higher at the lowest numerology scheme; this improvement becomes less significant at higher schemes. This can be explained by the fact that ASE behaves asymptotically without transmission power increase, *i.e.*, it is power-limited. Simply put, the increase in RB bandwidth does not translate to a proportional increase in throughput, hence the lackluster improvement (*i.e.* degradation) of ASE with higher schemes. Lastly, at the lowest uplink loads, the extremes of the ASE increasingly diverge from one another. From the UEs' perspectives, this can indicate that they may face a widely varying quality-of-experience (QoE) than they would at higher uplink loads. More importantly, however, is the fact that a higher uplink load would force the duplex ratio to decrease for the downlink communications. This would mean that at higher schemes, it may be harder to satisfy the downlink requirements since the symbol duration is even shorter. Given that this shorter duration will require a UE to utilize more RBs and RBs which are wider on the frequency domain, the ASE drops sharply as the scheme progressively becomes higher. This explains the downward ASE trend in Figure 3.2 as the scheme moves upwards.

3.5.2 URLLC slice performance

We now present and discuss the results of the URLLC slice in the mmWave band where latency is of greater importance. The total channel bandwidth available at the mmWave band is 400 MHz while the uplink throughput requirement was set to 1 Mbps [55]. The latency requirement was set to 0.5 ms [56]. The numerology schemes that will be utilized are all but the lowest two schemes as per 5G NR specifications (*i.e.*, schemes 2 to 5).

The UE latency and transmission power with respect to the downlink load are investigated for each scheme in Figure 3.3. As the downlink load increases, the uplink

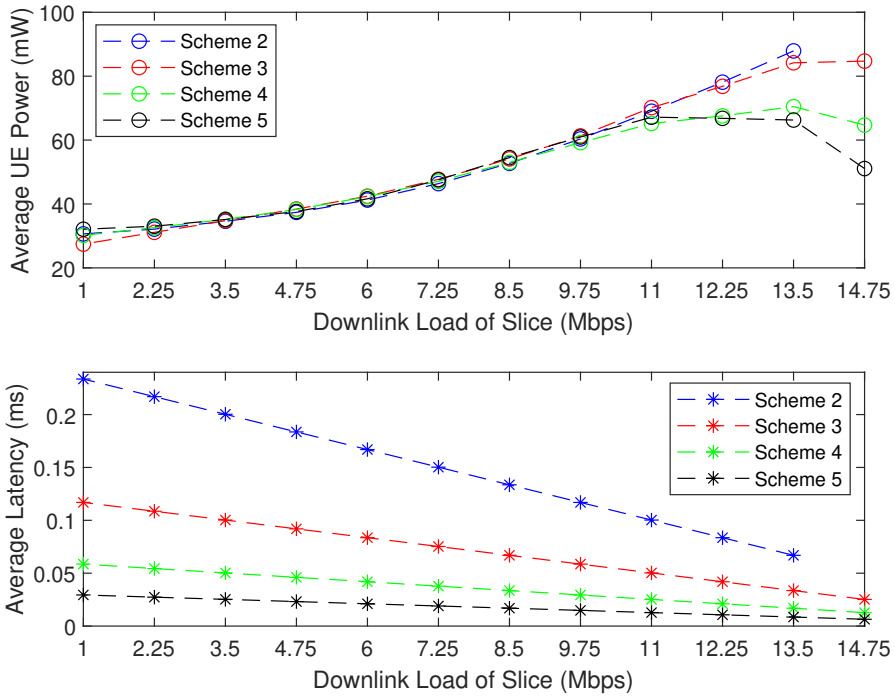

Figure 3.3 Average UE transmission power and latency vs downlink load.

transmission power generally increases. Although it seems counter intuitive, this is because more symbols are allocated to the downlink direction (higher downlink duplex ratio), thus forcing the UE to transmit at a higher power over the few remaining symbols allocated for the uplink direction. At lower downlink loads, the UEs can minimize their transmission power greatly due to the higher portion of the time slot allocated to uplink transmission. It is also clear that due to the large RB bandwidth and thus improved SNR, the highest scheme is able to minimize the uplink transmission power most effectively.

Figure 3.4 demonstrates how the UEs adjust their transmission power based on how much bandwidth is available in the lowest and highest possible schemes in the mmWave band. At minimal and full channel bandwidth access (100 MHz and 400 MHz), scheme 5 outperforms scheme 2. It also exhibits much better UE transmission power convergence as the downlink load decreases. This is due to the increased

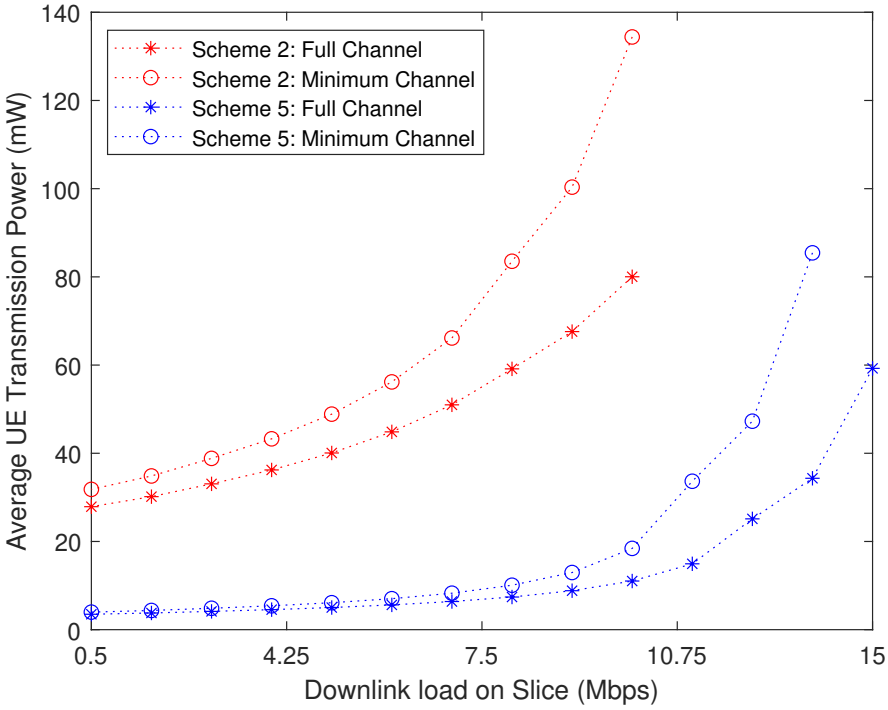

Figure 3.4 Average UE transmission power vs channel bandwidth and downlink load.

availability of uplink resources on the time domain (*i.e.*, via more symbols being allocated to uplink as a result of low downlink loads). Scheme 2 demonstrates much poorer performance in minimizing UE transmission power with increased downlink loads; there is a much wider gap between the UE transmission power at minimum channel bandwidth access (100 MHz) and maximum channel bandwidth access (400 MHz).

Throughout all the simulations for the URLLC slice, the UE throughput ranged from 1 to 1.06 Mbps. This is due to the power minimization objective; in order to minimize the power, the throughput must be throttled to the lowest possible level without violating the QoS. This especially allows the UE to reduce its transmission power when given access to more bandwidth, via higher schemes thus higher RB bandwidth or channel bandwidth access. If the UE has a minimal downlink load, it allows an even further reduction of power due to the increased time dedicated to

uplink transmission; all of the above have a compounding effect on UE transmission power. The only significant drawback is that this is spectrally very inefficient due to the very low throughput, high bandwidth usage, and temporal usage of resources. Nevertheless, from both the power and latency perspectives, scheme 5 appears to be the best choice for mission-critical applications since it demonstrates the best latency and uplink transmission power performance.

3.6 Summary

To summarize this chapter, we extensively study FDM of slices, each utilizing a unique numerology scheme, over both the sub-6 GHz and mmWave bands in a TDD RAN. Two optimization problems were formulated where each was concerned with a slice with respect to the prioritized direction of transmission and transmission band. The TDD parameters were then optimized to enhance ASE and minimize the UE transmission power for the EMBB and URLLC slices, respectively. It was shown, among many other observations, that the highest numerology schemes do not necessarily translate to the highest ASE or result in exclusively achievable ATPs. On the other hand, it was shown that the UE transmission power and latency drastically decrease with higher schemes. It was also demonstrated that at higher downlink loads, the highest numerology scheme effectively minimizes the transmission power and latency regardless of channel bandwidth.

CHAPTER 4

DUAL-BAND UAV NETWORKS FOR PRIORITY-BASED TRAFFIC

UAV networks typically suffer from lackluster performance due to line-of-sight issues as well as resource scarcity. Network slicing, multi-band transmission, and numerology show great potential in mitigating such limitations. In this chapter, we propose the use of the sub-6 GHz and mmWave bands, the latter of which requires careful consideration of line-of-sight, with numerology for aerial network slicing to provision time-critical services and broadband access. Accordingly, we formulate a user admission control policy to regulate band access after which we formulate a joint resource block and power allocation problem, an MINLP problem, to minimize the QoS gap of the users of the throughput-dependent service, which is considered to be best-effort traffic, and meet the time-critical service requirements. Most importantly however, we consider the channel condition on each RB individually during resource allocation through our low-complexity algorithm, PRiority BasED Resource AllocatIon in Adaptive SliCed NeTwork (**PREDICT**), which is proposed to tackle the formulated problem. In short:

1. We propose the deployment of a UAV equipped with dual-band transceivers for the fronthaul. The fronthaul utilizes both the sub-6 GHz and mmWave bands and supports all the 5G NR numerology schemes.
2. We design a channel and service-aware user admission control (UAC) policy to regulate the UEs' band access which is primarily dependent on their channel conditions, LoS, and service request.
3. We formulate a joint power and resource allocation problem, a MINLP problem, to minimize the provisioning gap of a best-effort throughput-dependent service while meeting the QoS requirements of a high-priority latency-sensitive service.

4. We propose our low-complexity algorithm, PREDICT, which allocates the sub-6 GHz and mmWave band resource blocks after the UAC policy is executed, to efficiently solve the MINLP problem. Most importantly though, the RB allocation depends on the unique channel gain and throughput per each individual RB as opposed to a single carrier frequency, thus accounting for the sensitive channel conditions at the mmWave region.

5. We discuss the extensive simulation results to validate our UAC policy and PREDICT; we benchmark the results against the LTE-based scenarios.

Before proceeding to the next section, we outline the organization of this chapter. In Section 4.1, we present the downlink network model whereas in Section 4.2, we formulate our UAC policy. In Section 4.3, we formulate our best-effort average QoS gap minimization problem. We propose our low-complexity PREDICT algorithm in Section 4.4 to efficiently solve the MINLP problem formulated in the preceding section. We present simulation results, discussions, and analyses to validate our approach in Section 4.5. Finally, we offer a brief summary of the problems addressed, the adopted approaches, and the corresponding results in Section 4.6.

4.1 System Model

In this particular section, we present our system model considering a UAV utilizing an OFDMA scheme. The UAV is equipped with sub-6 GHz and mmWave band transceivers for the fronthaul (the backhaul is not considered here) and supports all the NR numerology schemes, as shown in Figure 4.1. The UEs are categorized as either throughput-sensitive (EMBB) or latency-sensitive (URLLC); the former is a best-effort service while the latter is a high-priority service. The UAV will triage the two service types, meaning that the UAV must uphold full reliability for the URLLC slice at all times no matter the overall condition of the network. As a result, the EMBB slice may experience a QoS provisioning gap. The UAV will assign each UE to one of two transmission bands based on its requested service type and channel condition (this mechanism is detailed in Section 4.2). It is assumed that the number

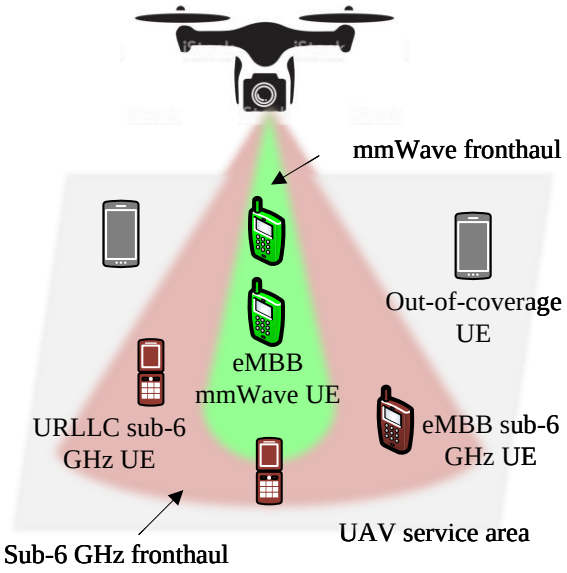

Figure 4.1 Dual-band numerology-enabled UAV network in a service area.

of RBs in the mmWave band is always greater than that of the sub-6 GHz band (the reasons as to why are clarified in Subsection 4.5.1).

4.1.1 Communication model

We now present our pathloss model which we utilize in conjunction with the communication model. The pathloss model is probabilistic and dependent on primarily two factors: Free Space Path Loss (FSPL) and probability of line-of-sight (PLoS). We first determine the LoS pathloss, $PL_{u,n}^{LoS}$, between the UAV and UE u on RB n whose carrier frequency is f_n [57]:

$$PL_{u,n}^{LoS}(dB) = 20\log\left(\frac{4\pi f_n d_u}{c}\right) + \eta_{LoS}, \tag{4.1}$$

where the first term is the FSPL and the second, η_{LoS}, is the additional average LoS link loss in dB. η_{LoS} depends on the environment where the network is situated, *i.e.*, rural, suburban, urban, etc. d_u is the distance (in meters) between the UAV and

UE u while c is the speed of light (meters per second). The non-line-of-sight (NLoS) pathloss can be determined as follows:

$$PL_{u,n}^{NLoS}(dB) = 20\log\left(\frac{4\pi f_n d_u}{c}\right) + \eta_{NLoS}, \tag{4.2}$$

where the second term, η_{NLoS}, is the additional average NLoS link loss (in dB) which also depends on the environment, *i.e.*, rural, suburban, urban, etc. For example, the average loss values for (η_{LoS}, η_{NLoS}) would be (0.1, 21), (1.0, 20), (1.6, 23), and (2.3, 34), for suburban, urban, dense urban, and high-rise urban environments, correspondingly [58]. Now that we have determined the path losses for the LoS and NLoS links, we need to determine the probabilities of occurrence for each type, PR_u^{LoS} and PR_u^{NLoS}, respectively [34]:

$$PR_u^{LoS} = \frac{1}{1 + a\exp(-b(\frac{180}{\pi}\theta_u - a))}, \tag{4.3}$$

$$PR_u^{NLoS} = 1 - PR_u^{LoS}, \tag{4.4}$$

where a and b are environmental constants while θ_u is the elevation angle (in radians) of the UAV with respect to UE u as shown in Figure 4.2. Therefore, the linearized LoS and NLoS channel gains can be determined from the logarithmic pathlosses (dB) respectively as follows:

$$g_{u,n}^{LoS} = 10^{-\frac{PL_{u,n}^{LoS}(dB)}{10}}, \tag{4.5}$$

$$g_{u,n}^{NLoS} = 10^{-\frac{PL_{u,n}^{NLoS}(dB)}{10}}. \tag{4.6}$$

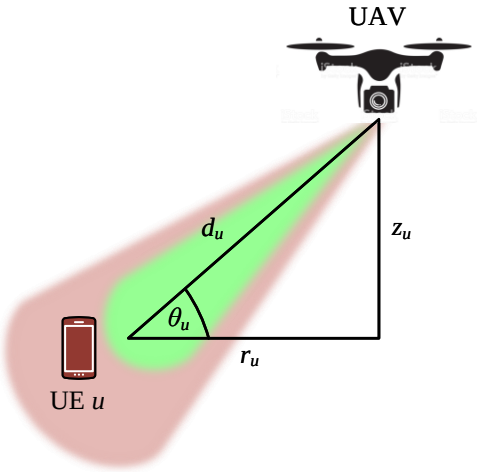

Figure 4.2 UAV location with respect to a UE.

The wireless throughput of the fronthaul for UE u over RB n, per the Shannon capacity is [59],

$$R_n^u = B^{RB} \log_2 \left(1 + \frac{g_{u,n}^{LoS} P_n^u}{B^{RB} N_0}\right)(PR_u^{LoS}) + B^{RB} \log_2 \left(1 + \frac{g_{u,n}^{NLoS} P_n^u}{B^{RB} N_0}\right)(PR_u^{NLoS}), \quad (4.7)$$

where B^{RB} is the RB bandwidth, P_n^u is the UAV transmit power on RB n for UE u, and N_0 is the noise spectral density. Therefore, the total data rate of UE u resultant of its allocated RBs would be:

$$R_u = \sum_{n=1}^{|\mathcal{N}^S|} a_n^u R_n^u, \quad (4.8)$$

where a_n^u is a binary indicator variable to represent if RB n is allocated to UE u and $\mathcal{N}^S$ is the set of RBs over the sub-6 GHz band. Accordingly, the latency of UE u is simply:

$$\tau_u = \frac{1}{R_u}. \quad (4.9)$$

Table 4.1 summarizes the notations utilized in this chapter.

