Optimal Resource Allocation in 5G RAN with Co-Existing eMBB and uRLLC Applications

Aqsa Sayeed and Samaresh Bera, *Senior Member, IEEE*
Department of Computer Science and Engineering
Indian Institute of Technology Jammu
Jammu and Kashmir, India - 181221
Email: 2023rcs1015@iitjammu.ac.in, s.bera.1989@ieee.org

***Abstract*—5G and beyond networks are designed to support heterogeneous applications that are broadly categorized as eMBB with high throughput requirements, uRLLC with stringent latency and reliability requirements, and mMTC with a massive number of end devices. Among these applications, the co-existence of eMBB and uRLLC applications is practical in 5G deployments. However, achieving optimal and near real-time radio resource allocation within the 5G Radio Access Network (RAN) remains a major challenge due to the inherently contrasting Quality-of-Service (QoS) requirements of eMBB and uRLLC applications.**

This paper formulates the joint resource allocation problem for coexisting eMBB and uRLLC services as an Integer Linear Programming (ILP) model that maximizes the overall admission rate while satisfying system constraints. A cell-free multi-connectivity framework is incorporated to enhance uRLLC reliability. The model captures the Signal-to-Interference-plus-Noise Ratio (SINR) between users and serving base stations (gNBs) to enable efficient Physical Resource Block (PRB) allocation with Adaptive Modulation and Coding (AMC) as per 3GPP standards. Three allocation strategies are analyzed: (i) with PRB reservation for uRLLC, (ii) shared PRB allocation among eMBB and uRLLC, and (iii) puncturing resources allocated to eMBB users for uRLLC. Simulation results highlight the trade-offs among these schemes. PRB reservation significantly improves uRLLC admission and reliability, but leads to resource under-utilization at low uRLLC traffic loads. Whereas non-reservation allocation enhances spectral efficiency but degrades eMBB throughput and admission percentage. Puncturing eMBB resources offers better resource utilization but introduces additional overhead and system complexity.

***Index Terms*—5G RAN, Optimization, Resource reservation, Enhanced mobile broadband, Ultra-reliable and low latency communications**

I. Introduction

Fifth-generation (5G) networks are designed to support a diverse range of use cases, broadly categorized as enhanced Mobile Broadband (eMBB), ultra-Reliable and Low-Latency Communications (uRLLC), and massive Machine-Type Communications (mMTC). Among these, uRLLC services demand ultra-low latency (as low as 10 ms) and high reliability (up to 99.99999%) [1], [2]. Whereas eMBB targets high throughput with moderate reliability to support stable connections with high data rates [3]. The coexistence of eMBB and uRLLC traffic is a natural and essential scenario in real-world deployments. Thus, effective service provisioning for these applications requires intelligent and efficient resource allocation techniques that balance resource utilization while satisfying their distinct reliability and data-rate requirements.

Achieving optimal and near real-time resource allocation at the 5G Radio Access Network (RAN) remains a significant challenge due to the fundamentally contrasting Quality-of-Service (QoS) demands of eMBB and uRLLC services [4]–[6]. Although network slicing offers a promising solution, it introduces operational complexity and may cause resource under-utilization with low request arrival rates. Moreover, inter-slice dynamic resource allocation increases system overhead and reconfiguration delays [7], which may not be suitable for uRLLC applications. Consequently, efficient resource allocation at the RAN level remains a central problem in the design and operation of 5G and beyond networks.

To address these challenges, several studies focused on enabling the coexistence of eMBB and uRLLC services. The authors in [3], [4], [8], [9] explored puncturing- and preemption-based resource allocation strategies. In [8], the resource scheduling problem is formulated as a utility maximization task under linear or convex eMBB rate loss assumptions, while ensuring uRLLC QoS satisfaction. Similarly, the authors in [4] proposed an enhanced puncturing mechanism that limits the extent of eMBB preemption to prevent significant throughput degradation, thereby maintaining eMBB performance while satisfying uRLLC requirements. The study in [3] further minimizes the eMBB rate loss resulting from uRLLC puncturing, while [9] introduced a joint preference metric-based scheduling approach for eMBB–uRLLC coexistence. In this approach, unpredictable uRLLC traffic is scheduled by preempting eMBB transmissions, which, although effective in meeting uRLLC latency requirements, leads to throughput degradation for eMBB users. Moreover, these time-domain preemption schemes can disrupt ongoing transmissions, and accurately predicting preempted eMBB traffic in multi-connectivity environments remains a challenge, limiting uRLLC reliability gains.

