Multi-Connectivity for URLLC and Coexistence With eMBB in Time-Varying and Frequency-Selective Fading Channels

Govindu Sai Kesava and Neelesh B. Mehta, *Fellow, IEEE*

***Abstract*— Multi-connectivity enables a 5G cellular system to meet the challenging reliability requirements of downlink ultra-reliable low-latency communication (URLLC) data traffic. In it, multiple base stations (BSs) transmit to the URLLC user by pre-empting time-frequency resources assigned to enhanced mobile broadband (eMBB) users. We derive insightful expressions for achievability, which is the probability that the URLLC user's block error rate (BLER) requirement is met by multi-connectivity. We do so for both joint transmission (JT) and orthogonal transmission (OT) modes of URLLC for the general case in which the transmissions occur over frequency-selective channels. We then propose a low-complexity algorithm to jointly select the set of cooperating BSs and their modulation and coding schemes (MCSs) to minimize the eMBB throughput loss. For time-varying channels with feedback delays, we present an alternate stochastic reliability requirement for URLLC traffic. The MCS selected on the basis of this requirement has a markedly higher probability of meeting the BLER target over the grid of URLLC user locations. Our results highlight the different trade-offs between URLLC achievability, eMBB throughput loss, and channel state information feedback overhead of OT and JT. They bring out the significant impact of feedback delays even at moderate Doppler spreads.**

***Index Terms*— URLLC, multi-connectivity, reliability, 5G, modulation and coding, feedback, time-varying.**

I. Introduction

5G SERVES a diverse set of use cases, namely, enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC) [2, Ch. 1]. eMBB offers high data rates for services such as 8K streaming, mMTC enables a large number of devices to be connected, and URLLC enables new applications such as factory automation, telesurgery, and autonomous driving.

URLLC stands out because of its unique and challenging requirements, which include block error rates (BLERs) as low as 10^{-6} and latencies as small as 1 ms [3]. New techniques and designs are essential for satisfying these critical requirements. To reduce latency, the slot duration for the URLLC data is reduced to two or more orthogonal frequency division multiplexing (OFDM) symbols. The base station (BS), which is referred to as eNB in 4G and gNB in 5G, immediately transmits by pre-empting or superposing a part of the eMBB data or stopping the eMBB transmission altogether. It also transmits a pre-emption indicator on the control channel to inform the eMBB user about the resource elements affected [2, Ch. 10].

Multi-connectivity is a technique to improve reliability. In the downlink, multiple BSs transmit data to the URLLC user. Its architectures include load balancing, packet duplication, and packet splitting [4]. We focus on packet duplication and the downlink, in which the BSs transmit the same information to improve reliability. The cooperating BSs can do so using either orthogonal transmission (OT) or joint transmission (JT) [5]. In OT, the BSs transmit on orthogonal frequency resources and can use different modulation and coding schemes (MCSs). The transmission is successful if the URLLC user decodes at least one of these transmissions. In JT, the BSs transmit on the same time and frequency resources using a common MCS and employ maximal ratio transmission (MRT). As a result, the signal-to-noise ratio (SNR), which is the sum of individual SNRs from the different BSs, improves.

The versatile 5G standard allows multi-connectivity to be implemented in the stand-alone mode, in which gNBs cooperate, or in the non-stand-alone mode, in which eNBs and gNBs cooperate. Coordinated multi-point (CoMP), which has been adopted in earlier releases of 3GPP, is a version of multi-connectivity [6]. Another version is dual connectivity, which was standardized in Release 16. In it, the traffic is duplicated at the packet data convergence protocol (PDCP) layer [7]. Multi-connectivity is also compatible with the open radio access network's (O-RAN) disaggregated Split 7.x architecture, in which the medium access control (MAC) and physical layer tasks are split across the centralized unit (CU), distributed unit (DU), and radio unit (RU). The RU carries out only physical layer and radio-frequency tasks, while the DU and CU do the rest [4]. Layers above and including the MAC layer are common to all the links. We note that multi-connectivity can be implemented in other splits as well.

A. Related Literature on Multi-Connectivity for Downlink

In [3], the trade-off between reliability and latency for URLLC and the principles for building access protocols

Manuscript received 13 January 2022; revised 2 August 2022 and 15 October 2022; accepted 31 October 2022. Date of publication 10 November 2022; date of current version 12 June 2023. This work was supported in part by the "Next Generation Wireless Research and Standardization on 5G and Beyond" Project though the Ministry of Electronics and Information Technology, Government of India. An earlier version of this paper was presented in part at the IEEE International Conference on Communications (ICC), May 2022 [DOI: 10.1109/ICC45855.2022.9839204]. The associate editor coordinating the review of this article and approving it for publication was G. C. Ferrante. *(Corresponding author: Neelesh B. Mehta.)*

The authors are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru 560012, India (e-mail: govindukesava@gmail.com; nbmehta@iisc.ac.in).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3219730.

Digital Object Identifier 10.1109/TWC.2022.3219730

are studied. The outage probability and throughput of multi-connectivity are characterized in [8]. Also studied are various combining algorithms at the receiver. In [9], tools required to address issues related to tail, risk, and scale of URLLC are discussed. A link management scheme based on multi-connectivity and OT that optimizes the number of links to be connected to the URLLC user is proposed in [10]. In [11], outage probability reduction and the increase in resource usage due to multi-connectivity are studied. In [5], a stable-matching algorithm that matches the available resources to multiple URLLC users is proposed, and single connectivity is compared with JT and a variant of OT. The effect of shadowing on inter-frequency and intra-frequency multi-connectivity is studied in [12].

In [13], different options for pre-empting the resources assigned to the eMBB users are studied. In [14], scheduling based on linear, convex, and threshold throughput loss models for eMBB-URLLC traffic is studied. In [15], deep learning techniques are used to allocate resources to eMBB and URLLC traffic. In [16] and [17], URLLC-eMBB scheduling and resource management to maximize the eMBB throughput are studied. In [18], a scheduling scheme to place URLLC data on eMBB resources that predicts the pre-empted eMBB user's decoding probability is proposed. In [19], resources are allocated to the URLLC data taking into consideration the increase in the error probability of the eMBB transmissions.

B. Fundamental System Design Issues and Our Contributions

Several system design challenges, which are common to JT and OT, need to be addressed while designing multi-connectivity. First, despite the tight error requirement, a BS cannot conservatively choose the lowest rate MCS for URLLC users because this leads to an excessive number of eMBB resource elements being pre-empted. Therefore, the subset of BSs that transmit the URLLC packet and the MCSs they employ must be jointly chosen to minimize the eMBB throughput loss. Second, in 5G, the same MCS must be used by the BS on all the subbands on which it transmits to a URLLC user. This is mandated in order to limit downlink control channel overhead [2, Ch. 10]. This leads to the challenging problem of choosing one MCS in frequency-selective fading despite the SNRs of different subbands being different.

Third, choosing the MCS to transmit to the URLLC user as soon as its packet arrives at the BSs requires the availability of timely channel state information (CSI) at the BSs. However, requesting CSI from the URLLC user on an as-needed basis leads to unacceptably large latencies. Thus, the CSI needs to be fed back periodically by the user. However, the CSI can now be partially outdated by the time it is used for transmission. Lastly, in time-varying, frequency-selective channels with non-negligible feedback delays, the BSs cannot know the instantaneous SNRs at the time of transmission. Thus, an instantaneous BLER target, which has been assumed widely in the URLLC literature, cannot be met.