Table 4.1 Summary of Notations for UAV Network

Notation	Definition
a_n^u	Binary variable for allocating RB n to UE u of EMBB.
b_m^v	Binary variable for allocating RB m to UE v of EMBB.
c_n^w	Binary variable for allocating RB n to UE w of URLLC.
D_E	Throughput requirement of EMBB.
J_v	QoS gap of UE v of EMBB over mmWave band.
P_{UAV}^S	Maximum UAV transmit power at the sub-6 GHz band.
P_{UAV}^M	Maximum UAV transmit power at the mmWave band.
P_n^u	UAV transmit power for UE u on RB n.
$P_{u,n}^S$	UAV transmit power for UE u of EMBB on RB n of sub-6 GHz.
$P_{v,m}^M$	UAV transmit power for UE v of EMBB on RB m of mmWave.
Q_u	QoS gap of UE u of EMBB over the sub-6 GHz band.
R_u	Achieved throughput of UE u over the sub-6 GHz band.
R_v	Achieved throughput of UE v over the mmWave band.
T	Air-interface bit latency requirement of URLLC slice.
τ_w	Latency of URLLC UE w.

4.2 User Admission Control Policy

We design our UAC policy in Algorithm 2 for the UEs to be assigned to either the sub-6 GHz or mmWave bands; this is executed prior to PREDICT. The policy is dependent on the service type and PLoS of the UE. The latter poses a significant challenge especially for the mmWave band [60, 61]. Since the URLLC slice does not generally require massive throughput, it does not need to employ mmWave links. As per Equations (4.1)-(4.2), it is clear that higher transmission frequencies lead to significantly higher path losses which do not bode well for sensitive traffic such as that of the URLLC slice. Moreover, a slight degradation in the mmWave LoS link will result in a much more pronounced deterioration of the channel gain than it would at the sub-6 GHz band. Therefore, to ensure maximum reliability and channel stability, our UAC policy places the URLLC slice exclusively on the sub-6 GHz band.

The EMBB slice does not have the same stringent requirements; hence, more liberties can be taken with its traffic as it is a best-effort slice. For each EMBB UE, if it is above a certain PLoS threshold, it is assigned to the mmWave band; otherwise,

it is associated with the sub-6 GHz band. The tuning of this PLoS threshold greatly impacts the overall resource allocation of the slices as the simulation results in Section 4.5 will demonstrate.

Algorithm 2: UAC Policy

Input: Unassociated EMBB and URLLC UEs

Output: UE-Band assignments

```
for all UEs do
    if UE requests URLLC service type then
        assign UE to sub-6 GHz band
    end
    else
        Calculate PLoS of UE
        if PLoS of UE ≤ PLoS Threshold then
            assign UE to sub-6 GHz band
        end
        else
            assign UE to mmWave band
        end
    end
end
```

4.3 Best-Effort Average QoS Gap Minimization

After the UAC policy is executed, the UAV must allocate bandwidth and transmission power to the users as per their priority. Thus, we formulate an average QoS gap (throughput gap in the context of EMBB traffic) minimization problem, which is a joint power and RB allocation problem, for the EMBB slice (best-effort). The URLLC slice is always guaranteed its QoS constraints.

We denote the sets of RBs for the sub-6 GHz and mmWave bands as $\mathcal{N}^S$ and $\mathcal{N}^M$, respectively. The superscripts S and M serve to identify which band the RB set

is from: S is for the sub-6 GHz band and M is for the mmWave band. Furthermore, we denote the set of EMBB UEs on the sub-6 GHz and mmWave bands as $\mathcal{U}_E^S$ and $\mathcal{U}_E^M$, respectively. We define the following decision variables: $\boldsymbol{A} = \{a_n^u\}$, $\boldsymbol{B} = \{b_n^v\}$, and $\boldsymbol{C} = \{c_n^w\}$ as RB allocation binary indicators. $\boldsymbol{P^S} = \{P_{u,n}^S\}$ and $\boldsymbol{P^M} = \{P_{v,m}^M\}$ are the power allocation variables on RB n for UE u of the sub-6 GHz band and RB m of UE v on the mmWave band, respectively. Q_u represents the QoS gap of EMBB UE u of the sub-6 GHz band while J_v represents the same for the mmWave EMBB UEs. They are written as decision variables in matrix form, $\boldsymbol{J}$ and $\boldsymbol{Q}$.

Prior to defining the problem, we need to explicitly define the QoS gap of each UE mathematically. In an ideal scenario where the network can always accommodate the needs of all its UEs, the UEs would be meeting and even exceeding their QoS requirements, and thus no QoS gap would exist. However, in a heavily-loaded network, this is highly improbable to achieve and as such, the resource allocation process will have to prioritize certain UEs over others. It follows that the best-effort UEs may suffer from a QoS gap which, for the EMBB UEs, would be the difference between their required and actual throughputs. Therefore, if D_E is the required EMBB slice throughput, the QoS gaps on the sub-6 GHz and mmWave bands can be defined as:

$$Q_u = D_E - R_u, \forall u \in \mathcal{U}_E^S, \tag{4.10}$$

$$J_v = D_E - R_v, \forall v \in \mathcal{U}_E^M. \tag{4.11}$$

Subsequently, we can present the problem as follows:

$$\textbf{P4:} \quad \min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{J},\boldsymbol{P}^S,\boldsymbol{P}^M,\boldsymbol{Q}} \frac{1}{|\mathcal{U}_E^S| + |\mathcal{U}_E^M|} \left(\sum_{u=1}^{|\mathcal{U}_E^S|} Q_u + \sum_{v=1}^{|\mathcal{U}_E^M|} J_v \right) \tag{4.12}$$

$$\text{s.t. } P^S_{u,n} \leq \frac{P^S_{UAV}}{|\mathcal{N}^S|}, \forall u \in \mathcal{U}^S, \ \forall n \in \mathcal{N}^S, \tag{4.13}$$

$$P^M_{v,m} \leq \frac{P^M_{UAV}}{|\mathcal{N}^M|}, \forall v \in \mathcal{U}^M, \ \forall m \in \mathcal{N}^M, \tag{4.14}$$

$$\sum_{u=1}^{|\mathcal{U}^S_E|} \sum_{n=1}^{|\mathcal{N}^S|} a^u_n + \sum_{w=1}^{|\mathcal{U}_U|} \sum_{n=1}^{|\mathcal{N}^S|} c^w_n \leq |\mathcal{N}^S|, \tag{4.15}$$

$$\sum_{v=1}^{|\mathcal{U}^M_E|} \sum_{m=1}^{|\mathcal{N}^M|} b^v_m \leq |\mathcal{N}^M|, \tag{4.16}$$

$$\sum_{u=1}^{|\mathcal{U}^S_E|} a^u_n + \sum_{w=1}^{|\mathcal{U}_U|} c^w_n \leq 1, \forall n \in \mathcal{N}^S, \tag{4.17}$$

$$\sum_{v=1}^{|\mathcal{U}^M_E|} b^v_m \leq 1, \forall m \in \mathcal{N}^M, \tag{4.18}$$

$$a^u_n \in \{0,1\}, \forall u \in \mathcal{U}^S_E, \forall n \in \mathcal{N}^S, \tag{4.19}$$

$$b^v_m \in \{0,1\}, \forall v \in \mathcal{U}^M_E, \forall m \in \mathcal{N}^M, \tag{4.20}$$

$$c^w_n \in \{0,1\}, \forall w \in \mathcal{U}_U, \forall n \in \mathcal{N}^S, \tag{4.21}$$

$$\tau_w \leq T, \forall w \in \mathcal{U}_U. \tag{4.22}$$

In Equations (4.13)-(4.14), we enforce the power constraint per band where the maximum transmit power of the UAV on the sub-6 GHz and mmWave bands is denoted by P^S_{UAV} and P^M_{UAV}, respectively. Through Equations (4.15)-(4.16), we ensure that, in each band, the allocated bandwidth cannot exceed the total channel bandwidth available (total number of RBs available). The OFDMA constraint per band is dictated in Equations (4.17)-(4.18). We enforce the binary nature of the

indicators in Equations (4.19)-(4.21) where a_n^u denotes if RB n of the sub-6 GHz band RBs, $\mathcal{N}^S$, is allocated to UE u of the EMBB slice, $\mathcal{U}_E^S$. b_m^v denotes if RB m of the mmWave band RBs, $\mathcal{N}^M$, is allocated to user v of the EMBB slice, $\mathcal{U}_E^M$. Lastly, c_n^w denotes if RB n of the sub-6 GHz band, $\mathcal{N}^S$, is allocated to user w of the URLLC slice, $\mathcal{U}_U$. These are all written in matrix form in the objective function. Finally, we ensure that the per-bit latency of UE w of the URLLC slice is below the deadline, T, in Equation (4.22).

Note that the URLLC UEs in this formulation, denoted by the set $\mathcal{U}_U$, are always granted the slice's minimum required QoS. The URLLC service is the most stringent of all, and thus does not tolerate any QoS gaps (hence the lack of any URLLC term in the objective function). Moreover, since this slice always utilizes the sub-6 GHz band, there is no need for a superscript S (unlike the EMBB slice, which can utilize both the sub-6 GHz and mmWave bands, that has a band indicator in the superscript, accordingly).

Regarding the primary decision variables $\mathbf{J}$ and $\mathbf{Q}$, there are three possible cases and respective implications to consider:

1. Q_u (or J_v) is positive: indicating that a user's throughput is below D_E,
2. Q_u (or J_v) is zero: indicating that a user's throughput is equal to D_E, and
3. Q_u (or J_v) is negative: indicating that a user's throughput is above D_E.

Minimizing the average QoS gap implies the maximization of its negative value which is equivalent to minimizing its positive value. To minimize its positive value, the network will seek to maximize the average throughput of the UEs. We can then conclude that minimizing the average QoS gap of the network is equivalent to, both in meaning and mathematically, maximizing the average throughput of the network. Accordingly, we use the phrases average throughput maximization and average QoS gap (or degradation) minimization, interchangeably.

4.4 Dual-Band Resource Allocation Policy: PREDICT

To solve **P4**, we propose the PRiority BasED Resource AllocatIon in Adaptive SliCed NeTwork (**PREDICT**) algorithm, which works in the following fashion: first, it calculates the channel gain of each UE on each band's RBs (Line 1). Subsequently, as per Line 2, the UAV allocates its maximum transmission power on each of the RBs as dictated by Equations (4.13)-(4.14) so that each UE's data rate per RB can be calculated in Line 3 via Equation (4.7). The UEs of all slices on all the bands are then sorted from the best channel conditions to the worst (Line 4). The URLLC UEs are provisioned RBs first due to their high-priority status. For each URLLC UE, the RBs are sorted from the lowest-throughput RBs to the highest-throughput RB (Line 5). A URLLC UE is allocated RBs sequentially from the lowest-throughput RB to the highest-throughput RB in that order until its latency requirements are met (Line 8). These allocated RBs are removed from the sub-6 GHz RB set (Line 9). This is repeated for each URLLC UE until they are all satisfied (Lines 6-12).

Next, the EMBB UEs on both bands are assigned to RBs; arbitrarily, we start with the sub-6 GHz band. For each EMBB UE, the RBs are sorted from the highest-throughput RBs to the lowest-throughput RBs (Line 12). Each EMBB UE is allocated RBs sequentially from the lowest-throughput RBs to the highest-throughput RBs in that order, until its minimum throughput is met or there are no RBs left on that band; the allocated RBs are removed from the sub-6 GHz RB set (Lines 13-18). The same process is repeated for the mmWave band EMBB UEs. Finally, as for the surplus RB allocation for the EMBB UEs in Lines 19-22, if there are any remaining RBs which have yet to be allocated, on either band, they are to be allocated to the single UE that achieves the highest throughput on those RBs in order to minimize the objective function in Equation (4.12).

Now that we have explained PREDICT, we can present its complexity analysis below.

Algorithm 3: PREDICT Algorithm

Input: UE band associations, network parameters

Output: UE RB assignments, QoS degradation

Calculate LoS and NLoS channel gains (Equations (4.5)-(4.6))

Allocate maximum transmit power to each RB (Equations (4.13)-(4.14))

Calculate the throughput per UE on each RB (Equation (4.7))

Sequence the UEs from the best channel gains to the worst

Sub-6 GHz RB allocation for URLLC UEs:

For each UE, sequence RBs from lowest to highest throughput

for $w = 1$ *to* $|\mathcal{U}_U|$ **do**

  while *UE* w *is not satisfied* **do**

    assign RBs from lowest to highest throughput per RB consecutively until UE latency satisfies Equation (4.22)

    Set $c_m^w = 1$ and remove corresponding RBs from $\mathcal{N}^S$

  end

end

Baseline sub-6 GHz and mmWave RB allocation for EMBB UEs:

For each UE, sequence RBs from lowest throughput to highest

for $u = 1$ *to* $|\mathcal{U}_E^S|$ **do**

  while *RBs are available && UE* u *is not satisfied* **do**

    assign RBs from lowest to highest throughput per RB consecutively (Equations (4.17)-(4.18)) until UE u's data rate satisfies D_E.

    Set a_n^u and $b_m^v = 1$ and remove corresponding RBs from $\mathcal{N}^S$ and $\mathcal{N}^M$

  end

end

Surplus sub-6 GHz and mmWave RB allocation for EMBB UEs:

while *surplus RBs available && all EMBB UEs are satisfied* **do**

  Allocate remaining RBs of $\mathcal{N}^S$ and $\mathcal{N}^M$ to UE with highest throughput per RB

  Set a_n^u and $b_m^v = 1$ and remove corresponding RBs from $\mathcal{N}^S$ and $\mathcal{N}^M$

end

- The complexities of sorting $|\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}|$, $|\mathcal{U}_{\mathcal{E}}^{\mathcal{M}}|$, and $|\mathcal{U}_U|$ UEs are $\mathcal{O}(|\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}| \log |\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}|)$, $\mathcal{O}(|\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}| \log |\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}|)$, and $\mathcal{O}(|\mathcal{U}_U| \log |\mathcal{U}_U|)$, respectively (Line 4).
- The complexities of sorting $|\mathcal{N}^S|$ RBs of the sub-6 GHz band for the URLLC UEs is $\mathcal{O}(|\mathcal{N}^S| \log |\mathcal{N}^S|)$ (Line 5).
- *Sub-6 GHz RB allocation for URLLC slice*: The complexity of allocating the sub-6 GHz RBs, $|\mathcal{N}^S|$, to the URLLC UEs is $\mathcal{O}(|\mathcal{N}^S|(|\mathcal{N}^S| - 1)/2)$ which can be simplified to $\mathcal{O}(|\mathcal{N}^S|^2)$ (Lines 6-11).
- The complexities of sorting the remaining $|\mathcal{N}^S - |\mathcal{U}_U||$ sub-6 GHz RBs and $|\mathcal{N}^M|$ mmWave RBs for the EMBB UEs are $\mathcal{O}((|\mathcal{N}^S| - |\mathcal{U}_U|) \log\big(|\mathcal{N}^S| - |\mathcal{U}_U|\big))$ and $\mathcal{O}(|\mathcal{N}^M| \log |\mathcal{N}^S|)$, respectively (Line 13).
- *Baseline RB allocation for EMBB slice on both bands*: Similar to that of Lines 6-11, the complexities of allocating $(|\mathcal{N}^S| - |\mathcal{U}_U|)$ remaining sub-6 GHz band RBs and $|\mathcal{N}^M|$ remaining mmWave band RBs to the EMBB UEs are $\mathcal{O}(|\mathcal{N}^S| - |\mathcal{U}_U|)^2$ and $\mathcal{O}(|\mathcal{N}^M|^2)$, respectively (Lines 14-20).

The overall complexity of PREDICT algorithm can be written as: $\mathcal{O}(|\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}| \log |\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}| + |\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}| \log |\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}| + |\mathcal{U}_U| \log |\mathcal{U}_U| + |\mathcal{N}^S| \log |\mathcal{N}^S| + |\mathcal{N}^S|^2 + (|\mathcal{N}^S| - |\mathcal{U}_U|) \log\big(|\mathcal{N}^S| - |\mathcal{U}_U|\big) + |\mathcal{N}^M| \log |\mathcal{N}^S| + (|\mathcal{N}^S| - |\mathcal{U}_U|)^2 + |\mathcal{N}^M|^2$. As $|\mathcal{N}^S| \geq |\mathcal{U}_{\mathcal{E}}^{\mathcal{S}}|$, $|\mathcal{N}^S| \geq |\mathcal{U}_{\mathcal{E}}^{\mathcal{M}}|$, $|\mathcal{N}^S| \geq |\mathcal{U}_U|$, and $|\mathcal{N}^M| > |\mathcal{N}^S|$, we can then simplify the overall complexity of PREDICT and write it as: $\mathcal{O}(|\mathcal{N}^M| \log |\mathcal{N}^M| + |\mathcal{N}^M|^2)$. We can then further simplify the complexity as it tends to $\mathcal{O}(|\mathcal{N}^M|^2)$.