Alternative approaches based on network slicing are also investigated. Korrai et al. [7] proposed a slicing-based resource allocation framework to maximize overall user data rates while meeting uRLLC latency and reliability targets. Similarly, the authors in [10] introduced a hybrid reservation-based strategy for the 5G uplink to enhance eMBB–uRLLC coexistence. To overcome the limitations of both orthogonal

979-8-3315-7896-1/26/$31.00 ©2026 IEEE

and non-orthogonal multiple access schemes, they proposed allocating reserved resources for uRLLC that partially overlap with eMBB allocations. Although this approach improves uRLLC reliability, it necessitates careful resource balancing to prevent under-utilization. Furthermore, multi-connectivity is employed in [10] to enhance uRLLC reliability by diversifying transmission paths and improving link robustness.

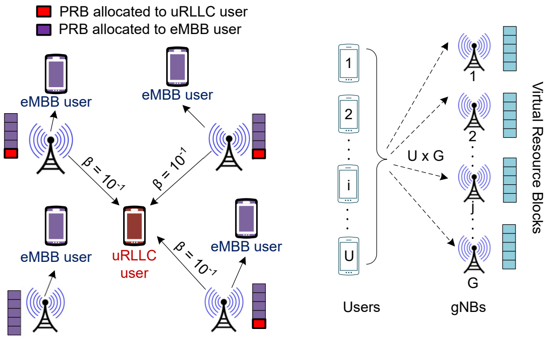
This paper focuses on a multi-connectivity-based PRB allocation strategy for uRLLC users to meet their stringent reliability requirements. Instead of relying solely on time-domain preemption, we consider multi-connectivity, as supported in a cell-free deployment scenario [5], [6], for uRLLC users. This approach leverages spatial diversity to improve uRLLC reliability while having minimal impact on ongoing eMBB transmissions. We model the scheduling problem as an integer linear programming (ILP), with the objective of maximizing admission rates of uRLLC and eMBB users, subject to PRB capacity constraints and the number of coordinating gNBs required to meet reliability targets. To enable this approach, we consider three scenarios: (a) without PRB reservation, where all users share the total resource pool; (b) with PRB reservation, where a fixed portion of resources is dedicated to uRLLC users at each gNB; and (c) puncturing resources allocated to eMBB users to admit uRLLC requests. This formulation bridges the gap between rigid slicing and dynamic preemption, providing a feasible and QoS-compliant scheduling solution that aligns with resource availability and traffic variety. The model provides fine-grained control to operators over performance trade-offs across different load conditions by allowing them to configure reservation ratios and priorities. Furthermore, this concept enables a fine-grained trade-off between the reliability of uRLLC users and the data rate of eMBB users. The simulation results demonstrate how combining multi-connectivity and PRB reservation improves user admission rates, system utilization, and data rates under varying load conditions, offering a practical direction for robust 5G RAN scheduling. The key considerations in this work are as follows:

- Diverse data rate and reliability requirements for eMBB and uRLLC requests, respectively.
- Use of a detailed mathematical model for signal-to-interference-plus-noise-ratio (SINR) calculations and adaptive modulation and coding (AMC) as per the 3GPP standard [2].
- Calculation of SINR threshold values for different modulation and coding schemes (MCSs) as per 5G NR [2] by deriving curve-fit parameters [11], [12]. We note that existing works considered SINR threshold values defined for LTE that are less efficient when compared with 5G NR. The MCS in 5G NR provides much more flexibility and improved spectral-efficiency.

The rest of the paper is organized as follows. Section II presents the detailed system model. Section III discusses the optimization problems for the three scenarios – with and without PRB reservation for uRLLC users, and puncturing eMBB users. Section IV demonstrates the trade-offs between resource reservation for uRLLC users and the data rate for eMBB users supported by simulation results. Finally, Section V concludes the work with future research directions.

II. Network Model

Figure 1 presents the network model, where multiple gNBs jointly transmit (downlink transmission) to a uRLLC user to meet its reliability requirement in terms of block error rate (BLER). The example scenario in the figure considers a multi-connectivity scenario to meet desired BLER as 10^{-3}, with the BLER of a single transmission link as 10^{-1}. On the other hand, an eMBB user is served by a single gNB. Accordingly, the figure shows the PRB allocation to uRLLC users (in red) and to eMBB users, following the orthogonal frequency-division multiple access (OFDMA) scheme. The figure (on the right) shows that, for a given uRLLC and eMBB user requests, the network operator aims to maximize the number of served requests by allocating the PRBs while considering the reliability and data-rate requirements for uRLLC and eMBB users, respectively. We discuss the parameters and the optimization problem in the subsequent sections.