Contributions: We present a comprehensive treatment of multi-connectivity for URLLC that addresses the above issues. We consider both OT and JT, and time-varying channels with negligible and non-negligible feedback delays.

- We formulate the optimization problem of jointly selecting the subset of BSs that cooperate and their MCSs to minimize the sum throughput loss of the eMBB users while satisfying the URLLC reliability constraint. We derive an expression for achievability, which is the probability that the above problem has a feasible solution, i.e., the URLLC constraint can be satisfied by at least one choice of the subset of BSs and their MCSs. We do so for both OT and JT. Achievability is a fundamental and a more fine-grained performance metric than coverage or outage probability to assess the suitability of a given deployment of BSs for supporting URLLC. We propose a low-complexity multi-connectivity MCS selection algorithm (MCMSA) that solves the above problem with fewer computations than exhaustive search. It does so by eliminating several possibilities without incurring a loss in optimality.
- For a system with non-negligible feedback delays, we propose an alternate stochastic reliability constraint. It requires the probability of the BLER being below a target value, given the CSI fed back, to be more than a threshold. We present a novel analytically tractable expression for this conditional probability. It leads to a higher probability of meeting the BLER target than the conventional approach that ignores feedback delays.
- For the many possible design choices, we evaluate and compare URLLC achievability and eMBB sum throughput loss. Our numerical results show that JT outperforms OT in both respects, but requires more CSI feedback. Our results bring out the trade-off between achievability and throughput loss, and the significant impact of feedback delays even at moderate Doppler spreads.

Comparison With Literature: Our work differs from the literature in multiple respects. This is summarized concisely in Table I. First, we do not assume a linear loss model, in which the eMBB throughput loss is assumed to be directly proportional to the number of pre-empted resources [15], [17], [19], [21]. This model does not take into account the non-linear and not necessarily convex increase in the error rates when more resource elements are pre-empted. In [13], the throughput loss is calculated using the mean mutual information per bit. Instead, we adopt a simulation-driven approach to accurately tabulate the throughput loss as a function of the MCS.

Second, in [5], [10], [14], [21], and [19], the SNRs between the BSs and the user are assumed to be all known. Our approach avoids this assumption, which requires excessive feedback, for OT. For any choice of MCSs and any subset of the cooperating BSs, it employs a BLER inference approach that enables the BLER to be determined from the conventional single link 4-bit CSI fed back to each BS. While quantized feedback is considered in [13] and [18], the focus is on single-connectivity.

Third, eMBB-URLLC co-existence is not considered in [7], [10], an [5]. Fourth, we consider discrete rate adaptation, which is used in practice. This requires a different problem formulation and analysis than the idealized continuous adaptation approaches pursued in [5], [11], [14], [15], [19], [20], and [21]

TABLE I
COMPARISON OF LITERATURE ON EMBB-URLLC RESOURCE ALLOCATION AND MULTI-CONNECTIVITY

Reference	*Connectivity*	*CSI Model*	*Fading*	*eMBB-URLLC Coexistence*	*eMBB Loss Model*	*Rate Adaptation*
Mahdi et al. [10]	Multi	Full CSI	No	No	NA	Discrete
Pedersen et al. [13]	Single	Quantized	Frequency-selective	Yes	MMIB based	Discrete
Shang et al. [18]	Single	Quantized	Flat	Yes	Neural network	Discrete
Yin et al. [17]	Single	Quantized	Flat	Yes	Linear	Discrete
Hößler et al. [5]	Multi	Full CSI	No	No	NA	Continuous
Alsenwi et al. [19]	Single	Full CSI	Flat	Yes	Linear	Continuous
Anand et al. [14]	Single	Full CSI	Flat	Yes	Linear, convex, threshold	Continuous
Alsenwi et al. [20]	Single	Full CSI	Flat	Yes	Linear	Continuous
Mahmood et al. [11]	Multi	Full CSI	Flat	No	NA	Continuous
Abdelsadek et al. [15]	Single	Full CSI	Flat	Yes	Linear	Continuous
Liu et al. [21]	Multi (uplink)	Full CSI	Flat	Yes	Linear	Continuous
Mahdi et al. [22]	Multi	Full CSI	Flat	No	NA	Discrete

that use the Shannon capacity formula or the finite block length capacity formula. Fifth, only flat fading is considered in [5], [14], [19], and [15]. To the best of our knowledge, the requirement of the MCS being the same for all the subbands used for transmission or the effect of feedback delays has not been addressed. We also note that issues related to eMBB-URLLC co-existence did not arise in the earlier studies on CoMP [6].

C. Outline

Section II describes the system model. Section III presents the achievability analysis and MCS selection algorithm when the feedback delays are negligible. Section IV extends the formulation to the case where the feedback delays are not negligible. Section V presents our numerical results. Our conclusions follow in Section VI.

Notation: The probability of an event A is denoted by $\Pr(A)$. The probability density function (PDF) and cumulative distribution function (CDF) of a random variable (RV) X are denoted by $f_X(.)$ and $F_X(.)$, respectively. The conditional probability of A given B is denoted by $\Pr(A \mid B)$. The conditional PDF and the conditional CDF of an RV X given Y are denoted by $f_{X|Y}(.)$ and $F_{X|Y}(.)$, respectively. Expectation is denoted by $\mathbb{E}[.]$, and expectation conditioned on Y is denoted by $\mathbb{E}[. \mid Y]$. The real part is denoted by $\Re\{.\}$. The moment generating function (MGF) of an RV X is denoted by $\Psi_X(s)$; it equals $\mathbb{E}[\exp(sX)]$. Vectors and matrices are denoted in bold font. $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$.

II. SYSTEM MODEL

In the 5G physical layer, the system bandwidth is divided into subcarriers in the frequency domain. A group of 12 subcarriers constitutes a physical resource block (PRB). A PRB spans 14 OFDM symbols in time for the eMBB user and only 2 OFDM symbols for the URLLC user due to its low latency requirement. q adjacent PRBs are grouped into a subband, where q depends on the system bandwidth. The total number of subbands is N. The transmission time interval (TTI) for eMBB data depends on the subcarrier bandwidth. For example, it is 1 ms when the subcarrier bandwidth is 15 kHz.

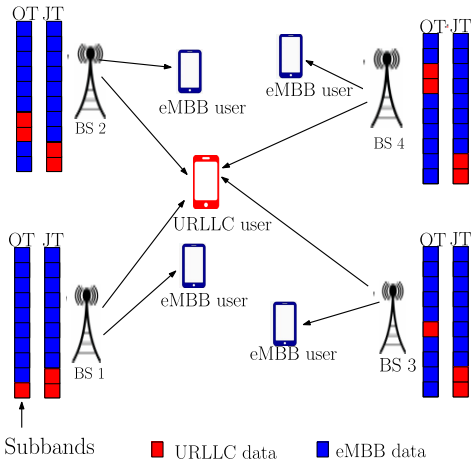


Fig. 1. System model consisting of multiple BSs, that serve a URLLC user and the eMBB users. The red colored boxes represent the subbands allocated to the URLLC user at a BS for JT and OT.

