Effective Resource Allocations Based on Network Slicing in O-RAN Networks

Soriya Prum[1] **Hongcheng Ngouch**[1] **Tha Khieng**[2] **Bondeth Hun**[2] **Sa Math**[2] **Tharoeun Thap**[2]
[1]Dept. of Telecommunication and Electronics Engineering
Royal University of Phnom Penh (RUPP), Cambodia
[2]Telecommunication Regulator of Cambodia (TRC),
Ministry of Post and Telecommunications, Cambodia
thaptharoeun@trc.gov.kh

Abstract

Recent research has increasingly focused on resource allocation strategies tailored for effective use of resources to meet the diverse Quality of Service (QoS) requirements of 5G networks, especially within the Open RAN (O-RAN) architecture. This paper presents an evaluation of the E2E QoS of millimeter wave communications in 5G O-RAN networks by investigating key QoS metrics, including throughput, latency, and signal-to-interference-plus-noise ratio (SINR). The result provides critical insights for the design and optimization of resource allocations in the O-RAN networks. Accordingly, we proposed an extensible Application (xApp) that resides in the near-real-time RAN intelligent controller (near-RT RIC), tackling the allocation of resource blocks based on QoS requirements of different services and varied traffic conditions in the near real-time scale.

Keywords: *Resource allocations, Network slicing, O-RAN, RIC*

1 Introduction

The ITU-R M.2083 defined three usage scenarios for IMT-2020, such as enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency communications (uRLLC) [1]. Specifically, eMBB requires substantial bandwidth and throughput, uRLLC demands ultra-low latency and high reliability, while mMTC emphasizes efficient resource allocations for massive device connectivity. These diverse requirements significantly make resource allocations in 5G networks, particularly with the O-RAN architecture, become even more complex that involving optimizing network resources to meet various service requirements. The rapid growth of diverse applications demands more resources in the 5G networks, including compute, storage, network, and radio resources (e.g., frequency, bandwidth, and physical resource blocks) that require both flexibility and precision in the resource management process.

To address this need, network slicing has emerged as the 5G's new features, enabling technology to meet the requirements of each service with abstracted of the physical resources. A network slice acts as a virtualized network that is isolated between different services to meet its requirements. Network slicing is a logical self-contained network across the network infrastructure, including RAN, transport, and core. However, this technology also introduces challenges in terms of isolation, customization, elasticity, and end-to-end lifecycle management.

Building on the foundations of C-RAN and V-RAN, O-RAN architecture advances the RAN through four guided key principles, including disaggregation, virtualization, open interfaces, and intelligent closed-loop control, enabling openness and interoperability, eliminating vendor lock-in [2]. Central to the O-RAN architecture are two types of RAN intelligent controller (RIC), which are categorized into near-real-time (near-RT) RIC and non-real-time (non-RT) RIC. Each RIC is designed to control and optimize RAN performance on different time sensitivities based on the decision-making process. The non-RT RIC, operating on time scales above 10 ms, is responsible for implementing the policy enforcement and training of AI/ML models. The near-RT RIC operates in the 1 ms to 10 ms time range and executes near-real-time control functions guided by the policies in the non-RT RIC through the A1 interface. Together, these components enable intelligent and flexible resource allocations

tailored to the QoS requirements of diverse network slices in O-RAN environments.

This paper presents a comprehensive E2E QoS analysis of mmWave communications within a 5G O-RAN network. Key performance metrics, including throughput, latency, and SINR, are evaluated to illustrate the important resource allocations tackling in the networks. We proposed a method that enables RAN to allocate the physical resource blocks (PRB) to each service with respect to their QoS, priority, and traffic conditions in near-real time. Additionally, a detailed discussion of the simulation configurations is provided to enhance the accuracy and reliability of experimental evaluations, providing deeper insights into the network performance in 5G O-RAN environments.