(a) Multi-connectivity to meet reliability for uRLLC

(b) Problem: Maximize eMBB/ uRLLC admission with resource allocation and QoS requirements

Fig. 1. (a) System model for PRB allocation with coexisting eMBB and uRLLC users using multi-connectivity. (b) The virtual resource block represents the radio resources to be allocated to the coexisting eMBB and uRLLC users, which is combinatorial in nature.

A. Network Model

We consider a 5G RAN deployed in an area with multiple gNBs and users. Let $\mathcal{G} = \{1, 2, \ldots, N\}$ denote the set of gNBs, where each gNB $g \in \mathcal{G}$ is randomly placed in the area, and N is the total number of gNBs. The set of users in the system is represented as $\mathcal{U} = \{1, 2, \ldots, U\}$, where U is the total number of eMBB and uRLLC users in the network. Furthermore, $\mathcal{U} = \mathcal{R} \cup \mathcal{M}$, where $\mathcal{R} = \{1, 2, \ldots, R\}$ is the set of uRLLC users, and $\mathcal{M} = \{1, 2, \ldots, m\}$ is the set of eMBB users. Each user $u \in \mathcal{U}$ has a specific data rate demand δ_u and is randomly deployed across the same area as the gNBs. Additionally, each uRLLC user $r \in \mathcal{R}$ has a reliability requirement β_r in terms of the block error rate (BLER) of the transmission link. And, each eMBB user $m \in \mathcal{M}$ has a data-rate requirements.

B. SINR Computation Model

The SINR $\zeta_{u,g}$ between a user u and a gNB g is computed as:

$$\zeta_{u,g} = \frac{P_{\text{tx}}\text{PL}(d)}{N_0 + \Gamma}, \forall g \in \mathcal{G}, u \in \mathcal{U}, \tag{1}$$

where P_{tx} denotes the transmit power, which is equally distributed across all PRBs. Mathematically,

$$P_{\text{tx}} = \frac{P_{\text{tx}}^{\text{total}}}{N_{\text{PRB}}}, \tag{2}$$

where N_{PRB} is the total number of PRBs.

In (1), PL(d) denotes the path loss, which is calculated as follows [12]:

$$\text{PL}(d) = \begin{cases} G\left(\frac{\lambda}{4\pi d_0}\right)^2 \left(\frac{d_0}{d}\right)^\eta, & \text{if } d \geq d_0, \\ G\left(\frac{\lambda}{4\pi d}\right)^2, & \text{otherwise,} \end{cases} \tag{3}$$

where d is the distance between the user and the serving gNB, G is the antenna gain, λ is the carrier wavelength, d_0 is the reference distance, and η is the path loss exponent.

In (1), N_0 is the noise power and is calculated as:

$$N_0 = KTBF(1 + \text{RoT}), \tag{4}$$

where K is the Boltzmann constant, T is the system temperature, B is the PRB bandwidth, F is the receiver noise figure, and RoT is the rise-over-thermal factor [12]. Finally, in (1), Γ is the inter-cell interference, which is considered as the aggregate received power from all other gNBs in the network except the serving one [13].

C. Reliability Model for uRLLC Users

Accordingly to the system model, each uRLLC user may be served simultaneously by multiple gNBs. This concept leverages spatial diversity to minimize the probability of service failure exceeding a desired BLER. The number of cooperating gNBs Ψ_r required by a uRLLC user $r \in \mathcal{R}$ is computed as:

$$\Psi_r = \lceil -\log_{10}(\beta_r) \rceil, \tag{5}$$

where, β_r denotes the target BLER for the uRLLC user to meet its reliability requirement. This helps to identify to determine how many gNBs needed to cooperate to meet the reliability.