We consider a set $\mathcal{B} = \{1, 2, \ldots, K\}$ of K BSs that serve eMBB users and a URLLC user. A controller that is connected to all the BSs determines which subset $\mathcal{S}$ of BSs serve the URLLC user and their MCSs. In 5G, the master gNB can itself serve as the controller. A BS can use a maximum of R subbands to serve the URLLC user, where R depends upon the bandwidth and K. The MCSs are chosen from the set $\mathcal{M} = \{1, 2, \ldots, M\}$; the MCSs are indexed in the increasing order of their rate.[1] Since the URLLC transmission occurs over a minislot for all MCSs, the latency does not depend on the choice of the MCS. Due to pre-emption, the latency, which includes the time required to schedule the minislot and its transmission duration, is a fraction of a slot.

A. Channel Model

We first describe the model and notation for the scenario where the feedback delays are small relative to the coherence time of the channel. Let H_{ij} denote the gain of subband i between the j^{th} BS and the URLLC user. It is a circularly symmetric complex Gaussian RV, which models Rayleigh fading. Therefore, the fading power $|H_{ij}|^2$ is exponentially distributed

[1] $\mathcal{M}$ only includes the MCSs that can carry the URLLC payload in at most R subbands.

with mean σ_j^2. For tractability, we assume that it is constant over a subband and $|H_{ij}|^2$, for $1 \leq i \leq N$ and $1 \leq j \leq K$, are statistically identical and independent. This is justified when the BSs are sufficiently far apart and the subband and channel coherence bandwidths are comparable [23, Ch. 3], [24]. Let $\mathcal{I}_j$ denote the set of interfering BSs for BS j.

B. Transmission Modes: JT and OT

1) OT: In it, the BSs in $\mathcal{S}$ transmit the URLLC packet on different subbands with the same or different MCSs. In general, let $\mathcal{D}_j$ denote the set of subbands assigned for URLLC at BS j.

For the URLLC user, the SNR Γ_{ij} of subband i due to a transmission by BS $j \in \mathcal{S}$ is[2]

$$\Gamma_{ij} = \frac{P_T |H_{ij}|^2}{P_T \sum_{p \in \mathcal{I}_j} |H_{ip}|^2 + \omega^2}, \tag{1}$$

where P_T is the transmit power of a BS per subband and ω^2 is the noise power per subband. When $|\mathcal{S}|$ increases, the SNR increases and the BLER decreases, but more eMBB resources are pre-empted. Hence, $\mathcal{S}$ needs to be optimized.

Interference Settings: There are two possible settings for $\mathcal{I}_j$. In the *high interference setting*, $\mathcal{I}_j = \mathcal{B} \setminus \{j\}$, $\forall j \in \mathcal{S}$. Here, all the remaining BSs (including those in $\mathcal{S}$) continue to transmit to eMBB users on the subband that BS j uses to transmit to the URLLC user. In the *low interference setting*, $\mathcal{I}_j = \emptyset$, $\forall j \in \mathcal{S}$. Here, no other BS transmits to eMBB users on the subband that BS j uses to transmit to the URLLC user.

For example, let $\mathcal{B} = \{1, 2, 3, 4\}$, $\mathcal{S} = \{1, 2\}$, and $R = 1$, with BSs 1 and 2 transmitting the URLLC packet on subbands 1 and 2, respectively. In the high interference setting, BSs 1, 3, and 4 transmit to their respective eMBB users on subband 2, and BSs 2, 3, and 4 do the same on subband 1. On the other hand, in the low interference setting, BS 1 does not transmit on subband 2, BS 2 does not transmit on subband 1, and BSs 3 and 4 do not transmit to the eMBB users on both subbands during the TTI of the URLLC user. eMBB transmission is pre-empted on a total of 2 subbands across the 4 BSs in the high interference setting, and on 8 subbands in the low interference setting.

2) JT: In JT, the BSs in $\mathcal{S}$ transmit the URLLC packet on the same resource elements with the same MCS. In MRT, BS i transmits the signal on subband j using the weight $w_{ij} = H_{ij}^* / \sqrt{\sum_{p \in \mathcal{S}} |H_{ip}|^2}$. The SNR on subband i is equal to $\sum_{j \in \mathcal{S}} \Gamma_{ij}$, where Γ_{ij} is given in (1).

Similar to OT, there are two possible interference settings. In the high interference setting, $\mathcal{I}_j = \mathcal{B} \setminus \mathcal{S}$, $\forall j \in \mathcal{S}$. Here, all BSs not in $\mathcal{S}$ transmit to eMBB users on the subband that BS j uses to transmit to the URLLC user. In the low interference setting, $\mathcal{I}_j = \emptyset$, $\forall j \in \mathcal{S}$. Since all the BSs use the same resource elements, it is easy to see that $\mathcal{S} \subset \mathcal{B}$ wastes resources without improving the URLLC user's BLER. We shall, therefore, set $\mathcal{S} = \mathcal{B}$ for the second setting.

[2]For ease of exposition, we do not distinguish between signal-to-interference-plus-noise ratio and SNR.

TABLE II
CURVE-FIT PARAMETERS FOR SEVERAL MCSS IN [26, TABLE 5.1.3.1-1] (32 bytes PAYLOAD)

MCS	c_m	d_m	λ_m *(dB)*
QPSK, $r = 78/1024$	1.02×10^5	73.22	-8.20
QPSK, $r = 120/1024$	1.97×10^5	67.07	-6.10
QPSK, $r = 173/1024$	7.02×10^5	38.96	-4.61
QPSK, $r = 308/1024$	3.13×10^5	16.27	-1.01
QPSK, $r = 449/1024$	4.97×10^4	9.47	0.70
QPSK, $r = 602/1024$	5.22×10^5	7.42	2.48
16-QAM, $r = 378/1024$	4.50×10^4	3.40	4.97
16-QAM, $r = 490/1024$	4.65×10^4	2.19	6.90
16-QAM, $r = 616/1024$	5.34×10^4	1.46	8.71
64-QAM, $r = 466/1024$	1.56×10^4	0.90	10.30
64-QAM, $r = 567/1024$	8.77×10^3	0.54	12.36
64-QAM, $r = 666/1024$	4.09×10^3	0.29	14.44
64-QAM, $r = 772/1024$	1.86×10^3	0.12	17.94
64-QAM, $r = 873/1024$	91.55	0.04	20.06
64-QAM, $r = 948/1024$	30.10	0.02	21.90

C. BLER Model

Let $\mathsf{BLER}_m(\gamma)$ denote the BLER of MCS $m \in \mathcal{M}$ at SNR γ. It can be accurately approximated by the following truncated exponential function [25]:

$$\mathsf{BLER}_m(\gamma) = \begin{cases} 1, & 0 < \gamma \leq \lambda_m, \\ c_m \exp\left(-d_m \gamma\right), & \gamma > \lambda_m, \end{cases} \tag{2}$$

where $c_m > 0$ and $d_m > 0$ are MCS-dependent parameters and $\lambda_m = \log(c_m)/d_m$. Table II shows the curve-fit parameters for several MCSs in [26, Table 5.1.3.1-1]. The MCSs are defined in terms of the modulation and code rate r. The truncated exponential function accurately approximates the BLER over a three order of magnitude range from 1 to 10^{-3} for all the MCSs.

D. CSI Feedback and URLLC BLER Computation

OT and JT markedly differ in the CSI that needs to be fed back and the resultant BLER.