2 Literature Review

2.1 Resource Allocations

Recent studies in 5G network slicing have explored various resource allocation mechanisms to meet the diverse requirements across slices. Two approaches have emerged in the literature are share-based and reservation-based allocation schemes [3]. Share-based approaches assign a predetermined network share to each slice, which can be dynamically redistributed across nodes according to traffic variations. This model supports high efficiency with statistical multiplexing and is particularly suitable for elastic services, such as eMBB. However, share-based models do provide only statistical guarantees and poor isolation, thus making them less appropriate for mission-critical services requiring strict QoS compliance. Reservation-based approaches contrast and allow slices to request and pre-reserve guaranteed resources, with high isolation and hard performance guarantees. Though good for latency-critical services like URLLC, such methods have lower efficiency and higher complexity due to traffic prediction and admission control overhead.

The Xu et al. [4] survey details the 5G networks' distribution of power and bandwidth resources between various base stations. Methods like OFDMA and NOMA allow for the best possible speed with reduced interference. Systems use relays or cloud control to have more control over the resources. The mechanism is designed to cater to many users with varying

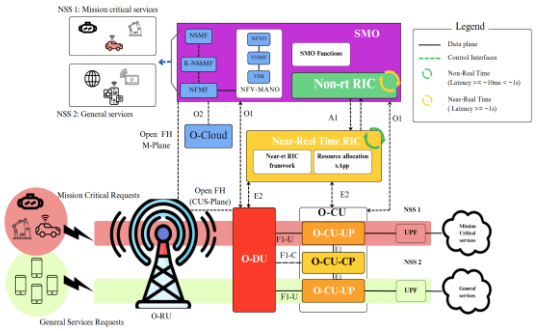


Figure 1. Slicing Architecture in O-RAN

demands. Most of the techniques are, however, elaborate, need accurate channel information, and are hard to make real-time modifications. They are also prone to being overcome by network slicing and ensuring fair service to all classes of users. These weaknesses limit their use in dynamic environments like O-RAN.

2.2 Network Slicing

Key enabling technologies of network slicing are Network function virtualization (NFV) and Software-defined networking (SDN). Their combined implementation offers greater scalability, flexibility, availability, and programming requirements while lowering both capital expenditures (CAPEX) and operational expenditures (OPEX). The decoupling of control and data plane in SDN enables the independent deployment of the traffic forwarding and decision-making. A centralized SDN controller maintains a global, logical view of the network infrastructure, optimizing traffic management, service provisioning, and resource allocation. The NFV-MANO manages the resource allocations for VNFs and the life cycle of the network slices, including the creation, updating, and termination of network functions by VM or containers [5].

In accordance to O-RAN Alliance slicing architecture [6], the near-RT RIC, and the O-CU Control Plane (O-CU-CP) are shared across multiple network slices, while each slice is assigned a dedicated O-CU User Plane (O-CU-UP) instance to ensure isolation at the user-data level as illustrated in Figure 1. As defined by ETSI NFV-MANO [7], NFV Orchestrator (NFVO) can reside in the Service Management Orchestrations (SMO) responsible for providing FCOM functionality, termination of O2 IMS, and

Table 1. Proposed Algorithm Scheme

1:	**SET** NSS 1 = mission-critical services
2:	**SET** NSS 2 = general services
3:	**Phase 1:** Resource Allocation
4:	Preconfigure resources for slices, RB_{total}
5:	**If** Slice = NSS 1 **then**
6:	**SET** RB_{crit} && Latency < 100 ms && Priority level = "high"
7:	**else**
8:	**SET** $RB_{general}$ && Latency < 300 ms && Priority level = "low"
9:	**end if**
10:	**Phase 2:** Operation
11:	Clustering U = {U_{crit}, U_{gen}}
12:	**If** NSS 1's traffic > threshold, **then**
13:	$RB_{crit\_congest} = (U_{crit} \times RB_{ms}) + RB_{crit}$
14:	**else** NSS 2's traffic > threshold
15:	$RB_{gen\_congest} = (N_{gen} \times RB_{gen}) + RB_{general}$
16:	**end if**

potentially NFO and O2 dms support, while also offering some element management capabilities. The Virtual Network Function Manager (VNFM) also maps to SMO, handling NFO-related operations and potentially terminating parts of the O2 DMS interface. The Virtualized Infrastructure Manager (VIM) maps to O-Cloud and provides IMS and DMS functionalities for managing VM-based virtualized resources. In O-RAN, network slice components are followed by 3GPP [8], including the Network Slice Management Function (NSMF), Network Slice Subnet Management Function (NSSMF), and Network Function Management Function (NFMF). NSMF interprets network slice requirements into network slice subnet requirements and distributes those requirements to the NSSMF. NSSMF is responsible for the management and orchestration of the network slice subnet instances (NSSIs). Connected to the NFV-MANO, they can dynamically allocate resources to network functions. Finally, NFMF is responsible for managing the lifecycle and configuration of individual network functions (such as VNFs and PNFs) involved in the network slice.