D. PRB Calculations for uRLLC and eMBB Users

We consider AMC based on the obtained SINR to achieve the best spectral-efficiency [14]. We evaluate the SINR threshold values for each MCS as defined in Table 5.1.3.1-1 [2]. Table I presents the threshold SINR values required to achieve BLER of a single transmission link as 10^{-1}. We note that the threshold values are obtained by deriving curve-fit parameters associated with BLER and SINR [11], [12]. Let T_{sym} denote the number of orthogonal frequency division multiplexing (OFDM) symbols per PRB. A scaling factor τ, is used with respect to PRB time duration and symbol rate. The required number of PRBs $\Upsilon_{u,g}$ for a user-gNB pair is calculated as:

$$\Upsilon_{u,g} = \left\lceil \frac{\delta_u}{\epsilon_{u,g} T_{\text{sym}} \tau} \right\rceil, \forall u \in \mathcal{U} \text{ and } \forall g \in \mathcal{G}, \tag{6}$$

where, $\epsilon_{u,g}$ is the spectral efficiency for the used modulation and coding scheme (MCS) as per the Table 5.1.3.1-1 in [2].

TABLE I
SINR THRESHOLD VALUES WITH BLER 10^{-1} FOR DIFFERENT MCS SCHEMES

Modulation	Code rate	SINR (dB)	Efficiency (bps)
QPSK	120/1024	-5.58	0.2344
QPSK	157/1024	-4.07	0.3066
QPSK	193/1024	-3.16	0.377
QPSK	251/1024	-1.94	0.4902
QPSK	308/1024	-0.97	0.6016
QPSK	379/1024	0.16	0.7402
QPSK	449/1024	1.07	0.877
QPSK	526/1024	2.02	1.0273
QPSK	602/1024	2.89	1.1758
QPSK	679/1024	3.73	1.3262
16-QAM	340/1024	4.17	1.3281
16-QAM	378/1024	4.82	1.4766
16-QAM	434/1024	5.75	1.6953
16-QAM	490/1024	6.65	1.9141
16-QAM	553/1024	7.66	2.1602
16-QAM	616/1024	8.63	2.4063
16-QAM	658/1024	9.29	2.5703
64-QAM	466/1024	10.47	2.7305
64-QAM	517/1024	11.31	3.0293
64-QAM	567/1024	12.27	3.3223
64-QAM	616/1024	13.07	3.6094
64-QAM	666/1024	14.08	3.9023
64-QAM	719/1024	15.04	4.2129
64-QAM	772/1024	16.07	4.5234
64-QAM	822/1024	16.92	4.8164
64-QAM	873/1024	18.78	5.1152
64-QAM	910/1024	19.64	5.332
64-QAM	948/1024	20.66	5.5547

III. Optimization Problems

We formulate the optimization problem considering three scenarios – with and without PRB reservation for uRLLC users, and puncturing eMBB users. We use the following decision variables: $x_{u,g} \in \{0, 1\}$ and $y_u \in \{0, 1\}, \forall u \in \mathcal{U}, \forall g \in \mathcal{G}$. Here, $x_{u,g} = 1$, if user u is served by gNB g, and $y_u = 1$ if user u is admitted with all requirements fulfilled. Furthermore, y_u is denoted as y_r and y_m, and $x_{u,g}$ as $x_{r,g}$ and $x_{m,g}$, to represent decision variables pertaining to uRLLC and eMBB users, respectively. Additionally, $z_{m,g} \in \mathbb{Z}^+$, $\forall m \in \mathcal{M}$ and $\forall g \in \mathcal{G}$, denotes the number of PRBs punctured from eMBB user m served by gNB g to admit uRLLC requests.

A. Case A: Without PRB Reservation for uRLLC

In this case, PRBs are allocated to uRLLC and eMBB users without any restrictions. The admission priority of uRLLC users is controlled by a configurable parameter, α. We consider a set of gNBs with limited PRB capacity and two types of users, uRLLC and eMBB, each with a specific data rate and PRB requirement, as discussed in Section II. The goal is to maximize the total number of admitted users. The mathematical formulation of the problem is as follows:

$$\text{Maximize} \sum_{u\in\mathcal{U}} y_u = \alpha \sum_{r\in\mathcal{R}} y_r + \sum_{m\in\mathcal{M}} y_m, \tag{7}$$

subject to

$$y_r, y_m \in \{0,1\}, \forall r \in \mathcal{R}, \forall m \in \mathcal{M}, \tag{8a}$$

$$x_{r,g}, x_{m,g} \in \{0,1\}, \forall r \in \mathcal{R}, \forall m \in \mathcal{M}, \forall g \in \mathcal{G}, \tag{8b}$$

$$\sum_{r\in\mathcal{R}} \Upsilon_{r,g} x_{r,g} + \sum_{m\in\mathcal{M}} \Upsilon_{m,g} x_{m,g} \le C_g, \forall g \in \mathcal{G}, \tag{8c}$$