1) OT: Let $T_{m_j}(\varepsilon_s)$ denote the SNR threshold at which the BLER of a single link from the BS to the user that uses MCS m_j is equal to ε_s. The URLLC user feeds back to BS $j \in \mathcal{B}$ the MCS index m_j for subband i, when m_j is the highest rate MCS for which $T_{m_j}(\varepsilon_s) \leq \Gamma_{ij}$. Note that the BLER with multi-connectivity will be different from ε_s since it depends on $\mathcal{S}$. The controller determines the BLER for any $\mathcal{S}$ using the following inference approach.

Consider first the high interference setting. Let the CSI fed back to the j^{th} BS for the i^{th} subband be the MCS index q_{ij}. The SNR due to a transmission by the j^{th} BS on the i^{th} subband is conservatively taken to be the lower threshold $T_{q_{ij}}(\varepsilon_s)$. The URLLC packet experiences different SNRs on different subbands. However, as mentioned, the 5G standard requires the same MCS to be used on all the subbands assigned to a user in order to limit the control channel overhead. To systematically evaluate the BLER of any MCS, we employ exponential effective SNR mapping (EESM), which has been extensively used in 3GPP system simulations [27], [28]. It maps the vector of SNRs Γ_{ij}, $\forall i \in \mathcal{D}_j$, into a single effective SNR $\gamma_{m_j}^{(j)}$, which is the equivalent

SNR in a frequency-flat channel that results in the same BLER for MCS m_j. It is given by

$$\gamma_{m_j}^{(j)} = -\beta_{m_j} \log\left(\frac{1}{N_{m_j}} \sum_{i \in \mathcal{D}_j} \exp\left(\frac{-\Gamma_{ij}}{\beta_{m_j}}\right)\right), \quad (3)$$

where $N_{m_j} = |\mathcal{D}_j| \leq R$ is the number of subbands allocated by a BS to transmit the URLLC packet with MCS m_j and β_{m_j} is an MCS-dependent constant. Its values are available in [29]. These are found using bit-level calibrations and need to be determined only once. We set $\Gamma_{ij} = T_{q_{ij}}(\varepsilon)$ conservatively. The BLER is then given by $\prod_{j \in \mathcal{S}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}\right)$.

For the low interference setting, the corresponding approach remains the same except that q_{ij} is generated using measurements made when no BS other than i is transmitting on subband j.[3] With 4 bits required per subband per BS, the feedback overhead of OT is $4KR$ bits.

2) JT: Since JT uses MRT, the controller needs to know the channel gains H_{ij}, $\forall j \in \mathcal{B}$ and $1 \leq i \leq N$. In the high interference setting, the SNR of the URLLC packet on subband i is $\sum_{j \in \mathcal{S}} \Gamma_{ij}$. Without loss of generality, let subbands $1, \ldots, N_m$ be assigned for URLLC at each BS. Hence, the effective SNR $\gamma_m^{\mathcal{S}}$ over N_m subbands of a common MCS m is

$$\gamma_m^{\mathcal{S}} = -\beta_m \log\left(\frac{1}{N_m} \sum_{i=1}^{N_m} \exp\left(\frac{-\sum_{j \in \mathcal{S}} \Gamma_{ij}}{\beta_m}\right)\right), \quad (4)$$

and the BLER is $c_m \exp\left(-d_m \gamma_m^{\mathcal{S}}\right)$. In the low interference setting with $\mathcal{S} = \mathcal{B}$, the SNR on subband i is $\sum_{j \in \mathcal{B}} \Gamma_{ij}$, where Γ_{ij} is the SNR when no other BS is transmitting. The effective SNR $\gamma_m^{\mathcal{B}}$ of MCS m is given by (4) and the BLER is $c_m \exp\left(-d_m \gamma_m^{\mathcal{B}}\right)$. Assuming that 6 bits per subband per BS are required to feed back the amplitude and phase of the complex MRT weights, the feedback overhead of JT is $6KR$ bits.

The interference settings, BLER expressions, and CSI fed back for OT and JT are summarized in Table III.

III. With Negligible Feedback Delays

At the j^{th} BS, let L_{m_j} denote the average eMBB throughput loss when it uses MCS m_j to transmit the URLLC data payload over N_{m_j} subbands. It is determined using bit-level simulations. As the MCS index increases, the number of resource elements required to transmit a given URLLC data payload decreases and so does the eMBB throughput loss.[4] In OT, the total eMBB throughput loss across all BSs is $K \sum_{j \in S} L_{m_j}$ for the low interference setting and $\sum_{j \in S} L_{m_j}$ for the high interference setting. For JT, all the cooperating BSs use the same MCS m and resource elements. Therefore, the above loss changes to KL_m and $|\mathcal{S}|L_m$ for the low and high interference settings, respectively.

The problem of minimizing the eMBB throughput loss subject to the URLLC error requirements can be stated as follows. Its form depends on whether OT or JT is used and the interference setting. For OT and the high interference setting, the problem can be stated as

$$\min_{\substack{\mathcal{S} \subseteq \mathcal{B} \\ m_j \in \mathcal{M}}} \left\{\sum_{j \in \mathcal{S}} L_{m_j}\right\}, \quad (5)$$

$$\text{s.t.} \prod_{j \in \mathcal{S}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}\right) \leq \varepsilon. \quad (6)$$

For the low interference setting, $\sum_{j \in \mathcal{S}} L_{m_j}$ is replaced with $K \sum_{j \in \mathcal{S}} L_{m_j}$ in the objective function. Similarly for JT and high interference setting, the problem can be stated as

$$\min_{\substack{\mathcal{S} \subseteq \mathcal{B} \\ m \in \mathcal{M}}} \{|\mathcal{S}| L_m\}, \quad (7)$$

$$\text{s.t. } \mathsf{BLER}_m\left(\gamma_m^{\mathcal{S}}\right) \leq \varepsilon, \quad (8)$$

where $\gamma_m^{\mathcal{S}}$ is defined in Section II-D.2. For the low interference setting, since $\mathcal{S} = \mathcal{B}$, only the MCS needs to be determined.

A. Achievability Analysis

We first evaluate the existence of a feasible solution to the above problem. To do this, we analyze achievability, which is the probability that at least one subset of cooperating BSs and MCSs can meet the URLLC BLER target. Intuitively, the target may not be met when the links from all the BSs to the user are in a deep fade. One option to overcome this issue is to use retransmissions. However, this increases latency. Therefore, the cellular operator should strive to ensure a high achievability over the set of all locations of the URLLC user.

It is clear that the system satisfies the URLLC error constraint if and only if the BLER when all the K BSs cooperate (i.e., $\mathcal{S} = \mathcal{B}$) and transmit with the lowest rate MCS 1 is at most ε. This is because the BLER increases when $\mathcal{S}$ is smaller or any BS uses a higher rate MCS.

1) OT: In terms of effective SNR, the achievability A is given by

$$A = \Pr\left(\prod_{j \in \mathcal{B}} \mathsf{BLER}_1\left(\gamma_1^{(j)}\right) \leq \varepsilon\right). \quad (9)$$

The following result presents a tractable lower bound for it.