4.5 Simulation Results

We now present an in-depth discussion of our simulation results. Table 4.2 summarizes the fixed simulation parameters utilized for this work. All the generated data points in the simulation figures are averaged over 500 Monte Carlo simulations. We investigate the performance of our network under five different cases:

1. Varying minimum required EMBB throughput and probability threshold,
2. Varying EMBB user load, required throughput, and probability threshold,
3. Varying the total channel bandwidth,

4. Varying numerology schemes of both bands, and

5. Legacy vs dual-band numerology-enabled aerial network performance comparison.

Where present in the figures, the satisfaction rate in the context of the EMBB slice is defined as the percentage of UEs that meet their minimum QoS requirement. Finally, the term PR represents the threshold probability of the mmWave band admission in the UAC policy.

Table 4.2 Simulation Parameters for UAV Network

Parameter	Value
EMBB users	40
URLLC users	25
URLLC air-interface per-bit latency	0.5 ms/bit [62]
Sub-6 GHz carrier frequency	2.4 GHz
mmWave carrier frequency	25 GHz
Sub-6 GHz channel bandwidth	20 MHz
mmWave channel bandwidth	80 MHz
Maximum UAV transmission power	40 dBm
Noise spectral density	-174 dBm/Hz
sub-6 GHz numerology scheme	2
mmWave numerology scheme	5
Environmental constants (a, b)	9.61, 0.16 [34]
Average LoS and NLoS attenuation	1, 20 dB [34]
Hotspot size	60m $\times$ 60m

4.5.1 Varying required EMBB throughput

In Figure 4.3, we investigate the EMBB slice's performance with varying probability thresholds and throughput requirements. Here, we bring to the reader's attention that since the maximum bandwidth allowed in 4G-LTE networks is 20 MHz only, we consider this amount as the minimum bandwidth for the sub-6 GHz band in our 5G UAV network. Furthermore, because the maximum bandwidth available for 5G networks at the sub-6 GHz band is 100 MHz while that of the mmWave band is 400 MHz, the 5G aerial networks in our simulation scenarios always maintain a

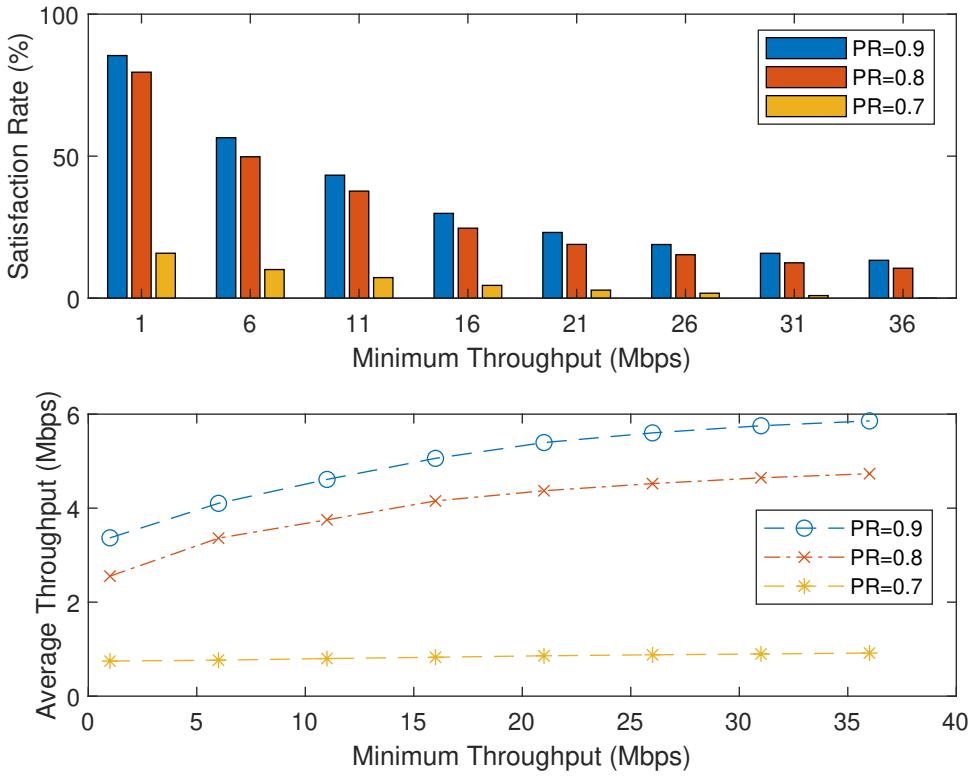

Figure 4.3 EMBB performance vs throughput requirement.

1-to-4 channel bandwidth ratio between said bands; and hence, we only explicitly mention the sub-6 GHz bandwidth in the results since it is implied that the mmWave bandwidth is four times that amount by default.

Figure 4.3 makes clear that lowering the PLoS threshold in the UAC policy incurs higher QoS gaps, and thus lower satisfaction rates, for the EMBB slice overall. In the case of the EMBB slice, it is much higher only when the users with near-perfect PLoSs are admitted into the mmWave band. Lowering this threshold leads to a higher dissatisfaction of the slice due to an increasing number of users with poorer channel conditions being admitted into the mmWave band, thus making it more difficult for the UAV to satisfy their requirements. Even a slight decrease in the PLoS threshold results in a massive dissatisfaction of the slice. Users that would have been better served by the sub-6 GHz band are now instead assigned to the mmWave band. We also see that increasing the minimum required throughput of the EMBB slice further

strains the UAV network. The pathloss at the mmWave band is much higher due to the high carrier frequencies of the mmWave RBs. Consequently, a slight decrease in LoS conditions at the mmWave band has a severe negative impact on the EMBB slice's performance.

Counter-intuitively, while the satisfaction rate worsens at higher requirements, the average throughput improves. The average throughput is actually driven by the UEs that can meet the higher throughput requirements; the higher the throughput that they can achieve, the higher the average throughput that is obtained. However, the average throughput curve eventually levels off because the UAV will exhaust all its resources and no longer be able to further increase any UE's throughput. The results for lower PLoS thresholds (below 0.7) are not shown since the network performance degrades too far.

4.5.2 Varying EMBB user load

In Figure 4.4, we assess the performance of the network under increased user loads at various throughput requirements, specifically at minimum requirements of 5, 10, and 15 Mbps. We see that as the minimum throughput requirement increases, even at a very high PLoS threshold, the network struggles to satisfy many of the users of the slice. The satisfaction of the slice suffers when the number of users in the EMBB slice increases; this is due to swift bandwidth exhaustion. This bandwidth exhaustion is brought on even faster at lower PLoS thresholds because more RBs are required to compensate for the lack of a strong LoS channel in the mmWave band. This further highlights the sensitivity of the mmWave band to LoS communication channels and underscores the need for a very stringent UAC policy.

4.5.3 Varying channel bandwidth

In Figure 4.5, we examine how the network responds to bandwidth availability. We investigate the performance for a 15 Mbps minimum EMBB throughput. The addition

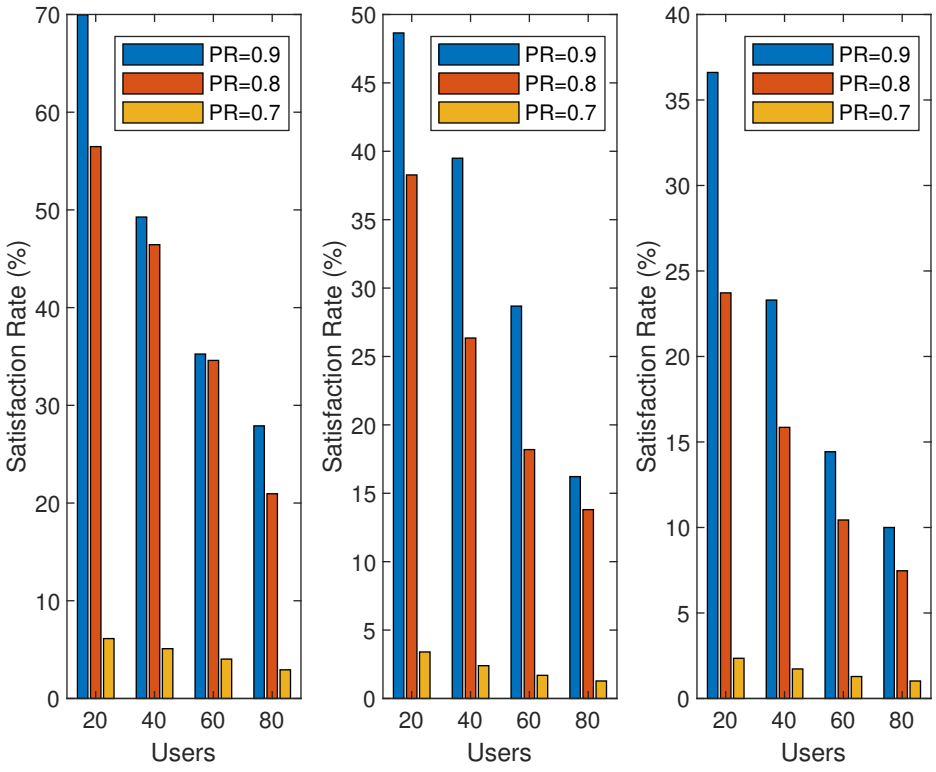

Figure 4.4 EMBB performance vs user load and throughput requirements of 5 Mbps (left), 10 Mbps (center), and 15 Mbps (right).

of bandwidth affords more RBs for allocation which increases the satisfaction and average throughput of the slice. However, at a high enough amount of bandwidth, the satisfaction rate will level off because with the addition of more RBs, the transmission power per RB will gradually decrease so much so that the throughput on each RB will be too little for it to be able to satisfy any UE. Therefore, increasing the channel bandwidth is not always favorable, especially for a UAV network (which already has limited transmission power). To offset the degraded performance at excessive channel bandwidths caused by the noise power, the UAV should be endowed with a higher transmission power to be able to allocate it over the added RBs without being spread out too thin. This is projected to be possible in the near future with the latest advancements in antenna design and UAV-related hardware.

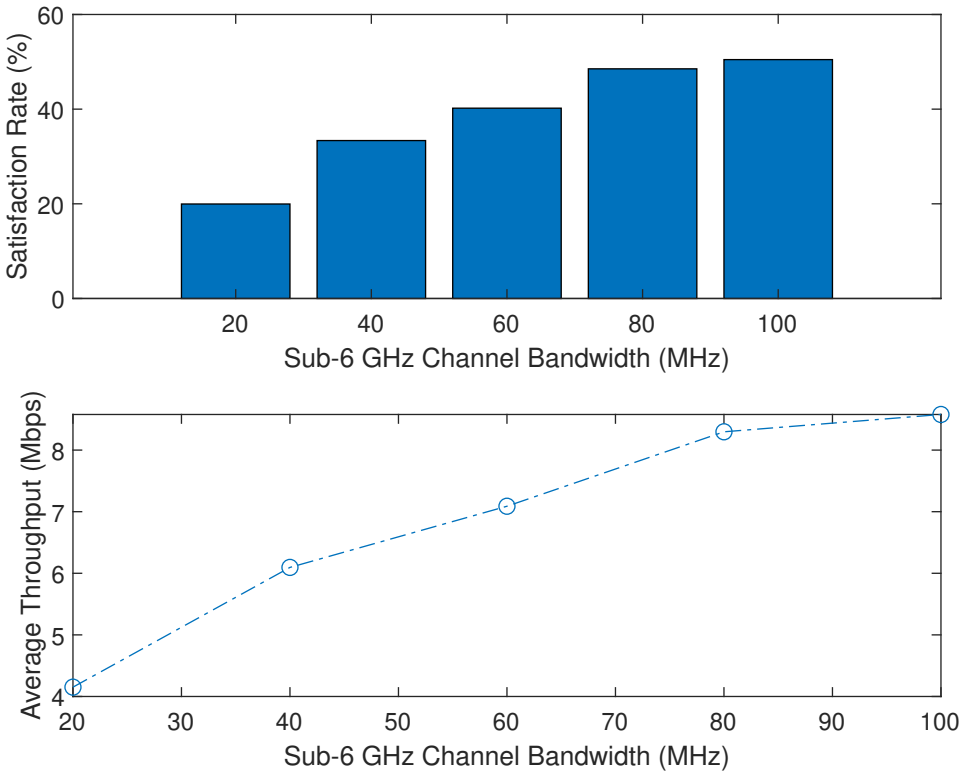

Figure 4.5 EMBB performance vs channel bandwidth. The mmWave bandwidth (not shown explicitly) is four times that of the sub-6 GHz band and increases proportionally.

4.5.4 Varying numerology schemes

In Figure 4.6, we investigate the EMBB slice's performance for varying numerology schemes for a 10 Mbps minimum requirement. Surprisingly, it is shown that the lowest pair of numerology schemes outperform the highest. It is deducible that the higher numerology schemes, while increasing the RB bandwidth, also increase the associated noise of the RB. Furthermore, at lower numerology schemes, there are more RBs available for allocation to the UEs; and hence, it is easier to satisfy more users. Higher schemes generally mean fewer RBs available for allocation. Furthermore, those fewer available RBs have a much higher noise factor; without additional transmission power to overcome the noise, the RB's SNR degrades, thus lowering its throughput. Therefore, the EMBB slice is even more sensitive to the PLoS thresholds at higher

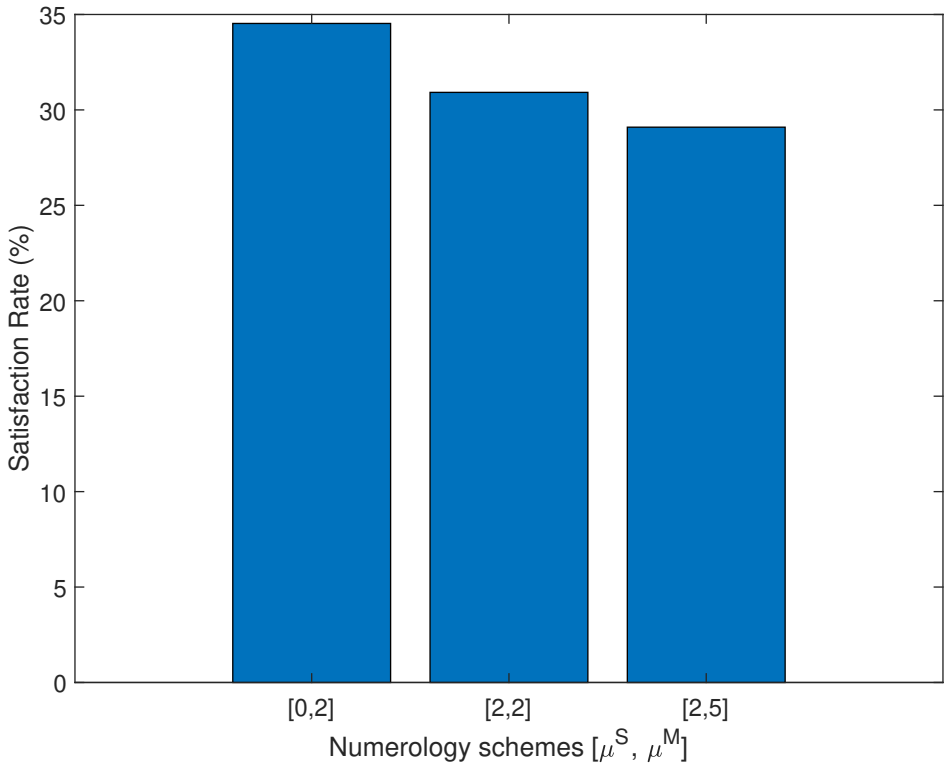

Figure 4.6 EMBB performance vs numerology scheme.

schemes. In other words, at the higher numerology schemes, there is a much lower tolerance for poor LoS conditions for users.