$$\sum_{g\in\mathcal{G}} x_{r,g} \ge \Psi_r y_r, \forall r \in \mathcal{R}, \tag{8d}$$

$$\sum_{g\in\mathcal{G}} x_{m,g} \ge y_m, \forall m \in \mathcal{M}. \tag{8e}$$

Eqn. (7) denotes the objective of maximizing the total number of admitted users, with a higher weight given to uRLLC users. We introduce the concept of prioritizing uRLLC requests and employ a parameter α to indicate weight regulating the relative priority of eMBB users. Eqn. (8c) ensures that the total PRB consumption at each gNB does not exceed its capacity, where $\Upsilon_{r,g}$ and $\Upsilon_{m,g}$ define the PRBs required by uRLLC and eMBB users from gNB g, respectively. The notation C_g defines the total PRB capacity of gNB $g \in \mathcal{G}$. Eqn. (8d) ensures that each admitted uRLLC user is served by the required number of gNBs, satisfying its reliability requirement. Eqn. (8e) states that if an eMBB user is admitted, it must be served by a gNB. This formulation describes a multi-constrained binary integer program, which is NP-hard in general.

B. Case B: With PRB Reservation for uRLLC

In this case, each gNB reserves a fixed proportion of its PRB budget exclusively for uRLLC traffic. The remaining PRBs are allocated exclusively to eMBB users. The goal remains the same as in the shared case, to maximize the total number of admitted users. However, uRLLC and eMBB traffic are now separated into distinct PRB pools. The optimization problem is formulated as follows:

$$\text{Maximize} \sum_{u\in\mathcal{U}} y_u = \sum_{r\in\mathcal{R}} y_r + \sum_{m\in\mathcal{M}} y_m, \tag{9}$$

$$\text{subject to} \sum_{r\in\mathcal{R}} \Upsilon_{r,g} x_{r,g} \le \sigma C_g, \forall g \in \mathcal{G}, \tag{10a}$$

$$\sum_{m\in\mathcal{M}} \Upsilon_{m,g} x_{m,g} \le (1-\sigma) C_g, \forall g \in \mathcal{G}, \tag{10b}$$

$$(8\text{b}), (8\text{a}), (8\text{d}), (8\text{e}). \tag{10c}$$

Here, $\sigma \in [0,1]$ represents the reservation factor of PRB resources at each gNB for uRLLC applications. The other variables remain the same as in the optimization problem described in III-A. Eqn. (9) denotes the objective of maximizing the total admitted users. Constraints (10a) and (10b) enforce PRB limits separately for uRLLC and eMBB traffic at each gNB. This problem structure separates uRLLC and eMBB traffic while ensuring the reliability requirements of uRLLC users, which is crucial for ultra-reliable low-latency applications.

C. Case C: Puncturing PRBs Allocated to eMBB Users

In this case, all PRBs are initially allocated to eMBB requests, similar to Section III-A, where PRBs are not reserved for uRLLC requests. Upon the arrival of uRLLC requests, the PRBs allocated to eMBB users are punctured to serve uRLLC requests instantly, ensuring that their stringent latency requirements are met. Furthermore, each uRLLC request can be served by multiple gNBs using multi-connectivity to meet its reliability requirement, as discussed earlier. Therefore, the objective of the service provider is to minimize the average loss in eMBB users' throughput while puncturing the resources allocated to eMBB requests to admit uRLLC requests. Mathematically,

$$\text{Minimize} \sum_{m\in\mathcal{M}} \sum_{g\in\mathcal{G}} z_{m,g}, \tag{11}$$

$$\text{subject to} \sum_{g\in\mathcal{G}} x_{r,g} \ge \Psi_r y_r, \quad \forall r \in \mathcal{R}, \tag{12a}$$

$$\sum_{m\in\mathcal{M}} z_{m,g} \ge \sum_{r\in\mathcal{R}} \Upsilon_{r,g} x_{r,g}, \forall g \in \mathcal{G}, \tag{12b}$$

$$z_{m,g} \le a_{m,g}, \forall m \in \mathcal{M}, \forall g \in \mathcal{G}, \tag{12c}$$

$$x_{r,g} \in \{0,1\}, \forall r \in \mathcal{R}, \forall g \in \mathcal{G}, \tag{12d}$$