Result 1: The achievability of OT is lower bounded by

$$A \geq \sum_{\mathcal{V} \subseteq \mathcal{B}} \left(\prod_{j \in \mathcal{V}} B\left(e^{-\frac{\lambda_1}{\beta_1}}, a_1^{(j)}, b_1^{(j)}\right) - \prod_{j \in \mathcal{V}} \left[B\left(e^{-\frac{\lambda_1}{\beta_1}}, a_1^{(j)}, b_1^{(j)}\right) - B\left(e^{-\frac{|\mathcal{V}| \log(c_1) - \log(\varepsilon) - (|\mathcal{V}|-1) d_1 \lambda_1}{d_1 \beta_1}}, a_1^{(j)}, b_1^{(j)}\right) \right] \right) \times \prod_{j \in \mathcal{B} \setminus \mathcal{V}} \left[1 - B\left(e^{-\frac{\lambda_1}{\beta_1}}, a_1^{(j)}, b_1^{(j)}\right)\right], \quad (10)$$

where $B(.,.,.)$ is the regularized incomplete Beta function [30, (6.6.2)] and λ_1 is defined in Section II-C. Here, $a_1^{(j)}$ and $b_1^{(j)}$

[3] In practice, these can be estimated using the downlink reference signals of the BSs.

[4] Section V presents a numerical example of this.

TABLE III
INTERFERENCE SETTINGS, BLER EXPRESSIONS, AND CSI REQUIREMENTS OF OT AND JT

Mode	*Interference Setting*		*BLER*	*CSI Feedback*
	Low	*High*		
OT	$\mathcal{I}_j = \emptyset$	$\mathcal{I}_j = \mathcal{B} \setminus \{j\}$	$\prod_{j\in\mathcal{S}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}\right)$	MCS index m_j for BS j, $\forall j \in S$
JT	$\mathcal{I}_j = \emptyset$ $(\mathcal{S} = \mathcal{B})$	$\mathcal{I}_j = \mathcal{B} \setminus \mathcal{S}$	$\mathsf{BLER}_m\left(\gamma_m^{\mathcal{S}}\right)$	$H_{ij}, \forall j \in \mathcal{B}, 1 \le i \le N$

are given by

$$a_1^{(j)} = \frac{g_j\left[g_j N_1 - s_j - (N_1 - 1)\, g_j^2\right]}{s_j - g_j^2}, \tag{11}$$

$$b_1^{(j)} = \frac{(1-g_j)a_1^{(j)}}{g_j}, \tag{12}$$

where $g_j = \omega^2\beta_1/(\omega^2\beta_1 + P_T\sigma_j^2)$ and $s_j = \omega^2\beta_1/(\omega^2\beta_1 + 2P_T\sigma_j^2)$.

Proof: The proof is given in Appendix A. ∎

2) JT: The following is the corresponding result on achievability for JT.

Result 2: The achievability of JT is given by

$$A = B\left(\left(c_1\varepsilon^{-1}\right)^{\frac{1}{d_1\beta_1}}, a_1^{\mathcal{B}}, b_1^{\mathcal{B}}\right), \tag{13}$$

where $a_1^{\mathcal{B}}$ and $b_1^{\mathcal{B}}$ are obtained from (11) and (12) by replacing g_j with $\prod_{j\in\mathcal{B}} g_j$ and s_j with $\prod_{j\in\mathcal{B}} s_j$.

Proof: The proof is given in Appendix B. ∎

B. MCMSA

1) OT: In (5), since there are K BSs and M MCSs, M^K choices of subsets and MCSs exist and require M^K BLER computations. Thus, the complexity increases exponentially in the number of BSs. We present a simple algorithm MCMSA below. The steps are similar for the two interference settings. It uses the following lemma to significantly reduce the number of computations without any loss in optimality.

Lemma 1: a) Let m_j^* be the highest rate MCS at BS j whose BLER is less than or equal to ε. Then, BS j will never use an MCS with a lower rate than m_j^*.

b) If the MCS vector $(m_j, j \in \mathcal{S})$ cannot meet the BLER target, then the MCS vector $(m_j + \nu_j, j \in \mathcal{S})$, where $\nu_j \ge 0$, $\forall j \in \mathcal{S}$, also cannot meet the error target.

Proof: The proof is given in Appendix C. ∎

Given $\mathcal{S}$, from Lemma 1, it is sufficient to only look at MCS vectors of the form $(m_j^* + \nu_j, \forall j \in \mathcal{S})$, where $\nu_j \ge 0, \forall j \in \mathcal{S}$. For a given MCS vector $\mathbf{m} = (m_j, \forall j \in \mathcal{S})$, MCMSA then computes the BLER for it as per Section II-D. If the BLER is less than or equal to ε, then the MCS vector is a feasible solution. In case it is not feasible, then the search space can be reduced using Lemma 1 because all MCS vectors of the form $(m_j + z_j, \forall j \in \mathcal{S})$, where $z_j \ge 0$, are also not feasible. MCMSA does this for every set $\mathcal{S} \subseteq \mathcal{B}$. For each feasible vector, it also computes the eMBB sum throughput loss, and selects the one with the smallest loss. When no feasible choice of MCSs exists for any $\mathcal{S} \subseteq \mathcal{B}$, the controller transmits the URLLC packet from all the BSs with the lowest rate MCS 1.

The pseudo-code of MCMSA is given in Algorithm 1.

Algorithm 1 MCMSA

Initialization `Count` $= 0$
for every $\mathcal{S} \subseteq \mathcal{B}$ **do**
  $\mathcal{T} = \left\{(m_j' + \nu_j, \forall \nu_j \ge 0, j \in \mathcal{S})\right\}$
  while $\mathcal{T} \neq \emptyset$ **do**
    Draw $\mathbf{m} = (m_j, j \in \mathcal{S}) \in \mathcal{T}$
    if $\prod_{j\in\mathcal{S}} c_{m_j} \exp\left(-d_{m_j}\gamma_{m_j}^{(j)}\right) \le \varepsilon$ **then**
      Compute $\sum_{j\in\mathcal{S}} L_{m_j}$, `Count` $\leftarrow$ `Count` $+ 1$
    else
      Remove all vectors of the form $\mathbf{m}' = (m_j + z_j, \forall z_j \ge 0, j \in \mathcal{S})$ from $\mathcal{T}$
    end if
  end while
end for
Select the set of BSs and their MCSs with the smallest sum throughput loss.
If `Count` $= 0$, set $\mathcal{S} = \mathcal{B}$ and select MCS 1 for all BSs.

2) JT: As above, we consider the low and high interference settings separately.

- *High Interference Setting:* The optimization problem in (7) requires $\left(2^K - 1\right) M$ possibilities to be examined.
- *Low Interference Setting*: Since all K BSs transmit with the same MCS, the number of possibilities is M.

We see that the complexity is lower for JT because the same MCS is used by all the BSs in $\mathcal{S}$. As in OT, the search complexity can be reduced significantly.

IV. MCS SELECTION WITH NON-NEGLIGIBLE FEEDBACK DELAYS

Let t_1 denote the time at which the CSI report is generated and $t_2 = t_1 + \tau$ be the packet transmission time. The MCSs and the cooperating BSs set that are selected become partially outdated because they are based on the CSI feedback at time t_1, while transmission occurs at time t_2. In the discussion that follows, we update the notation for the channel gains and SNRs to also show the time indices. As per the Jakes' model, $H_{ij}(t_1)$ and $H_{ij}(t_2)$ are jointly Gaussian with correlation coefficient $\rho(\tau) = J_0(2\pi f_d \tau)$, where $J_0(.)$ is the zeroth-order Bessel function of the first kind and f_d is the Doppler spread [31, Ch. 3]. For tractability, we focus on the low interference setting. The high interference setting, which has a lower achievability, is intractable due to the interference term in the denominator of the SNR expression in (1).