4.5.5 URLLC performance vs EMBB user load

In Figure 4.7, we investigate the URLLC slice's performance for varying numerology schemes for a 10 Mbps minimum EMBB throughput. We see that despite the increasing load on the EMBB slice, the URLLC slice is able to maintain its requirements comfortably. The worst case latency is slightly greater than 5 μs. Therefore, we have shown the resilience of the high-priority URLLC slice to poor performance despite the increasing load and/or poor channel conditions of the other heavily-loaded slices in the network. This is important because the URLLC slice is

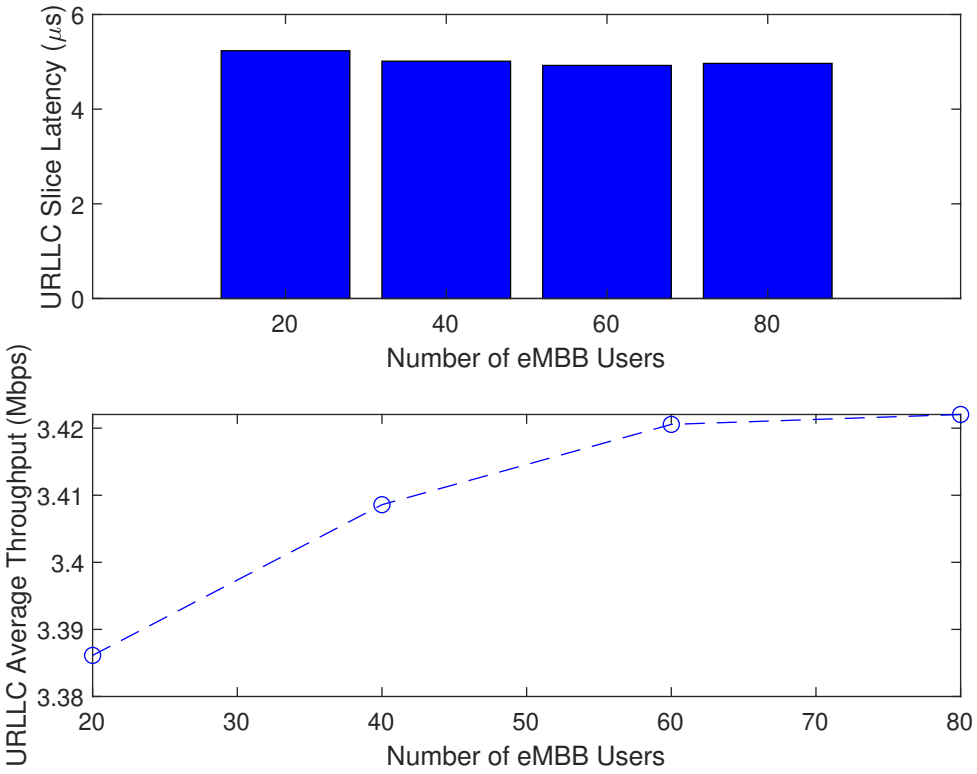

Figure 4.7 URLLC performance vs EMBB user load.

time-critical, has the highest priority in the network as per our system model, and requires the most effective slicing isolation.

4.5.6 Benchmarking PREDICT

In this section, we compare the performance of the legacy LTE scheme with PREDICT to validate the latter's advantages. Before doing so, we must note two inherent major differences between the two: firstly, in the legacy scheme, there is no need for a UAC policy since there is only a single transmission band (sub-6 GHz). Secondly, the RBs have a static bandwidth of 180 kHz only (which is equivalent to numerology scheme 0 of 5G networks) whereas in PREDICT, the RBs have wider bandwidths due to their numerology schemes. Along with other differences which are out of the scope of this work, these will impact and inform our analyses of the benchmark.

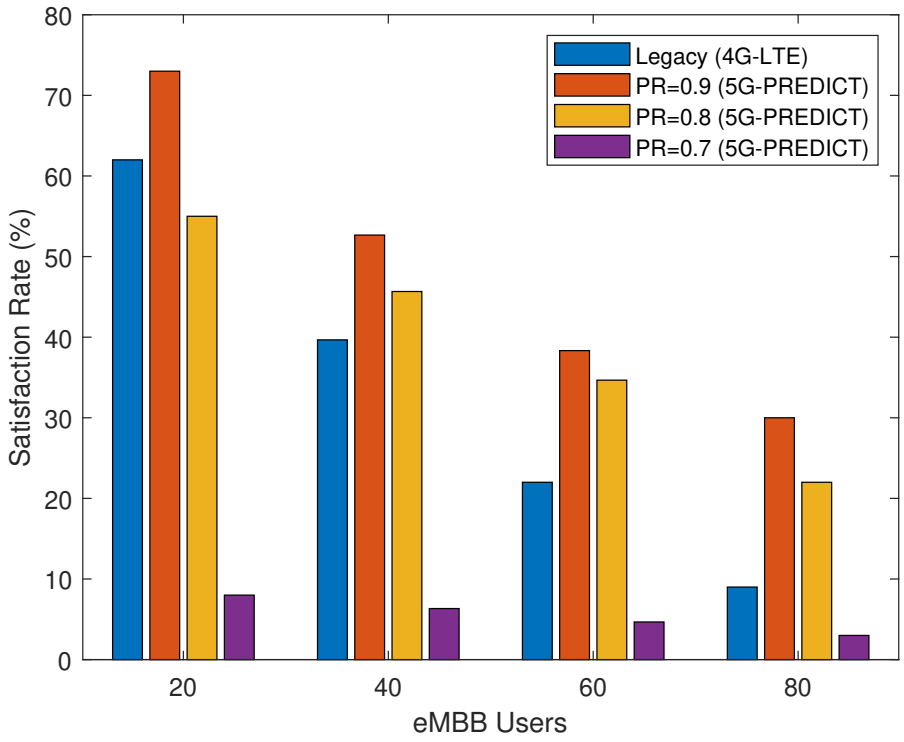

Figure 4.8 Legacy LTE and PREDICT performances vs EMBB user load.

We now outline the simulation settings for this benchmark. The sub-6 GHz channel bandwidth for both the legacy LTE scheme and PREDICT is set to 20 MHz; the mmWave bandwidth in PREDICT is set to 80 MHz. Next, in accordance to the performance observed in Figure 6, we set the numerology schemes under PREDICT to $\mu_S = 0, \mu_M = 2$. The legacy LTE network is set to scheme 0. The data rate of the EMBB slice is set to 5 Mbps; the URLLC slice latency requirement is the same as in the previous scenarios (0.5 ms/bit) and its load is fixed to 25 users. As for the EMBB slice load, we vary it to demonstrate the performance difference between the legacy LTE scheme and PREDICT.

In Figure 4.8, it is observed that the legacy scheme performs nearly identically as PREDICT at a minimum EMBB load; since both intra-slice (users within the same slice) and inter-slice (between different slices) contentions are negligible, both the legacy scheme and PREDICT have similar performances. However, there is indeed

a slight performance lag of the legacy scheme at lower loads due to the static RB bandwidth (PREDICT allocates RBs of much higher bandwidths affording higher SNR and resilience to poorer LoS). We notice that the performance advantages of PREDICT under a stringent UAC policy, *i.e.*, LoS threshold $\geq$ 0.8, become more pronounced as the EMBB slice load increases. This is explicitly due to the increasing contention, smaller RB bandwidth, and smaller channel bandwidth of the LTE network. Additionally, PREDICT assigns users to the mmWave band and thus is better at alleviating contention whereas the legacy scheme has only the sub-6 GHz band to work with and is limited to 20 MHz.

Although PREDICT's performance does degrade as the network load increases, it always outperforms the legacy scheme with the exception of when the LoS threshold $\leq$ 0.7. Recall that as the LoS threshold decreases, more users with poorer LoS conditions are admitted into the mmWave band. The mmWave band is highly sensitive to blockage especially in urban settings (η_{NLoS}=20 dB [34]). The throughput obtained on each RB for a user is extremely low and it makes it difficult for PREDICT to allocate enough RBs to meet its requirement thus allowing the legacy scheme to pull ahead. This underscores the great care that needs to be taken in designing a UAC policy when dual-band numerology-based UAVs are deployed. Essentially, there is a delicate balance between dual-band transmission, channel bandwidth and conditions, and numerology schemes in 5G aerial networks.

In traditional UAV communication schemes, a single wireless center frequency is assumed (around 2.4 GHz) for the channel modeling. In other words, the throughput for each RB on the sub-6 GHz band is calculated out to be identical and it only becomes a question of how many RBs (for integral constraints) or how much bandwidth (for continuous constraints) should be assigned to a UE. While this may be acceptable for approximate models in the sub-6 GHz region, this assumption significantly degrades the network performance in the mmWave region. This is

because even a slight increase in mmWave channel frequency massively worsens the resulting channel condition; and hence, the throughput on that RB. Moreover, up to 400 MHz bandwidth can be utilized without carrier aggregation at the mmWave region. Over this very large channel bandwidth, calculating a channel gain on one center frequency and applying the resulting throughput to the entire set of RBs will not work; the throughput of RBs at opposite ends of the channel bandwidth will have much disparity. Our proposed scheme adjusts for all such complications by calculating the channel condition and throughput on every RB on each band for each slice prior to determining the band-UE associations and transmit power allocations. This is one of catalysts behind PREDICT's superior performance.

Looking ahead to 6G networks which envision employing THz frequencies, such considerations will be even more imperative when modeling aerial-to-ground communication channels; PREDICT lays the groundwork for a more realistic implementation going ahead while incorporating numerology. Comparatively speaking, aerial-to-ground channels are much more susceptible to blockage than are ground-to-ground channels. Numerology does not negatively impact the resource allocation for ground-to-ground channels as much in mmWave bands, but this cannot be said for aerial networks for both mmWave and THz bands; our results have made that amply clear. We would further posit that aerial networks are more tolerant of sub-optimal UAV placements when not only utilizing dual-band schemes but also the 5G numerology schemes because the higher RB bandwidths can compensate for lower channel conditions to a certain extent. Furthermore, they also enable shorter time slots vital for URLLC use cases.

4.6 Summary

We have combined network slicing and the novel numerology schemes within the sub-6 GHz and mmWave transmission spectra to demonstrate how a UAV must adopt a

UAC policy and triage differing priorities of services to optimally allocate network resources. Furthermore, we observed the direct impact of PLoS on a best-effort service's satisfaction and throughput. We efficiently solved the proposed MINLP problem through our PREDICT algorithm to minimize the average QoS gap of the eMBB slice while maintaining full reliability of the URLLC slice. The results show that high network satisfaction is achievable with very stringent PLoS requirements even under strenuous conditions with the proper selection of the numerology scheme and transmission band. Furthermore, our proposed algorithm demonstrates superior performance against the conventional UAV scheme. Throughout our discussions of the network performance, whether it be from the perspective of throughput, latency, or satisfaction rate, it becomes clear that our scheme strategically exploits the flexibility in RB bandwidth, mmWave band, as well as the increased channel bandwidth in that region, to mitigate the LoS challenges which are dominant in aerial-to-ground communication systems.

CHAPTER 5

NEXT-GENERATION CORE NETWORK CONFIGURATION

Although 5G networks are still yet to be fully standardized, it is well-accepted by 3GPP that the 6G core network (CN) will be a complete overhaul of its predecessor. It is projected to exploit a system level artificial intelligence (AI) framework known as AI-Native which will oversee all aspects of management, orchestration, and operation. All CN NFs, tasks, and services will be executed autonomously. Currently as it stands, AI is designed as an external NF which is an after-the-fact service. However, in complete contrast, AI will be *the* system in 6G CNs. AI-Native will naturally have access to all NF data, performance metrics, statistics, and alerts to inform its decision making which will undoubtedly impact NF execution, resource allocation, network prediction, security, service restoration and redundancy. The specifications of AI-Native are still evolving [63]; the standardization bodies have yet to reach that milestone. Nevertheless, there have been preliminary investigations into integrating AI at a level expected to be conducive to AI-Native (also known as *Native-AI* in some literature) within the radio access network (RAN) [64, 65] and CN [66].

Optimal performance of the CN, which is split into the CP and UP, cannot be overstated. The CP consists of numerous vital NFs which have to conduct significant control signaling and specialized tasks in an extremely short period of time to ensure that E2E QoS requirements are met. Otherwise, QoS flows and user sessions cannot be established in time, thus hampering the UP routing from the RAN to the CN and then external networks. Consequently, it can even be argued that the CP latency constraints are perhaps among the most important and stringent constraints.

In this chapter, we propose slicing both the CP and UP while enforcing planar physical isolation to ensure maximum security, reliability, and slicing isolation. Under this configuration (and others), we exploit an AI-framework to minimize the CP

latency with the ulterior aim of satisfying E2E 6G slicing requirements. This area of research is extremely under explored especially considering that 3GPP envisions a complete overhaul of the CN in light of the ever so stringent QoS requirements of expected 6G services and applications and newly proposed NFs if any in future 3GPP standards and/or releases. Hence, our contributions are:

- We exploit knowledge of the traffic characteristics (from referenced virtual implementations of the CN) between the NFs from both CN planes and feed NF data sets into the AI-framework.
- We formulate an integer linear programming (ILP) problem to minimize the operational latency of the CN, instantiate CP and UP NF instances, assign them to CN servers, assign slices and UEs to NF instances, and allocate computational resources while maintaining virtual NF isolation and physical planar separation.
- We propose three additional CN operation configurations that should be considered for the envisioned 6G CNs, each offering an added degree of reliability, isolation, or both.

We conclude this section with a brief outline of this chapter: we present our system model in Section 5.1. In Section 5.2, we present our ILP problem while in Section 5.3, we discuss our AI-based solution. In Section 5.4, we offer a detailed discussion of our extensive simulation results. Finally, in Section 5.5, we summarize the problems addressed, the adopted approaches used to address them, and the obtained results.

5.1 System Model

In this section, we present the system model of our work (Figure 5.1). The CN consists of a pool of servers that run virtual instances of the CN NFs to service the 6G RAN slices known as further enhanced mobile broadband (feMBB) and extremely reliable low latency communications (ERLLC); while we provide a short summary of the NFs in Table 5.1, readers are referred to [66] for more details of the overall CN architecture. The servers, which are denoted by the set $\mathcal{N}$ and indexed by $n = 1, 2, ..., |\mathcal{N}|$, have a

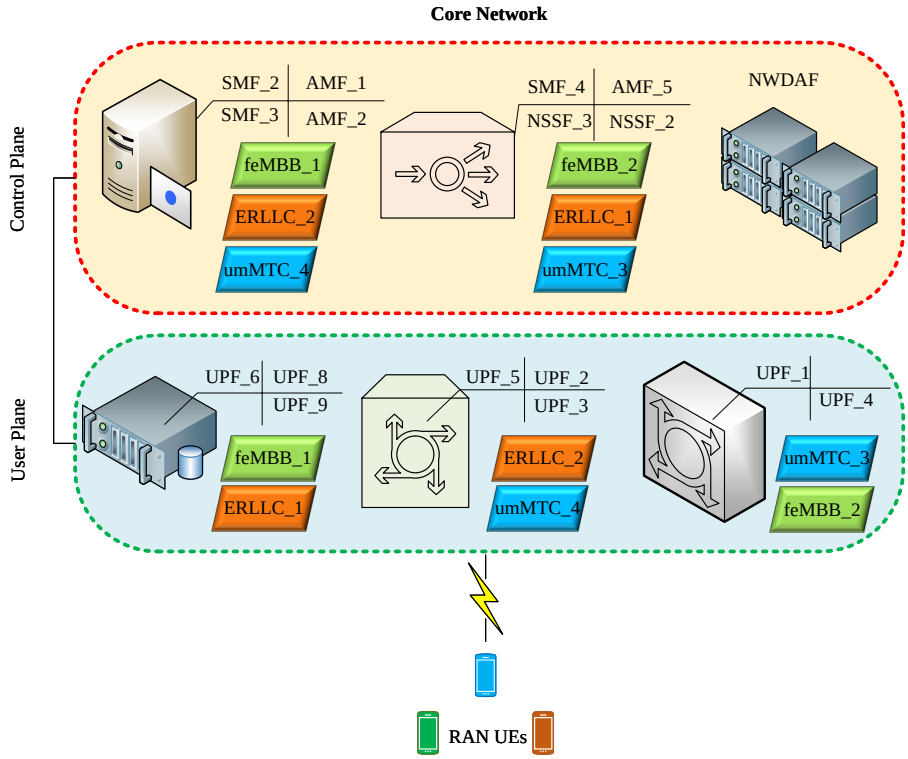

Figure 5.1 Core network system model.

maximum clock speed (cycles/sec) represented by C. Recall that the CN has both the control and user planes, and thus we designate two sets, accordingly, for the CP and UP NFs: $\mathcal{F}_{CP}$ and $\mathcal{F}_{UP}$ which are indexed as $f = 1, 2, ..., |\mathcal{F}_{CP}|$ and $g = 1, 2, ..., |\mathcal{F}_{UP}|$, correspondingly. The instances for each CP NF are represented by the set $\mathcal{I}_f$ while those of the UP are denoted by the set $\mathcal{I}_g$, and are indexed by $i = 1, 2, ..., |\mathcal{I}_f|$ and $j = 1, 2, ..., |\mathcal{I}_g|$, accordingly. $C_{f,i}^{CP}$ denotes the cycles per second allocated to instance i of CP NF f while $C_{g,j}^{UP}$ denotes the same for instance j of UP NF g; C denotes the maximum computing capacity of the individual servers (homogeneous). The set of slices and UEs are denoted by $\mathcal{S}$ and $\mathcal{U}_s$, respectively.