$$z_{m,g} \in \mathbb{Z}^+, \forall m \in \mathcal{M}, \forall g \in \mathcal{G}, \tag{12e}$$

$$y_r = 1, \forall r \in \mathcal{R}. \tag{12f}$$

In (11), $z_{m,g}$ denotes the number of PRBs punctured from eMBB user $m \in \mathcal{M}$ served by gNB $g \in \mathcal{G}$. Eqn. (12a) ensures that the reliability requirement of each uRLLC user is satisfied through multi-connectivity. Eqn. (12b) ensures that the total number of PRBs required by uRLLC users served by a gNB $g \in \mathcal{G}$ does not exceed the number of PRBs punctured from eMBB users. Eqn. (12c) ensures that the number of PRBs punctured from any eMBB user does not exceed the number of PRBs allocated to that user. Eqn. (12d) indicates whether a uRLLC request $r \in \mathcal{R}$ is served by gNB $g \in \mathcal{G}$. Eqn. (12f) ensures that the admission of uRLLC requests is considered when puncturing resources allocated to eMBB users.

IV. Performance Evaluation

In this section, we evaluate the performance of the proposed joint ILP-based resource allocation scheme for coexisting uRLLC and eMBB users within a bounded PRB environment to understand the trade-off among the three presented scenarios. Table II summarizes the key parameters used for SINR computation, while Table III lists the main simulation parameters employed in the experimental setup. The network scenario consists of randomly distributed users and gNBs to emulate realistic deployment conditions. In Case B, a fixed fraction of PRBs from each gNB is reserved exclusively for uRLLC traffic to satisfy its stringent latency and reliability requirements, ensuring immediate service upon request arrival. The remaining PRBs are dynamically allocated to eMBB users to maximize the overall user admission rate and improve resource utilization efficiency.

TABLE II
PARAMETERS FOR SINR CALCULATION

Symbol	Description	Value (unit)
R	Boltzmann constant	1.38×10^{-23} (J/K)
T	Temperature	300 (K)
B	Bandwidth per PRB	180×10^3 (Hz)
F	Noise figure	4 (linear)
$\text{noise}_{\text{RoT}}$	Rise over thermal	2 (linear)
G	Antenna gain	2 (linear)
$P_{\text{tx}}^{\text{total}}$	Total transmit power	0.251 (W)
f_c	Carrier frequency	5×10^9 (Hz)
c	Speed of light	3×10^8 (m/s)
λ	Wavelength	0.06 (m)
d_0	Reference distance	21 (m)
η	Path loss exponent	3.7

TABLE III
SIMULATION SETTINGS

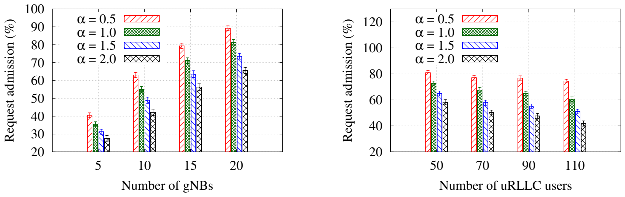
Parameter	Value / Range (unit)
Simulation area	1000 × 1000
Number of gNBs	{5, 10, 15, 20}
Bandwidth	20 MHz
Number of PRBs per gNB	100
Reservation factor (σ)	{0.05, 0.10, 0.15, 0.20}
Number of eMBB users	50 (fixed)
Number of uRLLC users	[50, 70, 90, 110]
eMBB data rate	100 kbps – 50 Mbps
uRLLC data rate	10 kbps – 100 kbps
uRLLC reliability	{1e-2, 1e-3, 1e-4, 1e-5}
Optimization problem solver	CBC [15]

A. Results and Discussions

To evaluate the performance of Case A, Case B, and Case C, we perform 50 simulation runs under varying network conditions, including different gNB densities, priority factors, and PRB reservation ratios, as applicable. We consider the following performance metrics: percentage of eMBB user admission and average data rate for eMBB users in Case A, PRB utilization percentage in Case B, and statistical measures of PRBs puncturing impact on eMBB users in Case C. In Case A, the system is analyzed using scaling factors $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$ for uRLLC request intensity. For Case B, PRB reservation factors $\sigma \in \{0.05, 0.1, 0.15, 0.2\}$ are applied to allocate fixed resource portions for uRLLC traffic. In Case C, all PRBs are initially assigned to eMBB users and then dynamically punctured to accommodate incoming uRLLC requests. The performance evaluation is conducted under two distinct simulation settings: (a) With different number of gNBs $\{5, 10, 15, 20\}$ with a random number of uRLLC users ranging between 50 and 70; (b) With different number of uRLLC users $\{50, 70, 90, 110\}$ with 15 gNBs. In both scenarios, the number of eMBB users is considered as 50 and the total number of PRBs per gNB is 100. We note that the eMBB user admission and average date rate are independent of the number of uRLLC users in Case B, i.e., with resource reservation. Hence, the results of eMBB user admission and data rate are not presented for Case B.