3 Proposed Scheme

This paper proposed a resource allocation based on network slicing, which is controlled by the xApp residing in the near-rt RIC of the O-RAN architecture. Two types of services are being sliced, one for mission-critical services (e.g., MCPTT voice, MCPTT video, and MCPTT data) and the other for general applications (e.g., internet traffic, streaming services). The purpose of this work is to allocate dedicated PRBs to each type of application in isolation to meet its QoS requirements, including throughput and latency. The network slice for critical applications is resource prioritized, providing lower latency, while the general services slice is given with lower resource priority with best-effort in terms of latency.

The resource management is divided into two critical phases: (i) Resource allocations and (ii) Operations, where the resource allocations xApp dynamically adapts resource blocks distribution in response to near real-time network conditions and traffic demands. The system process is illustrated in Table 1.

3.1 Resource Allocations

Both slices operate on a shared the same frequency spectrum and are provided with a pre-configured number of resource blocks. At the O-DU level, each slice is assigned a MAC scheduler for resource blocks allocations. The resource blocks are being distributed in terms of bandwidth in a normalized traffic condition, where 20 MHz bandwidth is preconfigured for the mission-critical services slice and 30 MHz is pre-allocated for the general services.

At a higher level, which is the near-rt RIC, the xApp is a container designed to perform resource allocation per slice to meet the QoS of each service in the network. The xApp monitors and collects the data from E2 nodes (e.g., O-CU and O-DU) and O-RU through E2, which is a logical interface in O-RAN. The A1 policy, which defined the optimization objectives such as latency and throughput target, guided the execution of the xApp through O1 interface.

3.2 Operations

During runtime, UEs are clustered into two different types based on their service requested characteristics: mission critical service users and general service users. UEs with similar service types are mapped to its VM, hosting a dedicated service, schedule with pre-allocated resource blocks. In case of a congestion event in both slices, the xApp is triggered and modifies the

update of the E2 performance through near-rt RIC guided by the A1 policy from the non-rt RIC. In addition, the number of resource blocks required to be reallocated is calculated in the accounts of the total number of users and the resource blocks needed for each user, in addition to the allocated resources to each service respectively. For instance, if the critical slice experiences overload, it temporarily borrows the reserve resource blocks to uphold its QoS guarantees. Once the network conditions are normalized, it releases those borrowed resources and restores the original resource allocation scheme.

Let RB_{total} denotes the total of resource blocks, $RB_{reserved}$ represent the reserved resource block, RB_{crit} as the resource block allocated to mission-critical services, and $RB_{general}$ for the resource block of general service. The RB_{total} can be expressed as:

$$RB_{total} = RB_{reserved} + RB_{crit} + RB_{general} \quad (1)$$

Let RB_{avail_cri} denote 80% of the total available resource block reserved for mission-critical services, and RB_{avial_gen} represents 20% of the total available resource block reserved for general services. The equation for resource allocation based on traffic conditions can be expressed as:

$$RB_{reserved} = RB_{avail\_cri} + RB_{avial\_gen} \quad (2)$$

$$RB_{avail\_cri} = RB_{reserved} \times \frac{80}{100\%} \quad (3)$$

$$RB_{avail\_gen} = RB_{reserved} \times \frac{20}{100\%} \quad (4)$$

Using (3) and (4), the $RB_{reserved}$ can be calculated as:

$$RB_{reserved} = (RB_{reserved} \times \frac{80}{100\%}) + (RB_{reserved} \times \frac{20}{100\%}) \quad (5)$$