Binary indicators $A_{f,i,n}$ and $B_{g,j,n}$ are used to indicate if instance i of CP NF f and instance j of UP NF g are running at server n, accordingly. To indicate if CP NF and UP NF instances are assigned to slice s, we utilize binary indicators $K_{f,i}^{s}$ and $V_{g,j}^{s}$, correspondingly. We denote $Y_{f,i,u}^{s}$ and $Z_{g,j,u}^{s}$ to indicate if user equipment (UE) u

Table 5.1 Description of Core Network Functions

NFs	Description
NSSF	The NSSF (Network Slice Selection Function) maintains a list of the network slice instances and facilitates slice access to the UEs based on their service requests.
NEF	The NEF (Network Exposure Function) broadcasts network services and capabilities to enable external developers to create their own specialized network services.
NRF	The NRF (Network Function Repository Function) tracks the instantiated NF instances and allows NFs to discover each other.
UPF	The UPF (User Plane Function) carries out packet routing, forwarding, inspection, QoS flows, protocol data unit session handling.
AMF	The AMF (Access and Mobility Management Function) manages user mobility, connection, registration, and anchors RAN subscribers to the CN.
AUSF	The AUSF (Authentication Server Function) verifies UEs' credentials to ensure that they are authorized to access the network.
SMF	The SMF (Session Management Function) performs session management, IP address allocation, control plane QoS management, and policy enforcement.
PCF	The PCF (Policy Control Function) conducts policy enforcement, billing and subscription, information access, and behavior governance.
UDM	The UDM (Unified Data Management) is responsible for UE ID handling, subscription management, and roaming access authorization.
AF	The AF (Application Function) advertises applications to UEs, interacts with PCF for application access control, operates similarly to the NEF but with respect to applications for UEs.

of slice s is assigned to instance i of CP NF f and instance j of UP NF g, accordingly. Lastly, τ_s denotes the CP deadline of the slice. For ease, Table 5.2 provides a list of notations and definitions utilized in this work.

Each NF when communicating with other NFs has a sequence of control packets with an average payload size and standard deviation (normal distribution) [45]. If we denote $S^{CP}_{f,i,x}$ and $N^{CP}_{f,i,x}$ as the average size (bits) and average number of control packets *received by* instance i of CP NF f from *CP NF* x per UE, respectively, $S^{UP}_{f,i,y}$ and $N^{UP}_{f,i,y}$ as the average size (bits) and average amount of control packets *received by* instance i of CP NF f from *UP NF* y per UE, respectively, and ω_f as the cycles per bit required by NF type f to execute the control tasks, then the total average computational time of instance i of CP NF f with respect to slice s is:

$$T^{CP,s}_{f,i} = \frac{(\sum_{x=1}^{|\mathcal{F}_{CP}|} S^{CP}_{f,i,x} N^{CP}_{f,i,x} + \sum_{y=1}^{|\mathcal{F}_{UP}|} S^{UP}_{f,i,y} N^{UP}_{f,i,y})}{C^{CP}_{f,i}(\omega_f K^s_{f,i} \sum_{u=1}^{|\mathcal{U}_s|} Y^s_{f,i,u})^{-1}}. \tag{5.1}$$

It should be stated that an NF instance does not communicate with itself; and hence, such a payload is essentially zero.

As for the UP payload, if we denote $S^{UP}_{z,g,j}$ and $N^{UP}_{z,g,j}$ as the average size (bits) and average number of control packets *received by* instance j of UP NF g from CP NF z per UE, respectively, and β_g as the cycles per bit required by NF type g to execute the control tasks, we can determine the total average computational time of instance j of UP NF g with respect to slice s as follows:

$$T^{UP,s}_{g,j} = \frac{(\sum_{z=1}^{|\mathcal{F}_{CP}|} S^{UP}_{g,j,z} N^{UP}_{g,j,z})}{C^{UP}_{g,j}(\beta_g V^s_{g,j} \sum_{u=1}^{|\mathcal{U}_s|} Z^s_{g,j,u})^{-1}}. \tag{5.2}$$

Notice how the UP latency equation differs from that of the CP because CP NFs coordinate with other CP NFs but UP NF instances do not communicate with other UP NF instances; they only coordinate with CP NFs. As a matter of fact, User Plane Function (UPF), which is the only UP NF, communicates with only one other CP NF: the Session Management Function (SMF). We point out here that we focus on

Table 5.2 Summary of Notations for Core Network

Notations	Definitions
$A_{f,i,n}$	Binary variable denoting if CP NF instance i of function f runs at CN node n.
$B_{g,j,n}$	Binary variable denoting if UP NF instance g of function j runs at CN node n.
$C_{f,i}^{CP}$	Actual CPU usage (cycles/sec) of instance i of CP NF f.
$C_{g,j}^{UP}$	Actual CPU usage (cycles/sec) of instance j of UP NF g.
C_n	Maximum computing capacity (cycles/sec) of servers.
$K_{f,i}^{s}$	Binary variable denoting if instance i of CP NF function f is assigned to slice s.
$N_{f,i,x}^{CP}$	Average number of control packets received by instance i of CP NF f from CP NF x per UE.
$N_{f,i,y}^{UP}$	Average amount of control packets received by instance i of CP NF f from UP NF y per UE.
$N_{g,j,z}^{UP}$	Average number of control packets received by instance j of UP NF g from CP NF z per UE.
$S_{f,i,x}^{CP}$	Average size (bits) of control packets received by instance i of CP f from CP NF x per UE.
$S_{f,i,y}^{UP}$	Average size (bits) of control packets received by instance i of CP NF f from UP NF y per UE.
$S_{g,j,z}^{UP}$	Average size (bits) of control packets received by instance j of UP NF g from CP NF z per UE.
$T_{f,i}^{CP,s}$	Average control task completion time of instance i of CP NF f at node n for slice s.
$T_{g,j}^{UP,s}$	Average control task completion time of instance j of UP NF g at node n for slice s.
$V_{g,j}^{s}$	Binary variable denoting if instance j of UP NF function g is assigned to slice s.
$Y_{f,i,u}^{s}$	Binary variable denoting if UE u of slice s is assigned to instance i of CP NF f.
$Z_{g,j,u}^{s}$	Binary variable denoting if UE u of slice s is assigned to instance j of UP NF g.
τ^{s}	Maximum latency tolerated by slice s.
ω_f	CPU (cycles per bit) required for (any instance of) CP NF f.
β_g	CPU (cycles per bit) required for (any instance of) UP NF g.
ϵ	Penalty factor for DRL cost function constraint violation.
μ_a	Learning rate of actor network.
μ_c	Learning rate of critic network.
ζ	Parameter for actor network.
θ	Parameter for critic network.
$\zeta^{'}$	Parameter for target actor network.
$\theta^{'}$	Parameter for target critic network.
δ	Temporal difference error.
λ	Discount factor.

the CN only, specifically the control traffic among its NFs; and hence, user traffic between the RAN and UPF is not considered here either. This is primarily due to the abundance of works which have already dealt with RAN traffic in the past.

5.1.1 Core network configurations

There are several CN operation configurations that impact the reliability and isolation of the CN. With respect to the virtual NF isolation of the CN, there are two configurations. The first one is the existing slicing method at the CN which we refer to as the *partial virtual isolation* (PVI) configuration. In this configuration, a UP virtual NF instance cannot service more than one slice (this restriction does not apply for any CP NF instance). Thus, the virtual isolation occurs only at the UP and not the CP. The second one, which we are proposing (along with an additional overlaid configuration explained shortly afterwards), is called the *full virtual isolation* (FVI) configuration. In this configuration, the CN operates with maximum slicing isolation in that both the CP and UP NF instances are isolated, *i.e.*, each virtual NF instance in both planes cannot serve more than one slice. While this guarantees a higher degree of slice isolation, this is more stringent and perhaps costly.

We propose one more configuration as well; it is a *physical* configuration called planar physical isolation (PPI) which means that a CN server cannot host NF instances of different planes simultaneously. If any CP NF instance runs at a server, no UP NF instance can co-exist at said server and vice versa. This can be integrated with the aforementioned *virtual* configurations, thus giving rise to four configurations (starting from the least stringent to the most): PVI, FVI, PPI+PVI, and PPI+FVI. As highlighted already, 3GPP utilizes only the PVI configuration which affords the least slicing isolation, security, or reliability. We note here that the names of these configurations, including the one used currently by 3GPP, are not standardized or conventional names found in any of 3GPP releases or academic literature; they are our

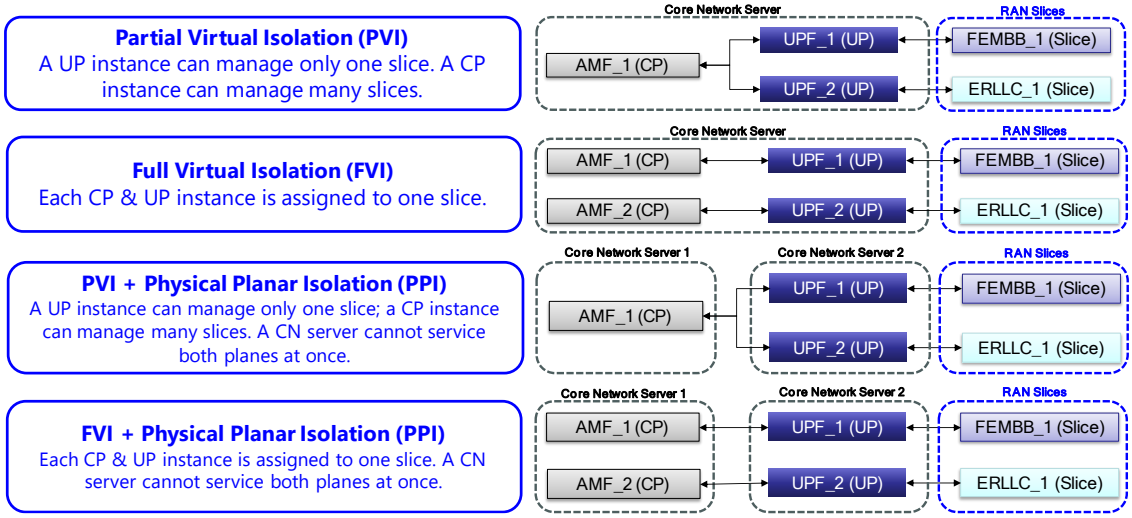

Figure 5.2 Core network configurations.

short hand forms for this work only. We propose PPI+FVI (and FVI and PPI+PVI by extension) which does offer the maximum virtual NF and physical planar isolation, security, and reliability. Examples of the discussed configurations are depicted in Figure 5.2.

5.2 Problem Formulation

We now present our aggregate latency minimization problem, **P5**, and the associated constraints below.

$$\textbf{P5:} \min_{A_{f,i,n},B_{g,j,n},C^{CP}_{f,i},C^{UP}_{g,j},K^{s}_{f,i},V^{s}_{g,j},Y^{s}_{f,i,u},Z^{s}_{g,j,u}} \sum_{s=1}^{|\mathcal{S}|} \left(\sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} T^{CP,s}_{f,i} + \sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} T^{UP,s}_{g,j} \right) \tag{5.3}$$

$$\text{s.t.} \sum_{s=1}^{|\mathcal{S}|} K^{s}_{f,i} \leq 1, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \tag{5.4}$$

$$\sum_{s=1}^{|\mathcal{S}|} V^{s}_{g,j} \leq 1, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \tag{5.5}$$

$$\sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,n} C^{CP}_{f,i} + \sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,n} C^{UP}_{g,j} \leq C_n, \forall n \in \mathcal{N}, \tag{5.6}$$

$$T_{f,i}^{CP,s} \leq \tau^s, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \forall s \in \mathcal{S}, \tag{5.7}$$

$$T_{g,j}^{UP,s} \leq \tau^s, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \forall s \in \mathcal{S}, \tag{5.8}$$

$$\sum_{i=1}^{|\mathcal{I}_f|} Y_{f,i,u}^s \leq 1, \forall f \in \mathcal{F}_{CP}, \forall u \in \mathcal{U}_s, \forall s \in \mathcal{S}, \tag{5.9}$$

$$\sum_{j=1}^{|\mathcal{I}_g|} Z_{g,j,u}^s \leq 1, \forall g \in \mathcal{F}_{UP}, \forall u \in \mathcal{U}_s, \forall s \in \mathcal{S}, \tag{5.10}$$

$$\text{sgn}\left(\sum_{f=1}^{|\mathcal{F}_{CP}|}\sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,n}\right) + \text{sgn}\left(\sum_{g=1}^{|\mathcal{F}_{UP}|}\sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,n}\right) \leq 1, \forall n \in \mathcal{N}, \tag{5.11}$$

$$A_{f,i,n} \in \{0,1\}, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \forall n \in \mathcal{N}, \tag{5.12}$$

$$B_{g,j,n} \in \{0,1\}, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \forall n \in \mathcal{N}, \tag{5.13}$$

$$C_{f,i}^{CP} \in \mathbb{Z}^+, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \tag{5.14}$$

$$C_{g,j}^{UP} \in \mathbb{Z}^+, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \tag{5.15}$$

$$K_{f,i}^s \in \{0,1\}, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \forall s \in \mathcal{S}, \tag{5.16}$$

$$V_{g,j}^s \in \{0,1\}, \forall s \in \mathcal{S}, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \tag{5.17}$$

$$Y_{f,i,u}^s \in \{0,1\}, \forall s \in \mathcal{S}, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \forall u \in \mathcal{U}_s, \tag{5.18}$$

$$Z_{g,j,u}^s \in \{0,1\}, \forall s \in \mathcal{S}, \forall g \in \mathcal{F}_{UP}, \forall i \in \mathcal{I}_g, \forall u \in \mathcal{U}_s. \tag{5.19}$$

The objective function is to minimize the aggregate average latency of the entire network. Equations (5.4)-(5.5) enforce isolation of slices at the CP and UP NFs,

respectively. Both are required for FVI while only the latter is needed for PVI. Equation (5.6) ensures that the maximum computing capacity is not violated while Equations (5.7)-(5.8) dictate the slice latency constraints. Equations (5.9)-(5.10) restrict a user to one instance of any CP and UP NF within one slice. Equation (5.11) enforces PPI. Finally, the integer decision variables are listed in Equations (5.12)-(5.19). This problem is evidently NP-hard and very complicated due to the three dimensional nature of several decision variables.

5.3 Proposed AI-Algorithm Solution

In this section, we outline a deep reinforcement learning (DRL) method known as the actor-critic technique utilized to efficiently solve **P5**. First, we must define the associated Markov Decision Process (MDP) $< \mathcal{W}, \mathcal{A}, \mathcal{F}, \mathcal{C} >$ which consists of state space $\mathcal{W}$, action space $\mathcal{A}$, state transition probability density function $\mathcal{F} : \mathcal{W} \times \mathcal{A} \times \mathcal{W} \mapsto [0, \infty)$, and cost function $\mathcal{C} : \mathcal{S} \times \mathcal{A} \mapsto [0, \infty)$. The network is initialized into state $\boldsymbol{s}(t)$, executes action $\boldsymbol{a}(t)$, incurs cost $\boldsymbol{c}(\boldsymbol{s}(t), \boldsymbol{a}(t))$, and transitions into state $\boldsymbol{s}(t+1)$.