1) Case A: eMBB User Admission: The objective is to maximize total user admissions in the network. Figure 2(a) illustrates that eMBB admission rates initially increase with gNB density owing to better PRB availability but later decline as PRB limits and uRLLC competition intensify. Lower priority weights (α = 0.5) favor eMBB access, while higher α values prioritize uRLLC reliability, restricting eMBB admissions. We note that nearly all uRLLC users are admitted due to their higher priority.

As shown in Figure 2(b), increasing number of uRLLC users reduces eMBB admission rates, especially at higher α values (e.g., α = 2.0). Lower α enables a fair resource sharing and higher eMBB acceptance. Overall, the results highlight a trade-off between uRLLC reliability and eMBB throughput, demonstrating the effectiveness of PRB isolation for QoS assurance.



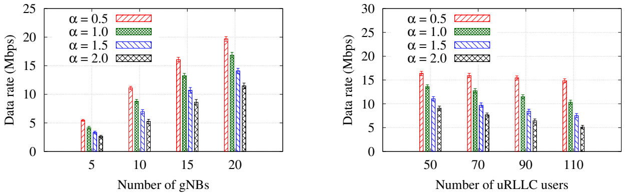
(a) With varying gNB (Number of uRLLC = 50 to 110)

(b) With varying uRLLC (Number of gNB = 15)

Fig. 2. Case A: eMBB admission percentage with different priorities (α).

2) Case A: Average Data Rate for eMBB: Figures 3(a) and 3(b) present the average data rate of admitted eMBB users in Case A under varying numbers of gNBs and uRLLC users, respectively. The data rate trend closely follows user admission behavior. Initially, the eMBB data rate increases with gNB density due to higher PRB availability and improved spatial diversity. However, beyond a certain point, it declines as resource contention and uRLLC prioritization intensify with increasing demand or priority weight (α). This reduction is further influenced by limited PRB availability and inter-cell interference at higher loads. Overall, the system achieves high throughput under dense gNB deployment and moderate uRLLC load, but performance degrades as uRLLC traffic dominates, reflecting the fundamental trade-off between reliability and throughput in shared 5G RAN environments.

3) Case B: PRB Utilization: Figures 4(a) and 4(b) show the PRB utilization for uRLLC users in Case B under varying gNB densities and uRLLC user counts. At low reservation factors (σ), reserved PRBs are efficiently used, with utilization approaching 100% at low gNB densities. As σ increases, PRBs are allocated more conservatively to meet reliability targets, causing partial under-utilization and reduced availability for eMBB users.



(a) With varying gNB (Number of uRLLC = 50 to 110)

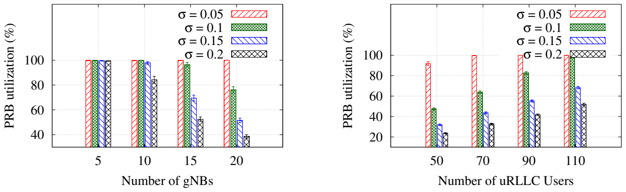
(b) With varying uRLLC (Number of gNB = 15)

Fig. 3. Case A: Average eMBB data rate without PRB reservation with different priorities (α)

TABLE IV
AVERAGE NUMBER OF PRBS PUNCTURED PER EMBB USER

#gNBs	Avg. #PRB Punctured	Std. Dev.
5	1.4876	0.428797845
10	1.1892	0.278515306
15	1.13	0.283743201
20	1.0916	0.229359

With more uRLLC users, Figure 4(b) shows that stringent reliability settings ($\sigma = 0.05$) quickly saturate PRB usage, whereas relaxed constraints ($\sigma = 0.2$) lower the PRB utilization, indicating less aggressive resource allocation. Overall, the results highlight a trade-off between reliability and resource efficiency, i.e., higher reliability demands consume more PRBs per uRLLC session, limiting spectrum availability and degrading eMBB performance.