The instantaneous BLER constraint in (6) cannot be satisfied as the BSs do not know the SNRs at the time of transmission. In the following, we propose a stochastic reliability constraint for URLLC, as per which the probability that the instantaneous BLER is less than the target value ε must be at least $1-\Delta$. The

value of $\Delta \ll 1$ depends on the application; smaller values of Δ and ε imply a tighter reliability constraint.

A. OT

Let $\mathbf{Q}_{t_1}$ be the CSI fed back at time t_1. It is a matrix whose $(i,j)^{\text{th}}$ element q_{ij} represents the MCS index fed back for the i^{th} subband at the j^{th} BS at time t_1. Therefore, $T_{q_{ij}}(\varepsilon_s) \leq \Gamma_{ij}(t_1) < T_{q_{ij}+1}(\varepsilon_s)$.

Let $\gamma_{m_j}^{(j)}(t_2)$ denote the effective SNR of MCS m_j at time t_2 due to a transmission by the j^{th} BS. It is given by

$$\gamma_{m_j}^{(j)}(t_2) = -\beta_{m_j} \log\left(\frac{1}{N_{m_j}} \sum_{i \in \mathcal{D}_j} \exp\left(\frac{-\Gamma_{ij}(t_2)}{\beta_{m_j}}\right)\right). \quad (14)$$

The stochastic reliability constraint can be stated as

$$\Pr\left(\prod_{j \in \mathcal{S}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}(t_2)\right) \leq \varepsilon | \mathbf{Q}_{t_1}\right) \geq 1 - \Delta. \quad (15)$$

Result 3: The conditional probability in (15) is lower bounded as follows:

$$\begin{aligned}
&\Pr\left(\prod_{j \in \mathcal{S}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}(t_2)\right) \leq \varepsilon | \mathbf{Q}_{t_1}\right) \\
&\geq \sum_{\mathcal{V} \subseteq \mathcal{S}} \left(\left[\prod_{j \in \mathcal{V}} \left(1 - F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\lambda_{m_j}\right)\right)\right]\right. \\
&\quad \left. - \prod_{j \in \mathcal{V}} \left[F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\frac{\theta_{\mathcal{V}} - \zeta_j}{d_{m_j}}\right) - F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\lambda_{m_j}\right)\right]\right) \\
&\quad \times \left[\prod_{j \in \mathcal{S} \setminus \mathcal{V}} F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\lambda_{m_j}\right)\right], \quad (16)
\end{aligned}$$

where $F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}(.)$ denotes the CDF of $\gamma_{m_j}^{(j)}(t_2)$ conditioned on $\mathbf{Q}_{t_1}$, $\zeta_j = \sum_{\substack{k \in \mathcal{V}, \\ k \neq j}} d_{m_k} \lambda_{m_k}$, and $\theta_{\mathcal{V}} = \log\left(\varepsilon^{-1} \prod_{j \in \mathcal{V}} c_{m_j}\right)$.

Proof: The proof is given in Appendix D. ∎

The last step is to derive an expression for $F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}(.)$.

Result 4: The conditional CDF $F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}(.)$ is given by

$$F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}(x) = B\left(e^{-\frac{x}{\beta_{m_j}}}, a_{m_j}, b_{m_j}\right), \quad \text{for } x \geq 0, \quad (17)$$

where the beta parameters a_{m_j} and b_{m_j} are given by

$$\begin{aligned}
a_{m_j} &= \frac{\sum_{i \in \mathcal{D}_j} g_{ij}}{\sum_{i \in \mathcal{D}_j} s_{ij} - \sum_{i \in \mathcal{D}_j} g_{ij}^2} \left[\left(\sum_{i \in \mathcal{D}_j} g_{ij}\right)\right. \\
&\quad \left. - \frac{1}{N_{m_j}} \left(\sum_{i \in \mathcal{D}_j} s_{ij} - \sum_{i \in \mathcal{D}_j} \sum_{l \in \mathcal{D}_j, l \neq i} g_{ij} g_{lj}\right)\right], \quad (18)
\end{aligned}$$

$$b_{m_j} = \frac{\left(N_{m_j} - \sum_{i \in \mathcal{D}_j} g_{ij}\right)}{\sum_{i \in \mathcal{D}_j} g_{ij}} a_{m_j}. \quad (19)$$

Here, $g_{ij} = (\beta_{m_j}/(\beta_{m_j} + 2\alpha_j)) \exp\left(-\delta_{ij}\alpha_j/(\beta_{m_j} + 2\alpha_j)\right)$, $s_{ij} = (\beta_{m_j}/(\beta_{m_j} + 4\alpha_j)) \exp\left(-2\delta_{ij}\alpha_j/(\beta_{m_j} + 4\alpha_j)\right)$, $\alpha_j = d_{m_j} P_T (1 - \rho^2(\tau)) \sigma_j^2/(2\omega^2)$, and $\delta_{ij} = 2\rho^2(\tau)|H_{ij}(t_1)|^2/(1 - \rho^2(\tau))$.

Proof: The proof is given in Appendix E. ∎

MCS Selection: The MCMSA algorithm remains the same as that in Section III-B.1, except that a feasible MCS vector must satisfy the stochastic constraint in (15). For the stochastic constraint, we use its lower bound due to its tractability. If the lower bound exceeds $1 - \Delta$, then (15) is satisfied.

B. JT

In JT, the weights used for MRT by the BSs become partially outdated by the time of transmission. The CSI $\mathbf{Q}_{t_1}$ fed back at time t_1 is a matrix with $(i,j)^{\text{th}}$ element $H_{ij}(t_1)$. The MRT weight of BS j for subband i is $H_{ij}^*(t_1)/w_i(t_1)$, where $w_i(t_1) = \sqrt{\sum_{j \in \mathcal{B}} |H_{ij}(t_1)|^2}$. Therefore, the SNR $\Gamma_i^{\mathcal{B}}(t_2)$ of subband i at time t_2 is

$$\Gamma_i^{\mathcal{B}}(t_2) = \frac{P_T}{\omega^2 (w_i(t_1))^2} \left| \sum_{j \in \mathcal{B}} H_{ij}^*(t_1) H_{ij}(t_2) \right|^2. \quad (20)$$

The effective SNR $\gamma_m^{\mathcal{B}}(t_2)$ of MCS m at time t_2 is then equal to $-\beta_m \log(Y'_m)$, where

$$Y'_m = \frac{1}{N_m} \sum_{i=1}^{N_m} \exp\left(-\frac{\Gamma_i^{\mathcal{B}}(t_2)}{\beta_m}\right). \quad (21)$$

The stochastic reliability constraint for MCS m is then given by

$$\Pr\left(\mathsf{BLER}_m\left(\gamma_m^{\mathcal{B}}(t_2)\right) \leq \varepsilon \mid \mathbf{Q}_{t_1}\right) \geq 1 - \Delta. \quad (22)$$

Result 5: The conditional probability in (22) is given by

$$\begin{aligned}
&\Pr\left(\mathsf{BLER}_m\left(\gamma_m^{\mathcal{B}}(t_2)\right) \leq \varepsilon \mid \mathbf{Q}_{t_1}\right) \\
&\qquad = B\left(\left(c_m \varepsilon^{-1}\right)^{\frac{1}{d_m \beta_m}}, a'_m, b'_m\right). \quad (23)
\end{aligned}$$