The network state describes the computing capacity remaining at each server below:

$$\boldsymbol{s}(t) = \boldsymbol{C} - [\sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,1}(t) C_{f,i}^{CP}(t) + \sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,1}(t) C_{g,j}^{UP}(t), \sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,2}(t) C_{f,i}^{CP}(t) + \sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,2}(t) C_{g,j}^{UP}(t), ..., \sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,|\mathcal{N}|}(t) C_{f,i}^{CP}(t) + \sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,|\mathcal{N}|}(t) C_{g,j}^{UP}(t)]. \tag{5.20}$$

where $\boldsymbol{C} = [C_1, C_2, ..., C_n]$. The network action assigns NF instances to servers, slices, and users, and then allocates computing resources to them as follows:

$$\boldsymbol{a}(t) = [\boldsymbol{A}(t), \boldsymbol{B}(t), \boldsymbol{C}^{CP}(t), \boldsymbol{C}^{UP}(t), \boldsymbol{K}(t), \boldsymbol{V}(t), \boldsymbol{Y}(t), \boldsymbol{Z}(t)], \tag{5.21}$$

where $\boldsymbol{A}(t), \boldsymbol{B}(t), \boldsymbol{C}^{CP}(t), \boldsymbol{C}^{UP}(t), \boldsymbol{K}(t), \boldsymbol{V}(t), \boldsymbol{Y}(t)$ and $\boldsymbol{Z}(t)$ are vectors that form the action space below:

$$\begin{aligned}\mathcal{A}(t) = \{&\boldsymbol{A}(t)|A_{f,i,n}(t) \in \{0,1\}, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \forall n \in \mathcal{N}, \\ &\boldsymbol{B}(t)|B_{g,j,n}(t) \in \{0,1\}, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \forall n \in \mathcal{N}, \\ &\boldsymbol{C}^{CP}(t)|C^{CP}_{f,i}(t) \geq 0, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \\ &\boldsymbol{C}^{UP}(t)|C^{UP}_{g,j}(t) \geq 0, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \\ &\boldsymbol{K}(t)|K^{s}_{f,i}(t) \in \{0,1\}, \forall s \in \mathcal{S}, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \\ &\boldsymbol{V}(t)|V^{s}_{g,j}(t) \in \{0,1\}, \forall s \in \mathcal{S}, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \\ &\boldsymbol{Y}(t)|Y_{f,i,u}(t) \in \{0,1\}, \forall f \in \mathcal{F}_{CP}, \forall i \in \mathcal{I}_f, \forall u \in \mathcal{U}_s, \forall s \in \mathcal{S}, \\ &\boldsymbol{Z}(t)|Z_{g,j,u}(t) \in \{0,1\}, \forall g \in \mathcal{F}_{UP}, \forall j \in \mathcal{I}_g, \forall u \in \mathcal{U}_s, \forall s \in \mathcal{S}\}. \end{aligned} \tag{5.22}$$

Therefore, given state $\boldsymbol{s}(t)$ and action $\boldsymbol{a}(t)$, the consequent cost is:

$$\begin{aligned} c(\boldsymbol{s}(t), \boldsymbol{a}(t)) = &\sum_{s=1}^{|\mathcal{S}|} \tau^{E2E,s}(t) + \epsilon \Bigg(\mathbb{B}\Bigg(\sum_{s=1}^{|\mathcal{S}|} K^{s}_{f,i}(t) > 1 \Bigg) + \mathbb{B}\Bigg(\sum_{s=1}^{|\mathcal{S}|} V^{s}_{g,j}(t) > 1 \Bigg) + \\ &\mathbb{B}\Bigg(\sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,n}(t) C^{CP}_{f,i}(t) + \sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,n}(t) C^{UP}_{g,j}(t) > C \Bigg) + \\ &\mathbb{B}\Bigg(\sum_{i=1}^{|\mathcal{I}_f|} Y^{s}_{f,i,u}(t) > 1 \Bigg) + \mathbb{B}\Bigg(\sum_{j=1}^{|\mathcal{I}_g|} Z^{s}_{g,j,u}(t) > 1 \Bigg) + \mathbb{B}\Bigg(\text{sgn}\Bigg(\sum_{f=1}^{|\mathcal{F}_{CP}|} \sum_{i=1}^{|\mathcal{I}_f|} A_{f,i,n}(t) \Bigg) + \\ &\text{sgn}\Bigg(\sum_{g=1}^{|\mathcal{F}_{UP}|} \sum_{j=1}^{|\mathcal{I}_g|} B_{g,j,n}(t) \Bigg) > 1 \Bigg) + \mathbb{B}(T^{CP,s}_{f,i}(t) > \tau^s) + \mathbb{B}(T^{UP,s}_{f,i}(t) > \tau^s) \Bigg). \end{aligned} \tag{5.23}$$

We exploit the the penalty method above, where $\epsilon > 0$ is the penalty parameter and $\mathbb{B}(\cdot)$ is the Boolean function (*i.e.*, $\mathbb{B}(x) = 1$ if x is true and 0 otherwise), to penalize the network when any of the CN constraints are violated [67].

The aim of the MDP is to find an optimal stochastic policy $\pi(\boldsymbol{s}, \boldsymbol{a}) = Pr\{\boldsymbol{a}(t) = \boldsymbol{a} | \boldsymbol{s}(t) = \boldsymbol{s}\}$ that minimizes the expected value of the discounted cost, $J(\pi)$, over all time steps starting from state $\boldsymbol{s}(0)$. Hence, we need to express the expected value of

the state-action value function as follows:

$$\mathrm{Q}(\boldsymbol{s}(t), \boldsymbol{a}(t)) = \mathbb{E}\{\sum_{i=t}^{\infty} \lambda^{i-t} c(\boldsymbol{s}(t), \boldsymbol{a}(t))\}, \tag{5.24}$$

where $\lambda \in [0, 1]$ is the discount factor to prioritize long-term rewards (*i.e.*, future costs). Assuming that we have a vast number of training episodes (not indexed here for simplicity), each having numerous training time slots, we can express the expected cost, $J(\pi)$, over all the training episodes as follows (this being a random process) [68]:

$$J(\pi) = \mathbb{E}\{\mathrm{Q}(\boldsymbol{s}(0), \boldsymbol{a}(0))\}. \tag{5.25}$$

Note that the state of the initial time slot in an episode, $\boldsymbol{s}(0)$, is randomly set; and hence, the initial states across all the training episodes are not identical. Hence, the DRL policy will seek to minimize the expected cost across all the training episodes that are initialized randomly.

In order to optimally solve the MDP and calculate the expected cost, the transition probabilities are required. However, their exact values are difficult to determine due to the vast state and action spaces resulting in a huge computational complexity. To make matters worse, the actual network conditions tend to deviate from any predefined state transition model. A more practical method would be to utilize a model-free reinforcement learning method, which the CN is expected to be able to execute, to solve this problem. This very challenge allows us to leverage the actor-critic technique since it can quickly approximate the state-action values, $\mathrm{Q}(\boldsymbol{s}, \boldsymbol{a})$, of the vast state and action spaces without having to exhaustively explore them [69].

5.3.1 Actor-critic method

In this section, we outline the actor-critic technique utilized to obtain the optimal policy and minimize the cost function value. It is a technique which has been

proven to handle mixed-domain (continuous and integer) action spaces quite well [70]. Unlike most other basic DRL methods, this technique exploits two pairs of neural networks: an actor network (and target actor network) which learns the parameterized policy and a critic network (and target critic network) which approximates the state-action values and evaluates the actor network's current policy. The actor leverages a parameterized function, $\pi_\zeta(\boldsymbol{s})$, where ζ is the actor network's parameter, to generate an action for state $\boldsymbol{s}$. The critic network evaluates the actor network's policy by updating the state-action value function, $Q_\theta(\boldsymbol{s}, \boldsymbol{a})$, and its (critic network) parameters, θ, via the temporal difference method [71]. Thereafter, the actor network updates its parameters based on the new state-action value function via the policy gradient method [69]. Both the policy gradient and temporal difference methods are explained next.

The actor network utilizes the policy gradient method to update parameter ζ through the gradients of $J(\pi)$ as expressed next:

$$\triangle_\zeta J(\pi_\zeta) = \frac{\partial J(\pi_\zeta)}{\partial \pi_\zeta} \frac{\partial \pi_\zeta}{\partial \zeta} = \mathbb{E}\{\triangle_a \mathrm{Q}_\theta(\boldsymbol{s}, \boldsymbol{a}) \triangle_\zeta \pi_\zeta(\boldsymbol{s})\}, \tag{5.26}$$

where $\mathrm{Q}_\theta(\boldsymbol{s}, \boldsymbol{a})$ is supplied by the critic network. If μ_a denotes the learning rate for the actor network, then ζ must be adjusted as follows:

$$\zeta = \zeta + \mu_a \triangle_\zeta J(\pi_\zeta). \tag{5.27}$$

As for the critic network, it evaluates the actor network's policy, $\pi_\zeta(\boldsymbol{s})$, and then updates its own parameter, θ, via the temporal difference method which entails using the temporal difference error, $\delta(t)$, to predict the state-action value function as follows (it is also known as the target value):

$$\delta(t) = c(\boldsymbol{s}(t+1), \boldsymbol{a}(t+1)) + \gamma \mathrm{Q}_\theta(\boldsymbol{s}(t+1), \boldsymbol{a}(t+1)) - \mathrm{Q}_\theta(\boldsymbol{s}(t), \boldsymbol{a}(t)). \tag{5.28}$$

If μ_c denotes the learning rate of the critic network, then by leveraging the temporal difference error, θ too can be updated via gradient descent as shown next:

$$\theta(t+1) = \theta(t) + \mu_c \delta(t) \bigtriangleup_\theta \mathrm{Q}_\theta(\boldsymbol{s}(t), \boldsymbol{a}(t)). \tag{5.29}$$

There is a slight issue however with the above process; notice how both of the previous equations contain the $\mathrm{Q}_\theta(\boldsymbol{s}, \boldsymbol{a})$ term. This may result in a conflict for the calculations and update process; thus, a target critic network, $\mathrm{Q}'_{\theta'}(\boldsymbol{s}, \boldsymbol{a})$ is initialized to serve as a copy of the actual critic network and calculate $\delta(t)$. The target critic network is then updated by $\theta^{'} = \tau\theta + (1-\tau)\theta'$, where $\tau \ll 1$ to gradually alter the target critic network to enhance its learning stability. As done with the target critic network, a target critic network, $\pi^{'}_{\zeta'}(\boldsymbol{s})$, is also initialized and updated by $\zeta^{'} = \tau\zeta + (1-\tau)\zeta'$.

We now summarize the entire process as shown in Algorithm 4. Lines 1-2 initialize the actor, critic, target actor, and target critic networks. Line 3 initializes the first time step and network state. Lines 4-11 adjust the stochastic policy and update all (neural) networks. Specifically, in Line 5, the actor network executes an action in the environment resulting in a cost which is calculated in Line 6 for the current state-action pair. Note that initially, the policy, actor, and critic networks' parameters are randomly initialized since the environment has not yet been explored. The action affects change in the network environment and leads to a new state in the next time step in Line 7. Line 8 updates the critic network via the temporal difference method while Line 9 updates the actor network via the policy gradient method. The target actor and critic networks are subsequently updated in Lines 10-11. Finally, the network transitions to the next time step in Line 12.

5.4 Simulation Results

In this section, we offer a detailed discussion of our extensive results. Specifically, we compare the average latency of the ERLLC and feMBB slices under all of the

Algorithm 4: Actor Critic Technique

Input: $\mathcal{S}$, $\mathcal{N}$, $\mathcal{U}_s$, $\mathcal{F}^{CP}$, $\mathcal{F}^{UP}$, μ_a, μ_c, τ^s, C, γ

Output: Policy π

Initialize actor network $\pi_\zeta(\boldsymbol{s})$ and critic network $\mathrm{Q}_\theta(\boldsymbol{s}, \boldsymbol{a})$

Initialize target actor network $\pi'_\zeta(\boldsymbol{s})$ and target critic network $\mathrm{Q}'_\theta(\boldsymbol{s}, \boldsymbol{a})$

Initialize time step $t = 0$ and corresponding state $s(0)$

for *every time step* t **do**

  Determine action $\boldsymbol{a}(t)$ via actor network $\pi_\zeta(\boldsymbol{s})$

  Calculate cost $c(s(t), \boldsymbol{a}(t))$ of associated state-action pair

  Observe resulting network state $\boldsymbol{s}(t+1)$

  Update critic network $\mathrm{Q}_\theta(\boldsymbol{s}, \boldsymbol{a})$ via the temporal difference method

  Update actor network $\pi_\zeta(\boldsymbol{s})$ via the policy gradient method

  Update the target actor network $\pi'_\zeta(\boldsymbol{s})$ via $\zeta' = \tau\zeta + (1-\tau)\zeta'$

  Update target actor network $\mathrm{Q}'_\theta(\boldsymbol{s}, \boldsymbol{a})$ via $\theta' = \tau\theta + (1-\tau)\theta'$

  Transition $t \leftarrow t+1$

end

four configurations discussed earlier. Because our focus here is limited to the control traffic between three of the most significant NFs, the various configurations and their performance results will be impacted accordingly. Our results are within 95 percent accuracy; the terminating fifty simulation iterations were within a five percent range. A more stringent criterion would have taken longer to converge without an appreciable difference in the results. The relevant simulation parameters are outlined in Table 5.3.

Figures 5.3 to 5.6 depict the average latency of the core network. We note that under no conditions for any CN configuration did any user ever violate any QoS constraints. Figure 5.3 demonstrates that *despite* very low computational capacity (25 servers only), PPI+FVI outperforms PVI in nearly all cases. Moreover, PPI+FVI always performs the best for both slices under the highest network load (50 UEs). As we augment more computational capacity, we observe that the feMBB slice

Table 5.3 Simulation Parameters for Core Network

Parameter	Value
Server Pool Size	[25 50 75 100]
Server Computing Capacity	12 GHz
ERLLC UEs	[10 20 30 40 50]
feMBB UEs	[10 20 30 40 50]
Slice Deadline	ERLLC: 0.5 ms, feMBB: 20 ms
Required CPU cycles/bit	AMF: 20, SMF: 30, UPF: 50
NF Control Traffic Dataset	Refer to [45]

performance under PPI+FVI improves greatly (the feMBB slice latency is 0.55 ms in a 25-server CN for 50 UEs but 0.33 ms in a 100-server CN).

For the ERLLC slice, in all cases except $\geq$ 10 UEs, PPI+FVI outperforms the rest of the configurations regardless of CN server count. The worst-case ERLLC slice latency occurs under PVI, the most relaxed configuration, in the 25-server CN at approximately 0.45 ms whereas for PPI+PVI, it is approximately 0.375 ms. We also notice for the ERLLC slice under 75 and 100-server CNs that as the network load increases, the performance degrades much faster under PVI than under PPI+FVI. In other words, the delta by which the PVI latency worsens (as the network load intensifies), increases much faster than that of PPI+FVI.

Generally speaking, because PPI+FVI imposes an added layer of stringency which results in the entirety of a CN server's computing resources being dedicated to one CN plane or another, we can conclude that PPI+FVI outperforms the other configurations for the ERLLC slice (the latency-sensitive slice), which is where it matters most. This stringency prevents two different CN planes from competing for resources within a server, ultimately improving performance for both. Therefore, to amply support low latency applications, it is justifiable to utilize our most stringent configuration, PPI+FVI. As for the feMBB slice, while PPI+FVI many not perform as optimally in some cases, it does always maintain superior performance for the highest-loaded networks ($\geq$ 50 UEs) regardless of the CN server count. This is

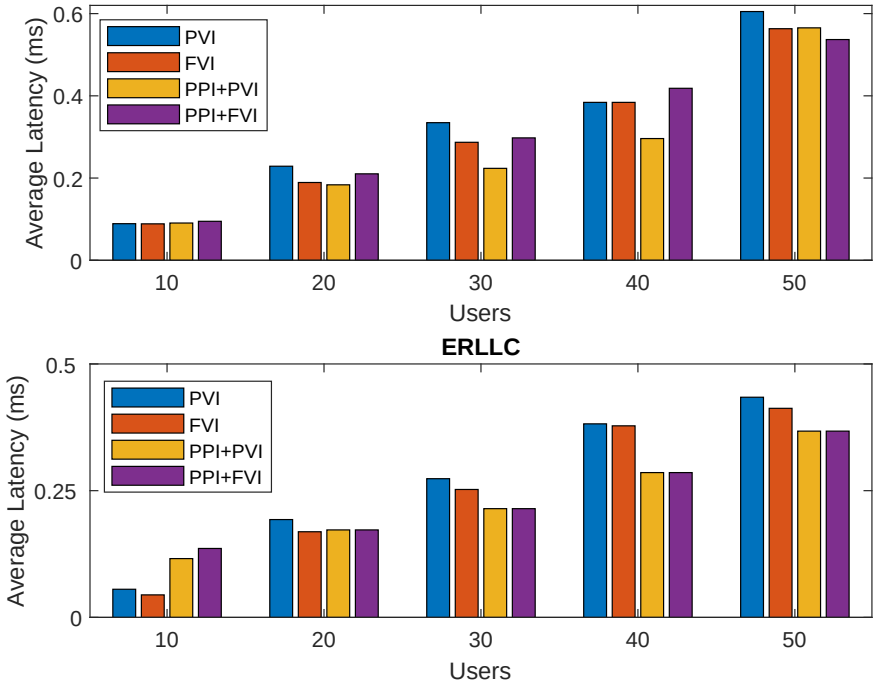

Figure 5.3 Average slice latency in a 25-server core network.