(a) With varying gNB (Number of uRLLC = 50 to 110)

(b) With varying uRLLC (Number of gNB = 15)

Fig. 4. Case B: Percentage of PRB utilization by uRLLC requests with different PRB reservation factor (σ)

Case C: Puncturing eMBB Users: Table IV presents the average number of PRBs punctured per eMBB user with the standard deviation. It is evident that puncturing PRBs allocated to eMBB users improves resource utilization while enabling uRLLC requests to be served instantly, thereby meeting their stringent latency requirements. However, such an approach puts additional complexity on RAN resource allocation in real-time.

V. CONCLUSION

In this paper, we studied the radio resource allocation problem in 5G RAN with coexisting uRLLC and eMBB services under diverse QoS requirements. The problem is formulated as a constrained optimization model and analyzed under three scenarios: without resource reservation, with resource reservation, and with resource puncturing for uRLLC. The model incorporated adaptive modulation and coding (AMC) based on instantaneous SINR to enhance spectral efficiency. Simulation results demonstrated the trade-offs among the approaches in terms of user admission rate, PRB utilization, and eMBB throughput. While PRB reservation improves uRLLC reliability, it reduces resource utilization; non-reservation enhances efficiency but degrades eMBB performance. The puncturing approach offers a balanced compromise between reliability and throughput. Future work will explore reinforcement learning–based methods for real-time RAN resource allocation in dynamic 5G environments.

VI. ACKNOWLEDGMENT

This work has been supported by the Indian Institute of Technology Jammu, Jammu-181221, India.

REFERENCES

[1] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2020.

[2] 3GPP, "Physical layer procedures for data (3GPP TS 38.214 version 17.3.0 Release 17)," 3GPP, Tech. Rep. ETSI TS 138 214 V17.3.0, 2022.

[3] H. Sun, J. Yang, J. Su, H. Wang, and D. Liu, "Joint Resource Scheduling for Coexistence of URLLC and eMBB in 5G Wireless Networks," in *Computing, Communications and IoT Applications*, 2021, pp. 53–58.

[4] E. Engin, I. Hökelek, and H. A. Çırpan, "Resource Allocation of eMBB and URLLC Traffic using Pre-emption Mechanism," in *Int. Conf. on Telecommun. and Signal Processing (TSP)*, 2023, pp. 129–133.

[5] J. F. Monserrat, F. Bouchmal, D. Martin-Sacristan, and O. Carrasco, "Multi-Radio Dual Connectivity for 5G Small Cells Interworking," *IEEE Commun. Stand. Mag.*, vol. 4, no. 3, pp. 30–36, 2020.

[6] A. Aijaz, "Packet Duplication in Dual Connectivity Enabled 5G Wireless Networks: Overview and Challenges," *IEEE Commun. Stand. Mag.*, vol. 3, no. 3, pp. 20–28, 2019.

[7] P. K. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, and B. Ottersten, "Slicing Based Resource Allocation for Multiplexing of eMBB and URLLC Services in 5G Wireless Networks," in *IEEE CAMAD*, 2019, pp. 1–5.

[8] A. Anand, G. De Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *IEEE INFOCOM*, 2018, pp. 1970–1978.

[9] A. Pradhan and S. Das, "Joint Preference Metric for Efficient Resource Allocation in Co-Existence of eMBB and URLLC," in *COMSNETS*, 2020, pp. 897–899.

[10] S. Zhao, Y. Wang, T. Wang, and Z. Wang, "A Reservation-Based Hybrid Multiple Access Scheme for URLLC Coexisting with eMBB," in *IEEE/CIC Intl. Conf. on Communications in China*, 2021, pp. 916–921.

[11] Q. Liu, S. Zhou, and G. Giannakis, "Cross-Layer combining of adaptive Modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sep. 2004.

[12] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[13] N. Saba, L. Mela, M. U. Sheikh, J. Salo, K. Ruttik, and R. Jäntti, "Rural Macrocell Path Loss Measurements for 5G Fixed Wireless Access at 26 GHz," in *IEEE 5G World Forum (5GWF)*, 2021, pp. 328–333.

[14] R. Fantacci, D. Marabissi, D. Tarchi, and I. Habib, "Adaptive modulation and coding techniques for OFDMA systems," *IEEE Trans. Commun.*, vol. 8, no. 9, pp. 4876–4883, 2009.

[15] "COIN-OR Branch-and-Cut solver." [Online]. Available: https://coin-or.github.io/Cbc/intro.html