The beta parameters a'_m and b'_m are given by

$$\begin{aligned}
a'_m &= \frac{\left(\sum_{i=1}^{N_m} g'_i\right)}{\sum_{i=1}^{N_m} s'_i - \sum_{i=1}^{N_m} (g'_l)^2} \left[\left(\sum_{i=1}^{N_m} g'_i\right)\right. \\
&\quad \left. - \frac{1}{N_m} \left(\sum_{i=1}^{N_m} s'_i + \sum_{i=1}^{N_m} \sum_{l=1, l \neq i}^{N_m} g'_i g'_l\right)\right], \quad (24)
\end{aligned}$$

$$b'_m = \frac{\left(N_m - \sum_{i=1}^{N_m} g'_i\right) a'_m}{\sum_{i=1}^{N_m} g'_i}, \quad (25)$$

where $g'_i = (\beta_m/(\beta_m + 2\alpha'_i)) \exp\left(-\delta'_i \alpha'_i/(\beta_m + 2\alpha'_i)\right)$, $s'_i = (\beta_m/(\beta_m + 4\alpha'_i)) \exp\left(-2\delta'_i \alpha'_i/(\beta_m + 4\alpha'_i)\right)$, $\delta'_i = 2\rho^2(\tau)(w_i(t_1))^4 / \left((1 - \rho^2(\tau)) \sum_{j \in \mathcal{B}} \sigma_j^2 |H_{ij}(t_1)|^2\right)$, and $\alpha'_i = P_T(1 - \rho^2(\tau)) \left(\sum_{j \in \mathcal{B}} \sigma_j^2 |H_{ij}(t_1)|^2\right)/(2\omega^2 (w_i(t_1))^2)$.

Proof: The proof is given in Appendix F. ∎

MCS Selection: The highest rate MCS that satisfies the stochastic constraint in (22) is selected at the time of transmission as it requires the least resources. This requires only M computations.

TABLE IV
eMBB THROUGHPUT LOSS PER PRB L_m (IN Mbps) AS A FUNCTION OF THE URLLC MCS INDEX m

m	MCS	L_m
1	QPSK, $r = 308/1024$	3.078
2	QPSK, $r = 449/1024$	2.052
3	QPSK, $r = 602/1024$	1.714
4	16-QAM, $r = 378/1024$	1.539
5	16-QAM, $r = 490/1024$	1.368
6	16-QAM, $r = 616/1024$	1.197
7	64-QAM, $r = 466/1024$	1.026
8	64-QAM, $r = 567/1024$	0.855
9	64-QAM, $r = 666/1024$	0.684
10	64-QAM, $r = 772/1024$	0.520
11	64-QAM, $r = 873/1024$	0.342
12	64-QAM, $r = 948/1024$	0.171

The above approaches can also be extended to account for multiple antennas at the transmitter or receiver. However, the statistics of the subband SNR are different [24].

V. NUMERICAL RESULTS AND BENCHMARKING

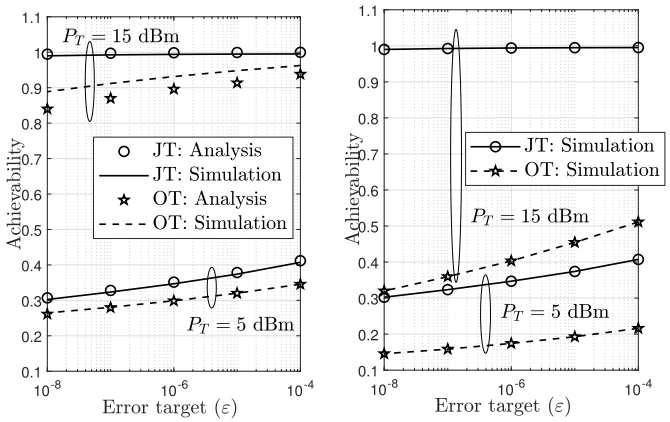
We now present Monte Carlo simulations to evaluate the URLLC achievability and the eMBB throughput loss. We consider a square grid of size 300 m $\times$ 300 m with four BSs that are located at the grid's corners. The noise power spectral density is -174 dBm/Hz, noise figure is 10 dB, system bandwidth is 20 MHz, and each subband consists of $q = 2$ PRBs. The URLLC data payload is 32 bytes [13] and $R = 4$. We set $\varepsilon_s = 0.1$ and use the urban macro scenario's pathloss model [32].[5] Each PRB is assigned to a different eMBB user. The results are averaged over 500 drops of the URLLC user and 2000 fade realizations per drop.

We ran bit-level simulations using Matlab's 5G toolbox to determine the eMBB throughput loss for each MCS. The increase in the BLER due to the pre-emption of two OFDM symbols is evaluated at the SNR at which the eMBB BLER is 0.1, which is the BLER target in 4G/5G. The eMBB user's MCS is chosen with equal probability from the MCSs in Table II. Table IV lists the eMBB throughput loss per PRB for the MCSs that can carry the URLLC payload in at most R subbands.

A. With Negligible Feedback Delays

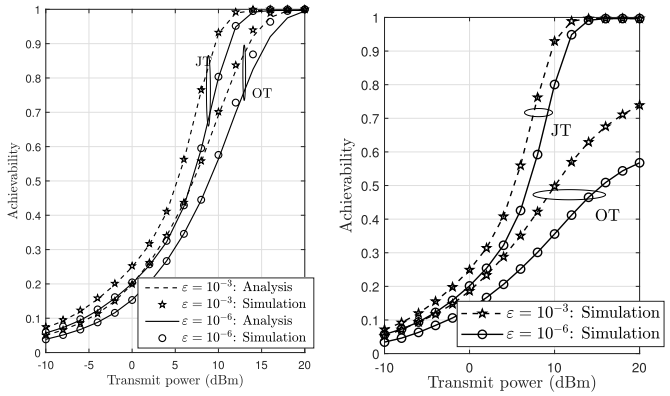
1) Achievability: Fig. 2a plots the achievability of OT and JT averaged over different user locations for the low interference setting as a function of the URLLC error target. It shows results for two values of P_T. Also plotted are the results from the analysis. For OT, the lower bound tracks the simulation curve. For JT, the analytical and simulation results match well. As ε increases, the achievability increases for both transmit powers because the SNR required to meet the error target decreases. The achievability of JT is greater than that of OT. Fig. 2b plots the corresponding results for the high interference setting. The achievability of JT remains the same as that for the low interference setting since feasibility is evaluated when all the BSs transmit to the URLLC user. For OT, the achievability drops significantly due to the interference from the remaining BSs.

[5]We have observed that the performance of MCMSA is insensitive to the choice of ε_s because of the BLER inference approach presented in Section II-D.



(a) Low interference setting (b) High interference setting

Fig. 2. Achievability averaged over different locations of OT and JT as a function of error target for low and high interference settings.