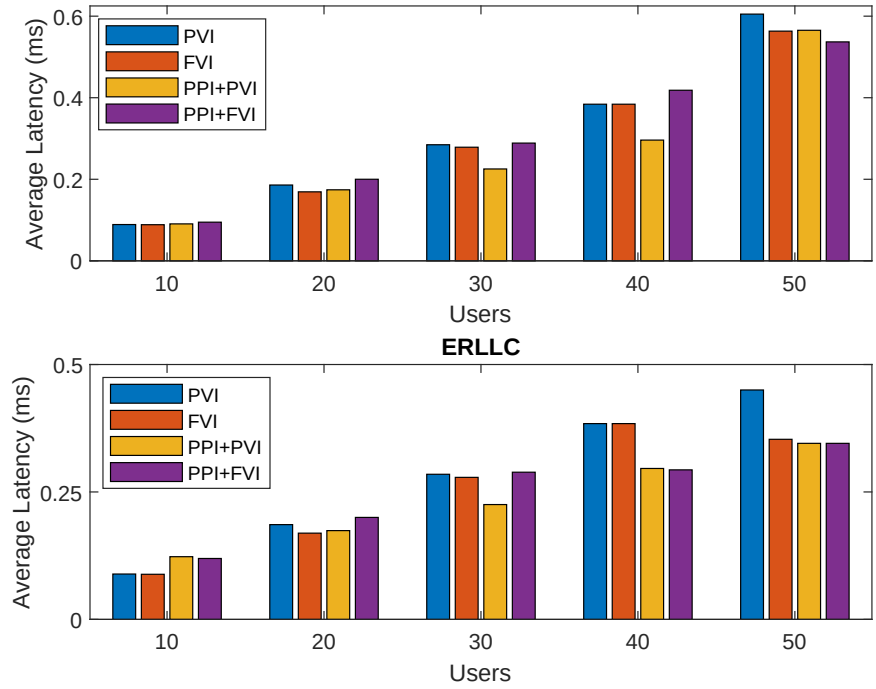

Figure 5.4 Average slice latency in a 50-server core network.

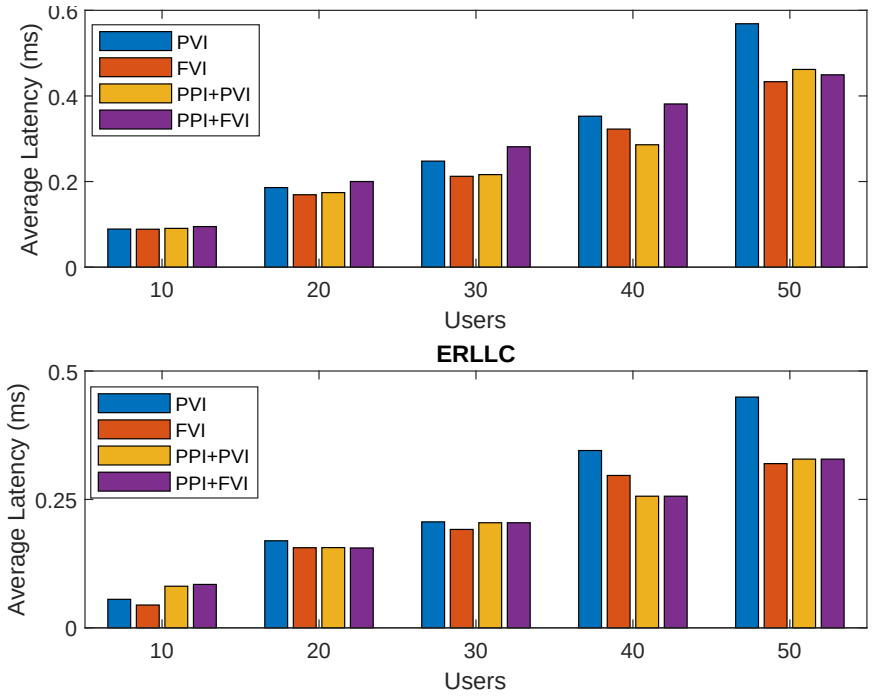

Figure 5.5 Average slice latency in a 75-server core network.

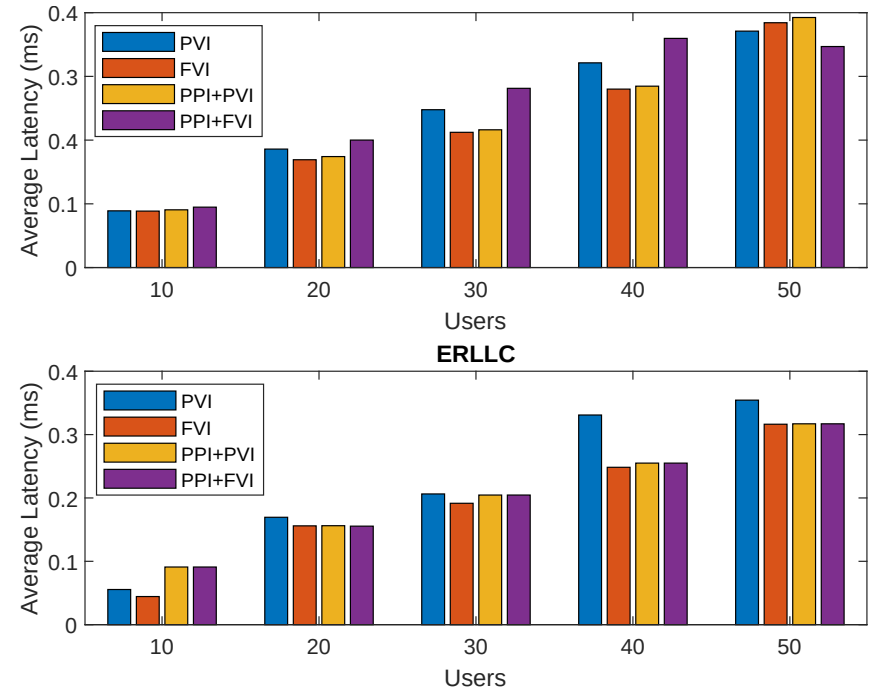

Figure 5.6 Average slice latency in a 100-server core network.

important to note especially for high-UE density situations. Since the feMBB slice can indeed tolerate a higher latency than the ERLLC slice, the same degree of latency minimization is definitely not required. Hence, the performance gap between PVI and PPI+FVI for lightly to moderately loaded networks ($\leq$ 40 UEs) is justifiable. We do note once again though that for heavily loaded networks ($\geq$ 50 UEs), the PPI+FVI latency is the least of all the configurations.

Additionally, we can conclude that FVI tends to be very similar to PPI+PVI in some cases and comparable to PPI+FVI in other cases. Consequently, there is no real incentive to use FVI over PPI+PVI or PPI+FVI unless there are not enough servers since PPI makes the control plane and user plane NF assignment to the CN servers more restricted. Numerous other conclusions can be drawn but our focus is majorly on the current slicing standard and our third proposed configuration (PPI+FVI).

5.5 Summary

In this chapter, we advanced the concept of 6G AI-Native for effective CN management. Specifically, we developed three distinct configurations that offer various degrees of virtual and planar physical isolation and analyzed their performances. We validated our proposed approach and provided detailed insights on the optimal configurations for specific service types and networking conditions. This study enhances understanding of the CN CP and UP NF provisioning for 6G networks. Specifically, we demonstrated that the most reliable and resilient configuration, PPI+FVI, performs comparably to the most relaxed standard configuration PVI under significant network loads. Furthermore, any performance gap is not significant except in extreme network load conditions which can be mitigated by augmenting computational capacity. We also noted that the latency constraints are always met by the CN under all conditions. Thus, we convincingly argued and amply demonstrated herein that the additional reliability and isolation offered by PPI+FVI is worth the

performance cost. Future work should include studies of the CN performance with all the CP NFs, while accounting for their bidirectional traffic distributions accurately, and considering the RAN to investigate the E2E latency as well as throughput for latency-sensitive and throughput-dependent services, simultaneously. Optimizing mapping of slices between the RAN and CN under different CN configurations should be studied as well.

CHAPTER 6

FUTURE WORK

In Chapter 4, we presented our dual-band UAV network in which we provision guaranteed and best-effort services across both the sub-6 GHz and mmWave bands. We minimized the QoS gap of the best-effort service while meeting the QoS constraints of the guaranteed services. The network topology was an elementary configuration consisting of a single BS and UAV which is a commendable starting point; this should however lead to the study of Coordinated Multi-Point (CoMP) techniques in aerial networks. CoMP consists of several sub-techniques which exploit the dynamic coordination of data transmission and reception at several geographically dispersed BSs to enhance system performance and QoS. Although it is not a new technology, it has not been implemented in 4G LTE networks with great success; and hence, is expected to become more mainstream in 5G and beyond networks.

6.1 Coordinated Multi-Point in Aerial Networks

The study of aerial networks from many perspectives including flight time maximization, location and trajectory optimization, free space optics for both communications and battery charging, has advanced greatly in recent years. However, to fully realize the highest potentials of aerial networking, network operators must seriously consider deploying multiple UAVs within a service area to maximize coverage, line-of-sight, and QoS for the users. Most aerial network topologies considered are quite elementary in that they generally consist of only a single base station and a UAV. In the interest of practicality, scenarios with multiple base stations and UAVs should be considered; this brings a whole set of unique challenges, not the least of which is hand over and CoMP.

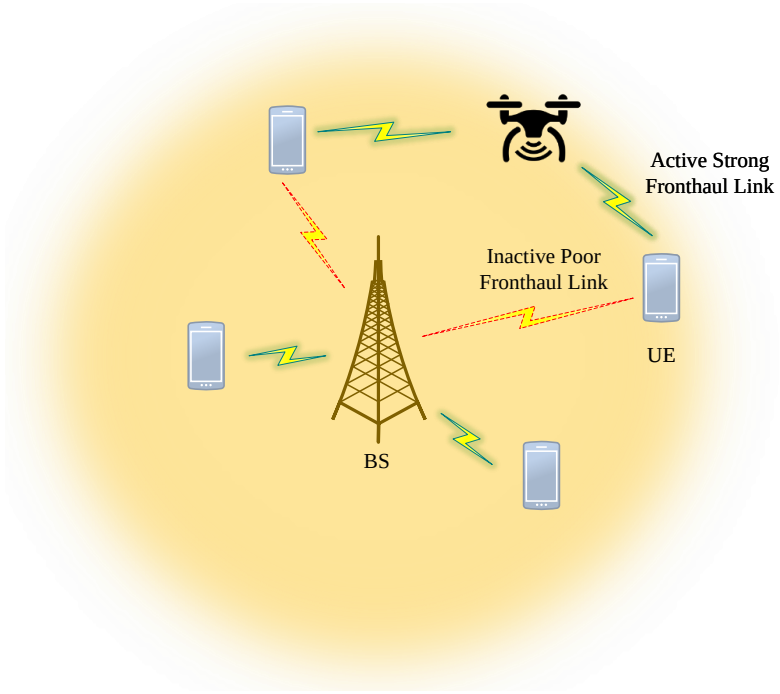

Figure 6.1 Legacy aerial network without CoMP.

CoMP entails the dynamic coordination of data transmission and reception at several physically dispersed nodes to enhance system performance and consists of three sub-techniques: Joint Processing (JP), Coordinated Scheduling and Beamforming (CSB), and Transmission Point Selection (TPS) [72]. JP is where multiple BSs coordinate to send data to a UE simultaneously; the UE's data is therefore available at all participating BSs. CSB is where a UE is served by a single BS but the scheduling decision is coordinated by multiple BSs. The UE's data is available at only one BS but the UE channel conditions are known by all BSs. Finally, TPS is where only one BS serves a UE at any given transmission time interval, but the UE's data is available at all BSs. Therefore, while all CoMP involves multiple BSs in some form, JP is the only CoMP technique that actually involves multiple BSs serving or transmitting to a single UE within a transmission time interval.

In conventional aerial networks, users can either connect to only a single UAV or a base station (Figure 6.1). However, by integrating CoMP with UAVs, in a bid

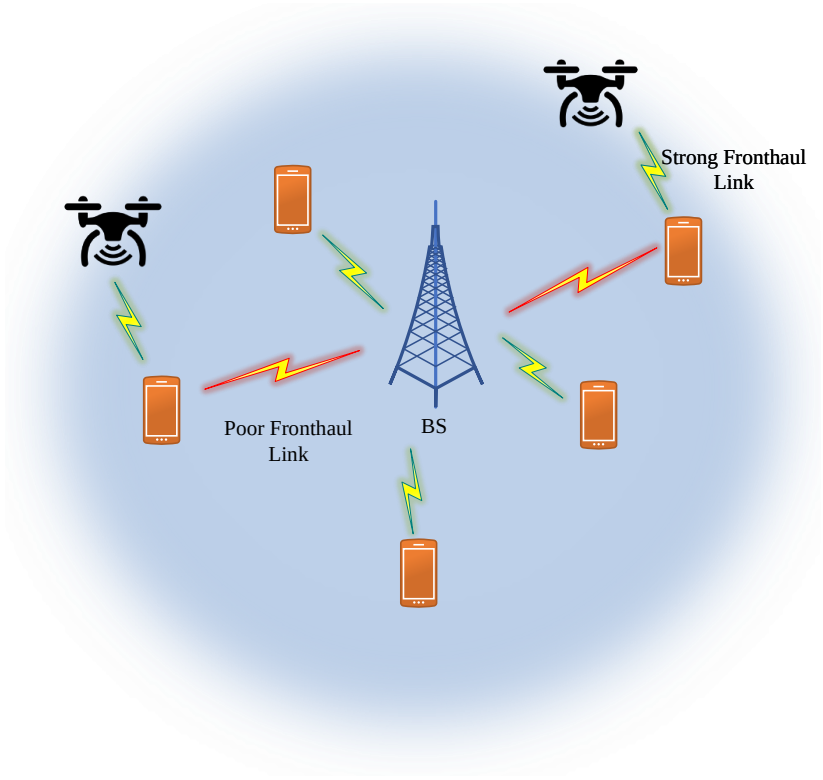

Figure 6.2 JP-CoMP aerial network with a single BS and UAV.

to maximize their throughput, channel quality, and performance, users can associate with several serving nodes concurrently, whether they be multiple UAVs, multiple base stations, or a combination of base stations and UAVs (Figures 6.2-6.3). While very robust, such networks are not without their unique issues such as but not limted to: 1) determining the optimal set or combination of nodes that users associate with, 2) resource allocation across multiple nodes for users, 3) coordination among the multiple UAVs in the hotspot (via master-slave or other hierarchies), 4) location and trajectory optimization for collision avoidance, and 5) optimizing handover procedures which need to be far more robust to handle not just user mobility but also the mobility of the aerial base stations, *i.e.*, UAVs. It must be kept in mind that user mobility is typically two-dimensional only whereas aerial mobility is three-dimensional which is far more complex. Exploiting AI to tackle such highly-dimensional problems has proven to be a promising solution in other contexts.

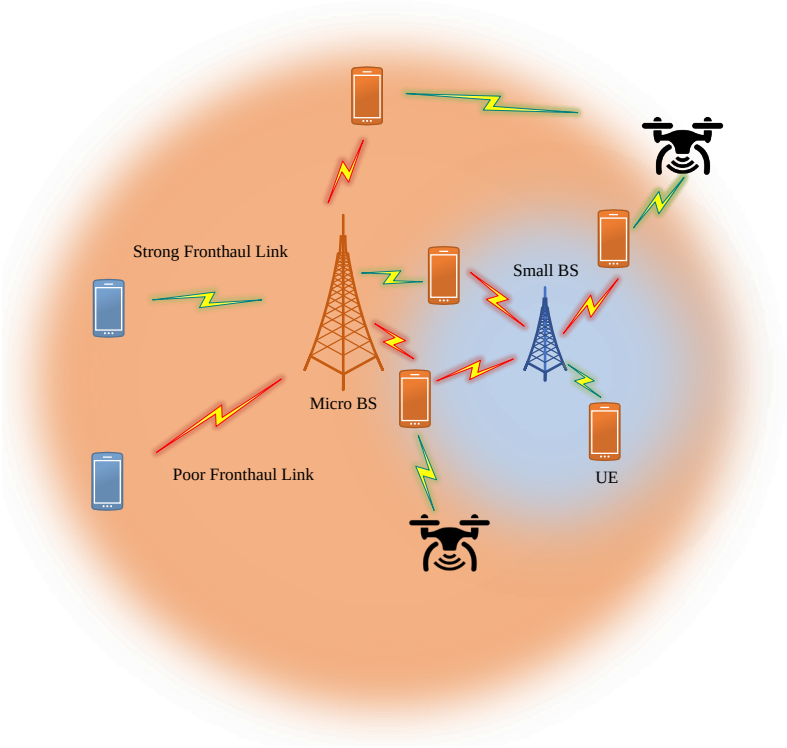

Figure 6.3 JP-CoMP aerial network with multiple BSs and UAVs.

6.2 Integrated Sensing and Communications

Integrated sensing and communications (ISAC) has received increased attention in light of the high-frequency bands to be employed by next-generation mobile networks; such waveform technologies natively support high-speed communications and high-resolution sensing, the latter of which, thus far, has been reserved exclusively for radar sensing platforms. However, because high frequency bands are expected to be utilized by 6G networks, the corresponding waveform technologies will bear strong resemblances to those of sensing platforms. Hence, it is deeply anticipated that mobile networks will conduct both sensing and communications in an integrated manner, hence, ISAC.