Let $RB_{crit\_congest}$ be the total number of resource blocks required for U mission-critical users, and $RB_{gen\_congest}$ denotes the total number of resource blocks required for N general service users. The equations can be expressed as:

$$RB_{crit\_congest} = (U_{crit} \times RB_{ms}) + RB_{crit} \quad (6)$$

Table 2. Simulation Parameters

Parameters	Value
Center frequency	30 GHz
Bandwidth	80 MHz
Numerology	2
BS Tx power	40 dBm
Antenna height	BS = 25 m, UE = 1.5 m
MIMO antenna	BS = 8×8, UE = 4× 4
Antenna gain	BS = 8 dBi, UE = 5 dBi
Noise Fig.	BS = 7 dB, UE = 10 dB
Pathloss model	Umi_Streetcanyon
Channel condition	Line of sight (LoS)
Peak data rate	20 Gbps
Max UDP packet	10, 000
UDP packet size	1024 bytes

$$RB_{gen\_congest} = (N_{gen} \times RB_{gen}) + RB_{general} \quad (7)$$

where U_{crit} is the number of mission-critical users, RB_{ms} is the number of resource blocks required for each mission-critical user, N_{gen} is the number of mission-critical users, RB_{gen} is the number of resource blocks required for each general user.

4 Simulation and Results

The simulation is conducted using the open-source ns-O-RAN Open RAN simulation platform, which enables large-scale 5G network simulations by incorporating 3GPP-compliant channel models and a detailed implementation of the full 3GPP RAN protocol stack within the ns-3 framework**,** augmented with an O-RAN-compliant E2 interface [9].

4.1 Simulation Setup

Figure 2 illustrates a simulation scenario of an urban environment with two gNBs positioned at the height of 25 meters and six user equipment's (UEs), each at the height of 1.5 meters, moving randomly within a 1000 square meters Line-of-sight (LOS) region at the speed ranging from 1 m/s to 9 m/s. The scenario is designed by the Dense Urban-eMBB test environment as specified in ITU-R M.2412 [10] and is illustrated in Table 2, which defines configuration

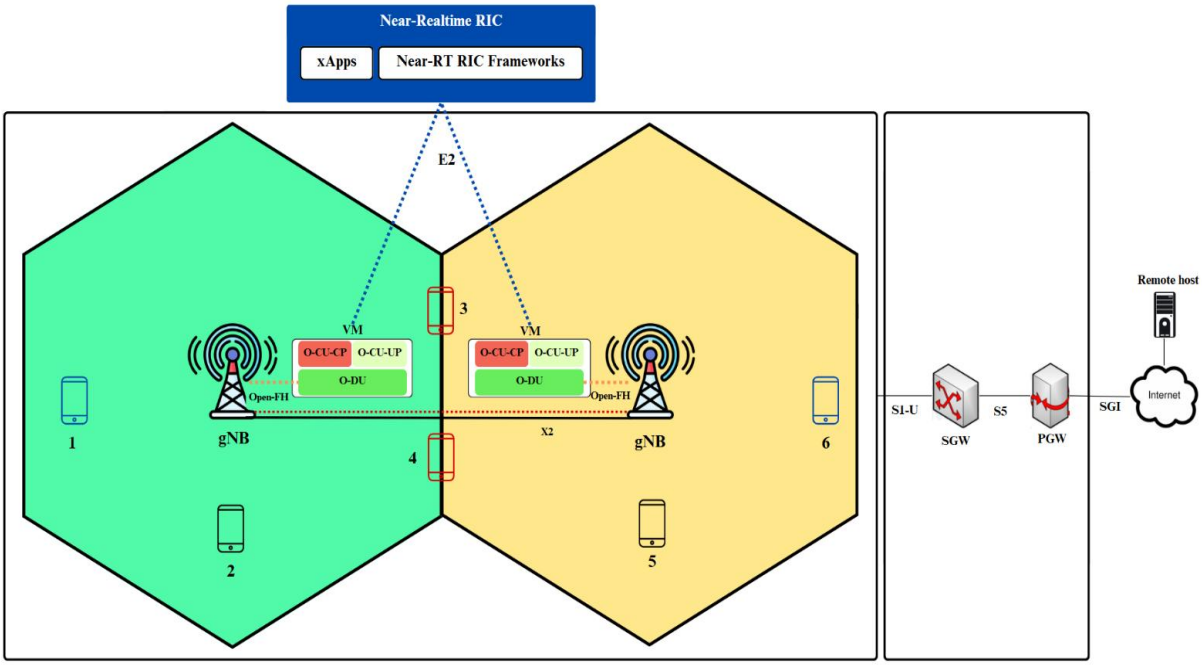


Figure 2. Simulation Scenario

parameters for this simulation. Additionally, the base station dynamically detects and manages HO based on an adaptive signal-to-noise ratio (SNR) threshold, initiating the process when the SNR drops below -5 dB. Finally, the experiment is conducted through a simulation spanning a total duration of 120 seconds.