The added dimension of sensing in mobile networks opens up a new horizon of research opportunities for both RANs and CNs. It has the potential to enhance localization methods, map the physical world for the network to see,

enhance target identification and tracking, and improve beamforming and interference mitigation. Despite the exciting new opportunities, there are still numerous outstanding challenges including how and precisely which sensing parameters are to be measured, and most importantly, how to leverage them to optimize networks.

In light of the above considerations, 3GPP envisions proposing a new NF known as the *Sensing Service Function* (SSF) to process sensing data, generate crucial sensing analytics, and thereby enhance network decision-making. There are many avenues of research that can be pursued in this area; some would be with a heavy emphasis on the physical layer, specifically, how to measure different sensing parameters, such as LoS within an environment, velocity of users, radar cross sections for object identification, etc. Other avenues seek to leverage this additional sensing information within the CP to optimize network resource allocation, edge computing, UAV trajectory/location, and QoS performance. The applications of ISAC are almost endless and have opened the door for standardization research as well.

CHAPTER 7

CONCLUSION

We have studied E2E network slicing, specifically at the RAN, aerial (UAV) networks, and CNs in a bid to better meet the QoS requirements of users subscribed to various slices. At each segment of the E2E network, we proposed novel integrations of networking technology and impactful modifications to existing networking schemes that improved the performance at each individual segment. Improving the performance at each individual segment will ultimately lead to a holistic performance improvement. Starting at the RAN, we integrated the 3GPP 5G NR numerology schemes and dual-band transmissions with a TDD ground network to better meet the requirements of the EMBB and URLLC slices' requirements. We formulated two separate convex optimization problems to maximize the throughput of downlink EMBB users and minimize the uplink transmission power of the URLLC users. We then proposed a low complexity algorithm to choose the optimal numerology scheme and TDD duplex ratio to achieve the aforementioned aims. Through our results, we clearly demonstrated that utilizing dual-band transmissions with the optimal numerology schemes, which are dependent on the aim (*i.e.*, latency minimization, throughput maximization, transmission power minimization, spectral efficiency maximization, etc.), results in a significantly improved slicing performance.

We then transitioned to aerial networks that utilize UAVs to service hot spots that are remotely far from BSs and have no direct access to ground networks. To truly transform aerial networks seamless extensions of ground networks that are transparent to end users, we borrowed the innovative ideas we proposed for ground networks. Specifically, we proposed the use of dual-band transceivers for UAVs to facilitate dual-band connectivity for users requiring EMBB and URLLC services. However, due to the limited transmission power of aerial nodes, we treated the EMBB slice as a best-

effort service while guaranteeing the QoS requirements of the high-priority URLLC slice. We formulated a QoS gap minimization problem, an MINLP problem, and enforced a band-access policy to optimize the user-band association. Subsequently, we designed a low-complexity algorithm, PRiority BasED Resource AllocatIon in Adaptive SliCed NeTwork (PREDICT), to tackle the MINLP problem. Through extensive results, we proved that our propositions improve aerial network performance far above the conventional aerial network architectures that have been investigated thus far; such conventional architectures are heavily-based on LTE resource allocation schemes which hamper the full potentials of UAVs.

We then highlighted the importance of CNs with respect to E2E slicing and connectivity. Hence, we investigated three proposed CN design configurations, which concerned themselves with different ways of isolating slice management to NF instances across the CP and UP, and benchmarked their latency performance against the state-of-the-art configuration utilized today. We formulated an ILP problem to minimze the operational latency of the CN and proved that to better support the ERLLC latency requirements, which are far more stringent that of URLLC in 5G networks, the CN should not mix and match the managing of different slices under an NF instance. Additionally, any physical server should not host both CP and UP NF instances, simultaneously, thus forcing each plane to contend with each other for computing resources. This intra-plane contention harms the CN operational latency which ultimately harms the access-end and E2E performance. Finally, we proposed two additional avenues of research: 1) Coordinated Multi-Point in Aerial Networks and 2) Integrated Sensing and Communications. We detailed the many open challenges associated with each of these areas and offered AI as a potential method of solution to handle their associated complexities.

REFERENCES

[1] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, 2016.

[2] B. Chatras, U. S. Tsang Kwong, and N. Bihannic, "NFV enabling network slicing for 5G," in *20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, 2017, pp. 219–225.

[3] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 38–45, 2017.

[4] R. Ni, X. Li, J. Chen, S. Chen, E. Wang, M. Zhu, W. Zhang, and Y. Chen, "An end-to-end demonstration for 5G network slicing," in *IEEE 89th Vehicular Technology Conference (VTC)*, 2019, pp. 1–5.

[5] P. H. A. Rezende and E. R. M. Madeira, "An adaptive network slicing for LTE radio access networks," in *Wireless Days (WD)*, 2018, pp. 68–73.

[6] S. Xiao and W. Chen, "Dynamic allocation of 5G transport network slice bandwidth based on LSTM traffic prediction," in *IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 2018, pp. 735–739.

[7] J. T. Infiesta, C. Guimarães, L. M. Contreras, and A. de la Oliva, "GANSO: Automate network slicing at the transport network interconnecting the edge," in *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2020, pp. 161–166.

[8] N. Shahriar, S. Taeb, S. R. Chowdhury, M. Zulfiqar, M. Tornatore, R. Boutaba, J. Mitra, and M. Hemmati, "Reliable slicing of 5G transport networks with bandwidth squeezing and multi-path provisioning," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1418–1431, 2020.

[9] W. Lee, T. Na, and J. Kim, "How to create a network slice? - a 5G core network perspective," in *21st International Conference on Advanced Communication Technology (ICACT)*, 2019, pp. 616–619.

[10] G. Wang, G. Feng, S. Qin, R. Wen, and S. Sun, "Optimizing network slice dimensioning via resource pricing," *IEEE Access*, vol. 7, pp. 30 331–30 343, 2019.

[11] G. Wang, G. Feng, W. Tan, S. Qin, R. Wen, and S. Sun, "Resource allocation for network slices in 5G with network resource pricing," in *IEEE Global Communications Conference (GLOBECOM)*, 2017, pp. 1–6.

[12] E. E. Tsiropoulou, G. K. Katsinis, and S. Papavassiliou, "Distributed uplink power control in multiservice wireless networks via a game theoretic approach with convex pricing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 1, pp. 61–68, 2012.

[13] I. da Silva, G. Mildh, A. Kaloxylos, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann, and N. Bayer, "Impact of network slicing on 5G radio access networks," in *European Conference on Networks and Communications (EuCNC)*, 2016, pp. 153–157.

[14] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, 2018.

[15] B. Khodapanah, A. Awada, I. Viering, A. N. Barreto, M. Simsek, and G. Fettweis, "Slice management in radio access network via iterative adaptation," in *IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[16] M. O. Ojijo and O. E. Falowo, "A survey on slice admission control strategies and optimization schemes in 5G network," *IEEE Access*, vol. 8, pp. 14 977–14 990, 2020.

[17] Q. Liu, T. Han, and N. Ansari, "Energy-efficient on-demand resource provisioning in cloud radio access networks," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 4, pp. 1142–1151, 2019.

[18] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5G wireless networks," in *IEEE Conference on Computer Communications (INFOCOM)*, 2018, pp. 1970–1978.

[19] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8036–8049, 2019.

[20] Q. Zhang, W. Saad, M. Bennis, X. Lu, M. Debbah, and W. Zuo, "Predictive deployment of UAV base stations in wireless networks: Machine learning meets contract theory," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 637–652, 2021.

[21] I. Purnomo, A. A. Muayyadi, and D. M. Saputri, "Numerology effect on 5G 28 Ghz communication system performance," in *2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2020, pp. 332–337.

[22] A. Yazar and H. Arslan, "Reliability enhancement in multi-numerology-based 5G new radio using INI-aware scheduling," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 110, May 2019. [Online]. Available: https://doi.org/10.1186/s13638-019-1435-z

[23] S. Wei, T. Li, and W. Wu, "Load optimization of joint user association and dynamic TDD in ultra dense networks," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020, pp. 545–550.

[24] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "User association and load balancing for cellular massive MIMO," in *2014 Information Theory and Applications Workshop (ITA)*, 2014, pp. 1–10.

[25] J. Flordelis, F. Rusek, F. Tufvesson, E. G. Larsson, and O. Edfors, "Massive MIMO performance—TDD versus FDD: What do measurements say?" *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2247–2261, 2018.

[26] E. Zeydan, O. Dedeoglu, and Y. Turk, "Experimental evaluations of TDD-based massive MIMO deployment for mobile network operators," *IEEE Access*, vol. 8, pp. 33 202–33 214, 2020.

[27] A. Sheikhi, S. M. Razavizadeh, and I. Lee, "A comparison of TDD and FDD massive MIMO systems against smart jamming," *IEEE Access*, vol. 8, pp. 72 068–72 077, 2020.

[28] S. Lagen, B. Bojovic, S. Goyal, L. Giupponi, and J. Mangues-Bafalluy, "Subband configuration optimization for multiplexing of numerologies in 5G TDD New Radio," in *IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1–7.

[29] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G New Radio numerologies and their impact on the end-to-end latency," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2018, pp. 1–6.

[30] V. N. Ha, T. T. Nguyen, L. B. Le, and J. Frigon, "Admission control and network slicing for multi-numerology 5G wireless networks," *IEEE Networking Letters*, vol. 2, no. 1, pp. 5–9, 2020.

[31] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, and P. Zhang, "Machine learning based flexible transmission time interval scheduling for EMBB and URLLC coexistence scenario," *IEEE Access*, vol. 7, pp. 65 811–65 820, 2019.

[32] L. Diez, A. Garcia-Saavedra, V. Valls, X. Li, X. Costa-Perez, and R. Agüero, "LaSR: A supple multi-connectivity scheduler for multi-RAT OFDMA systems," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 624–639, 2020.

[33] X. Sun, N. Ansari, and R. Fierro, "Jointly optimized 3D drone mounted base station deployment and user association in drone assisted mobile access networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2195–2203, 2020.

[34] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, 2014.

[35] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 434–437, 2017.

[36] P. Yang, X. Xi, K. Guo, T. Q. S. Quek, J. Chen, and X. Cao, "Proactive UAV network slicing for URLLC and mobile broadband service multiplexing," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2021.

[37] J. Gui, N. Jin, and X. Deng, "Performance optimization in UAV-Assisted wireless powered mmWave networks for emergency communications," *Wireless Communications and Mobile Computing*, vol. 2021, p. 9936837, Jun 2021. [Online]. Available: https://doi.org/10.1155/2021/9936837

[38] Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, and A. L. Swindlehurst, "A comprehensive overview on 5G-and-beyond networks with UAVs: From communications to sensing and intelligence," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2021.

[39] C.-W. Hsu, Y.-L. Hsu, and H.-Y. Wei, "Energy-efficient edge offloading in heterogeneous industrial IoT networks for factory of future," *IEEE Access*, vol. 8, pp. 183 035–183 050, 2020.

[40] T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3192–3206, 2021.

[41] N. Ansari, Q. Fan, X. Sun, and L. Zhang, "SoarNet," *Wireless Commun.*, vol. 26, no. 6, p. 37–43, Dec. 2019. [Online]. Available: https://doi.org/10.1109/MWC.001.1900126

[42] M. A. Hossain, A. R. Hossain, and N. Ansari, "Numerology-capable UAV-MEC for future generation massive IoT networks," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 860–23 868, 2022.

[43] Y. Yin, M. Liu, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Cross-layer resource allocation for UAV-Assisted wireless caching networks with NOMA," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3428–3438, 2021.

[44] H. Tadayyoni, M. H. Ardakani, A. R. Heidarpour, and M. Uysal, "Ultraviolet communications for unmanned aerial vehicle networks," *IEEE Wireless Communications Letters*, vol. 11, no. 1, pp. 178–182, 2022.

[45] D. M. Manias, A. Chouman, and A. Shami, "An NWDAF approach to 5G core network signaling traffic: Analysis and characterization," in *IEEE Global Communications Conference (GLOBECOM)*, 2022, pp. 6001–6006.

[46] A. Chouman, D. M. Manias, and A. Shami, "Towards supporting intelligence in 5G/6G core networks: NWDAF implementation and initial analysis," in *2022 International Wireless Communications and Mobile Computing (IWCMC)*, 2022, pp. 324–329.

[47] K. Du, X. Wen, L. Wang, and T.-T. Nguyen, "A cloud-native based access and mobility management function implementation in 5G core," in *IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp. 1251–1256.

[48] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, and D. Darche, "On the scalability of 5G core network: The AMF case," in *15th IEEE Annual Consumer Communications and Networking Conference (CCNC)*, 2018, pp. 1–6.

[49] D. Sattar and A. Matrawy, "Optimal slice allocation in 5G core networks," *IEEE Networking Letters*, vol. 1, no. 2, pp. 48–51, 2019.

[50] N. Salhab, R. Rahim, R. Langar, and R. Boutaba, "Offloading network data analytics function to the cloud with minimum cost and maximum utilization," in *IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[51] S. Gangakhedkar, H. Cao, A. R. Ali, K. Ganesan, M. Gharba, and J. Eichinger, "Use cases, requirements and challenges of 5G communication for industrial automation," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, 2018, pp. 1–6.

[52] S. E. Elayoubi, M. Fallgren, P. Spapis, G. Zimmermann, D. Martín-Sacristán, C. Yang, S. Jeux, P. Agyapong, L. Campoy, Y. Qi, and S. Singh, "5G service requirements and operational use cases: Analysis and mETIS II vision," in *European Conference on Networks and Communications (EuCNC)*, 2016, pp. 158–162.

[53] M. Gundall, J. Schneider, H. D. Schotten, M. Aleksy, D. Schulz, N. Franchi, N. Schwarzenberg, C. Markwart, R. Halfmann, P. Rost, D. Wübben, A. Neumann, M. Düngen, T. Neugebauer, R. Blunk, M. Kus, and J. Grießbach, "5G as enabler for industrie 4.0 use cases: Challenges and concepts," in *IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1, 2018, pp. 1401–1408.

[54] A. R. Hossain and N. Ansari, "Priority-based downlink wireless resource provisioning for radio access network slicing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9273–9281, 2021.

[55] N. A. Johansson, Y. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 1184–1189.

[56] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[57] A. Alzidaneen, A. Alsharoa, and M.-S. Alouini, "Resource and placement optimization for multiple UAVs using backhaul tethered balloons," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 543–547, 2020.

[58] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *IEEE Global Communications Conference*, 2014, pp. 2898–2904.

[59] W. Liu, L. Zhang, and N. Ansari, "Laser charging enabled DBS placement for downlink communications," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 4, pp. 3009–3018, 2021.

[60] S. Rajagopal, S. Abu-Surra, and M. Malmirchegini, "Channel feasibility for outdoor Non-Line-of-Sight mmWave mobile communication," in *IEEE Vehicular Technology Conference (VTC)*, 2012, pp. 1–6.

[61] W. Khawaja, O. Ozdemir, and I. Guvenc, "UAV air-to-ground channel characterization for mmWave systems," in *IEEE 86th Vehicular Technology Conference (VTC)*, 2017, pp. 1–5.

[62] R. Ali, Y. B. Zikria, A. K. Bashir, S. Garg, and H. S. Kim, "URLLC for 5G and beyond: Requirements, enabling incumbent technologies and network intelligence," *IEEE Access*, vol. 9, pp. 67 064–67 095, 2021.

[63] Y. Liu, Y. He, Y. Lin, and L. Tang, "Toward native artificial intelligence in 6G," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2022, pp. 1–6.

[64] M. A. Hossain, A. R. Hossain, and N. Ansari, "AI in 6G: Energy-efficient distributed machine learning for multilayer heterogeneous networks," *IEEE Network*, vol. 36, no. 6, pp. 84–91, 2022.

[65] W. Wu, C. Zhou, M. Li, H. Wu, H. Zhou, N. Zhang, X. S. Shen, and W. Zhuang, "AI-Native network slicing for 6G networks," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 96–103, 2022.

[66] M. A. Hossain, A. R. Hossain, W. Liu, N. Ansari, A. Kiani, and T. Saboorian, "A distributed collaborative learning approach in 5G+ core networks," *IEEE Network*, 2023, DOI: 10.1109/MNET.133.2200527, early access.

[67] J. Yao and N. Ansari, "Caching in dynamic IoT networks by deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3268–3275, 2021.

[68] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, 1999.

[69] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[70] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. M. O. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *CoRR*, vol. abs/1509.02971, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:16326763

[71] J. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.

[72] F. Irram, M. Ali, Z. Maqbool, F. Qamar, and J. J. Rodrigues, "Coordinated multi-point transmission in 5G and beyond heterogeneous networks," in *IEEE 23rd International Multitopic Conference (INMIC)*, 2020, pp. 1–6.