4.2 Performance Metrics

To evaluate the impact of mmWave communications within the O-RAN framework, this study investigates key QoS metrics, including throughput, latency, and signal-to-interference-plus-noise ratio (SINR).

- Throughput refers to the rate at which information is correctly delivered and received via a connection in a period. It can usually be measured in megabits per second (Mbps), depending on the manner and extent to which the network is utilized. Throughput is the performance seen by end users and applications, and hence is one of the most important metrics of end-to-end network quality of service and efficiency. Throughput can be affected by numerous factors along the transmission path. They are channel conditions, i.e., signal strength, noise levels, and interference from other users or proximate cells, which will degrade signal quality and affect successful data transfer. Such other factors as packet loss, i.e., packet loss due to congestion or transmission, and retransmissions, bandwidth-consuming, and delay, thus reducing the effective throughput.
- Latency also known as delay, is the time it takes for a packet of data to travel from the source (e.g., remote host) to reach the destination (e.g., mobile users), typically measured in milliseconds. Mean delay per user device is an indicative measure and can be used to quantify network performance.
- Signal-to-interference-plus-noise-ratio (SINR) is a performance metric used to evaluate the quality of a wireless communication link. The higher the SINR, the clearer, better, and more dependable the signal will be, and therefore, the increased data rate and decreased error rates are obtained. In contrast, unfavorable channel conditions signaled by low SINR can cause packet losses, retransmissions, low data rates, or even trigger handovers.

4.3 Results

Figure 3 presents the throughput performance of all six UEs under the simulated mmWave communications in a 5G O-RAN scenario. The mean throughput, measured in megabits per second (Mbps), is plotted against the individual UEs. The results show that UE 4 and UE 6 achieved the highest throughput, receiving

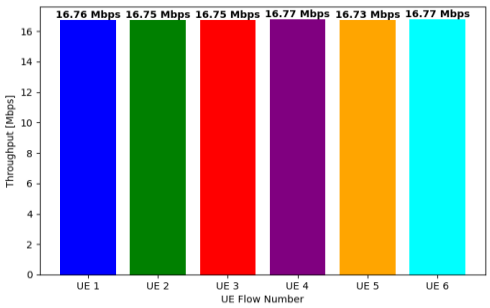


Figure 3. Comparison of Average Throughput

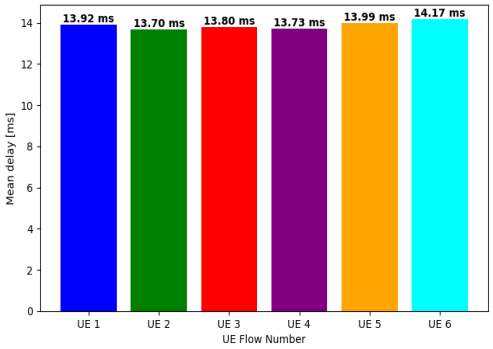


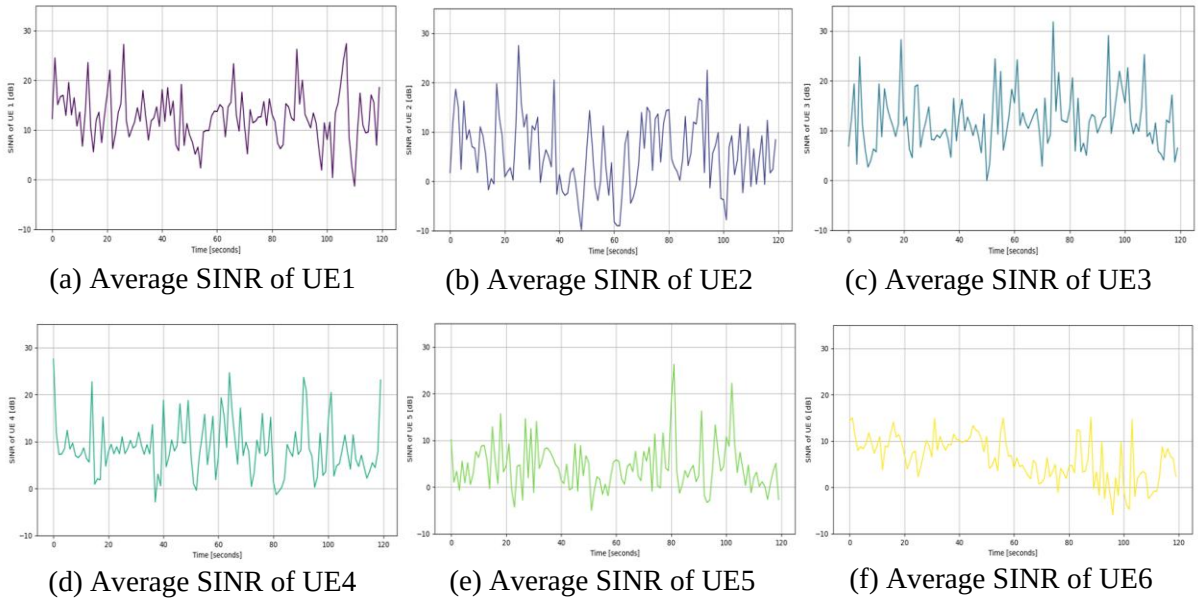
Figure 4. Comparison of Average Delay

packets from the remote host at 16.77 Mbps. UE 1 followed closely with a throughput of 16.76 Mbps. UE 2 and UE 3 recorded a slightly reduced throughput of 16.75 Mbps, while UE 5 experienced the lowest throughput at 16.73 Mbps, potentially attributable to link instability, network congestion, or channel coding.

Figure 4 illustrates the end-to-end delay comparison across all UEs. The mean delay, expressed in milliseconds (ms), reveals that UE 2 experienced the lowest delay at 13.70 ms, with UE 3 and UE 4 closely trailing at 13.80 ms and 13.73 ms, respectively. UE 1 displayed a marginally higher delay of 13.92 ms, while UE 5 recorded a delay of 13.99 ms. UE 6 exhibited the highest mean delay at 14.17 ms, suggesting momentary congestion or less favourable channel conditions during the transmission interval.

Figure 5 indicates the SINR for all UEs over a 120-second simulation period, where values are sampled at 1-second intervals. UE 1 recorded the highest average SINR at 12.61 dB, but its SINR pattern revealed pronounced fluctuations between 1 dB and 27 dB, particularly during the final 40 seconds, suggesting transient degradation likely driven by mobility-induced channel variation or degraded propagation conditions. UE 3 followed closely with an average SINR of 11.99 dB with values oscillating between 0 dB to 31 dB and showed a relatively smoother signal trajectory, with intermittent enhancements possibly reflecting the near distance between the base station and the UE, favorable link reestablishment or beam alignment following mobility events. UE 4, with an average of 8.66 dB, exhibited dense oscillations throughout the simulation from -2 dB to 27 dB, maintaining a consistent but moderately turbulent channel quality, which may indicate frequent transitions between coverage zones due to user mobility. In contrast, UE 2 displayed a lower average SINR of 5.55 dB and experienced persistent dips and variations from -9 dB to 27 dB, suggesting it encountered degraded coverage areas, unfavorable positioning, ineffective beamforming, or suboptimal handover performance. UE 6 averaged 6.27 dB and demonstrated a more gradual decline in SINR over time, punctuated by brief recoveries with the highest SINR value of 15 dB, indicative of partial adaptation to the dynamic environment. Radio resource utilization and favorable signal conditions. UE 5 registered the lowest average SINR performance at 4.40 dB, characterized by severe and frequent signal degradation from -4 dB to 26 dB, which maybe attributed to unfavorable positioning, ineffective beamforming, or repeated handover failures.

The overall performance of the simulated scenario reveals the measurements of QoS metrics mmWave communications within the 5G O-RAN, including throughput, latency, and signal quality. UEs located closer to their serving gNBs or in the beamforming path demonstrated minimal delay, underscoring the advantages of stable LOS conditions and reduced handover frequency. In contrast, UEs that experience lower SINR profiles experienced noticeable performance degradation. Notably, some UEs with moderate SINR values performed better in latency metrics than those with higher SINR, indicating that radio link quality alone does not dictate end-to-end performance. Instead, transient factors such as handover interruptions, resource scheduling delays, and varying channel conditions also play a crucial role and



(a) Average SINR of UE1 (b) Average SINR of UE2 (c) Average SINR of UE3

(d) Average SINR of UE4 (e) Average SINR of UE5 (f) Average SINR of UE6

Figure 5. Average SINR across UEs Over Time

significantly affect the perceived user experience. These conditions suggest that ensuring QoS in O-RAN environments requires not only managing radio conditions but also implementing adaptive resource allocation strategies that respond dynamically to real-time network behavior.

5 Conclusion

In this paper, a comprehensive analysis of end-to-end QoS for mmWave communications in a 5G O-RAN system was conducted, with multiple performance metrics under a dense urban deployment scenario. The simulation results reveal that factors such as handover frequency, link stability, and scheduling delays also have a substantial impact on overall performance, reinforcing the need for real-time channel adaptation and dynamic resource allocation to maintain service continuity. Thus, the future work will focus on developing a slice-aware xApp operating within the near-real-time RAN Intelligent Controller (Near-RT RIC), designed to dynamically allocate resource blocks based on the QoS requirements of different service types and real-time traffic conditions. This approach aims to enhance adaptability and resource efficiency in the O-RAN networks.

Ackowledgement

This research was supported by the Ministry of Post and Telecommunications, Kingdom of Cambodia.

References

[1] M Series, "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," Recommendation ITU-R M.2083-0, ITU Recommendation Sector, Geneva, Sep. 2015.

[2] M. Polese, L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research Challenges," in IEEE Communications Surveys & Tutorials, vol. 25, no. 2, pp. 1376-1411, Secondquarter 2023, doi: 10.1109/COMST.2023.3239220.

[3] A. Banchs, G. de Veciana, V. Sciancalepore and X. Costa-Perez, "Resource allocation for network slicing in mobile networks," in IEEE Access, vol. 8, pp. 214696-214706, 2020, doi: 10.1109/ACCESS.2020.3040949.

[4] Y. Xu, G. Gui, H. Gacanin and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," in IEEE Communications Surveys & Tutorials, vol.

23, no. 2, pp. 668-695, Secondquarter 2021, doi: 10.1109/COMST.2021.3059896.
[5] R. Su et al., "Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models," in IEEE Network, vol. 33, no. 6, pp. 172-179, Nov.-Dec. 2019, doi: 10.1109/MNET.2019.1900024.
[6] O-RAN Working Group 1, “Slicing architecture 14.00,” O-RAN, Alfter, Germany, document O-RAN.WG1. Slicing-Architecture-v14.00 Technical Specification, 2025.
[7] ETSI GR NFV-IFA 046. Network Functions Virtualisation (NFV) Release 5; architectural framework; report on NFV support for virtualisation of RAN, May 2023. Available: https://encr.pw/gv4bf.
[8] 3GPP, Management and Orchestration, Architecture framwork, Standard TS 28.533, Version 19.1.0, 3rd Generation Partnership Project (3GPP), March 2025.
[9] A. Lacava, M. Polese, R. Sivaraj, R. Soundrarajan, B.S. Bhati, T. Singh, T. Zugno, F. Cuomo, and T. Melodia, "Programmable and customized intelligence for traffic steering in 5G networks using open RAN architectures," arXiv:2209.14171 [cs.NI], pp. 1-15, October 2022. Available: https://arxiv.org/pdf/2209.14171v3.
[10] M Series, "Guidelines for evaluation of radio interface technologies for IMT-2020," Recommendation ITU-R M.2412-0, ITU Recommendation Sector, Geneva, Oct. 2017