(a) Low interference setting (b) High interference setting

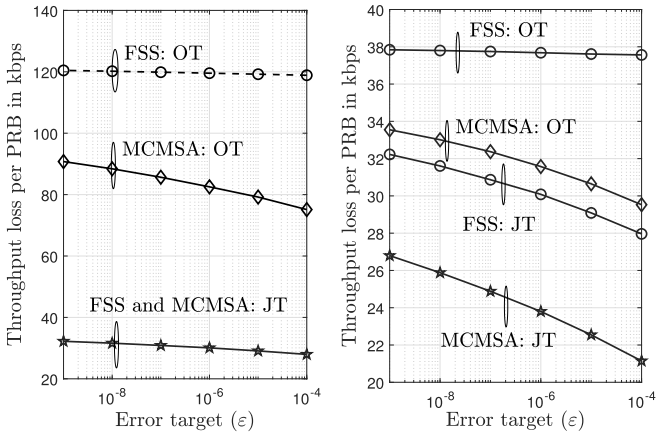
Fig. 3. Achievability averaged over different locations of OT and JT as a function of transmit power for low and high interference settings.

Fig. 3a plots the achievability averaged over different URLLC user locations as a function of P_T for two values of ε. It does so for OT and JT for the low interference setting. Also plotted are the results from analysis. For OT, the lower bound tracks the simulation curve well and is tight for $P_T \leq 10$ dBm. For JT, the analysis and simulation results match well. Fig. 3b plots the corresponding results for the high interference setting. As P_T increases, the achievability increases because the SNR increases. As before, the achievability of JT is the same for the two interference settings, while that of OT is lower for the high interference setting.

2) eMBB Throughput Loss: We benchmark MCMSA with the fixed subset size (FSS) approach. In it, $\mathcal{S}$ is set as $\mathcal{B}$ so as not to compromise on the achievability. Furthermore, the MCS at a BS is determined as follows. For OT, the BLER of the URLLC packet is $\prod_{j\in\mathcal{B}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}\right)$. Therefore, when all K BSs contribute equally to the BLER, the MCS m_j for the j^{th} BS is

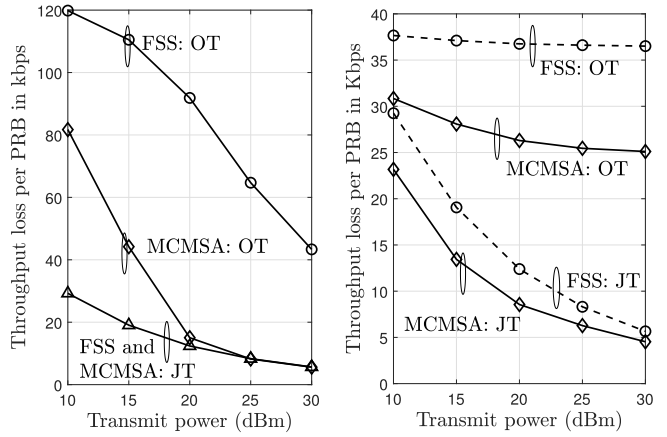
$$m_j = \max\left\{m \in \mathcal{M} : T_m\left(\varepsilon^{1/K}\right) < \gamma_m^{(j)}\right\}. \quad (26)$$

In JT, all the BSs transmit with the highest rate MCS that meets the BLER target. We note that a comparison with the



(a) Low interference setting (b) High interference setting

Fig. 4. Comparison of eMBB throughput loss per PRB of MCMSA and the FSS approach as a function of error target for low and high interference settings ($P_T = 10$ dBm).



(a) Low interference setting (b) High interference setting

Fig. 5. Comparison of eMBB throughput loss per PRB of MCMSA and the FSS approach as a function of P_T for low and high interference settings ($\varepsilon = 10^{-5}$).

algorithms in [5], [10], [21], [22], and [7] is not meaningful because of fundamental differences in the models, constraints, and objectives.

Fig. 4a benchmarks the eMBB throughput loss per PRB of OT and JT as a function of ε for the low interference setting for two values of P_T. As ε increases, the throughput loss decreases. This is because the probability that the BSs use a higher rate MCS, which requires fewer resource elements to transmit the URLLC packet, increases. OT has a larger throughput loss than JT. This is because eMBB transmissions do not occur on $|\mathcal{S}|K$ subbands in OT and on K subbands in JT. For OT, the throughput loss of MCMSA is much smaller than the FSS approach. For JT, the throughput losses of MCMSA and the FSS approach are the same because they choose the same MCS in the low interference setting. Fig. 4b plots the corresponding results for the high interference setting. MCMSA has a smaller throughput loss than the FSS approach for both OT and JT. The throughput loss for OT is now significantly smaller.

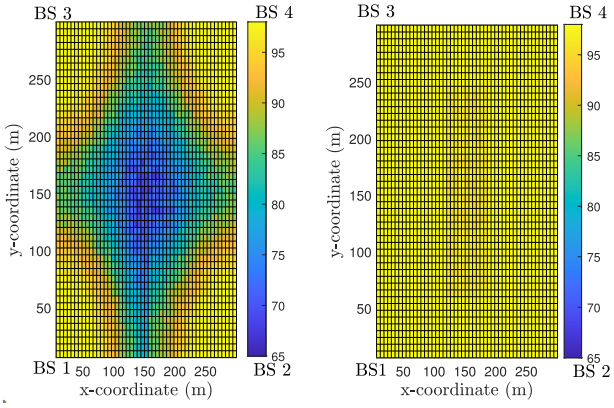
Fig. 5a benchmarks the eMBB throughput loss per PRB of OT and JT as a function of P_T for the low interference setting. As P_T increases, the throughput loss decreases because the BSs use higher rate MCSs. The throughput loss of MCMSA is less than that of the FSS approach for OT. For JT, the throughput loss is the same for MCMSA and the FSS approach. Fig. 5b plots the corresponding results for the high interference setting. As before, the throughput loss for OT is smaller, and the throughput loss of MCMSA is smaller than the FSS approach for both OT and JT. The throughput loss can be different for different BSs for a given URLLC user location. The odds that a BS is included in $\mathcal{S}$ depend on its distance from the user. We also note that the nearest BS need not always be included in $\mathcal{S}$ because the signal strength depends on shadowing and small-scale fading besides pathloss.

Complexity Evaluation: For OT, MCMSA requires $O(M^K)$ computations to determine the optimal subset of cooperating BSs and their MCSs. For JT, MCMSA requires $O\left((2^K-1)M\right)$ computations for the high interference setting and $O(M)$ for the low interference setting. However, numerically, it has a much lower complexity. We see this in Table V, which plots the number of computations of MCMSA and exhaustive search averaged across user locations and channel fades.

TABLE V

COMPARISON OF COMPLEXITY OF MCMSA AND EXHAUSTIVE SEARCH ($M = 12$, $K = 4$, AND $P_T = 10$ dBm)

Transmission Mode	Interference Setting	Exhaustive Search	MCMSA
OT	High	50625	2721
OT	Low	50625	1248
JT	High	225	43
JT	Low	15	7



(a) Ignoring feedback delays (b) Considering feedback delays

Fig. 6. Comparison of the probability that the BLER is below the target value when feedback delays are ignored and accounted for (low interference setting, $f_d\tau = 0.3$, $\Delta = 0.05$, $\varepsilon = 10^{-6}$, and $P_T = 20$ dBm).

B. With Non-Negligible Feedback Delays

Fig. 6a shows a heat map of the probability that the BLER is below the target value at different locations in the grid, when the MCS is based on the feedback at time t_1 and the effect of feedback delays ignored. At the corners of the grid, the probability is close to one because the pathloss from the closest BS is sufficiently small to ensure that the BLER is below ε. However, closer to center of the grid, the pathloss

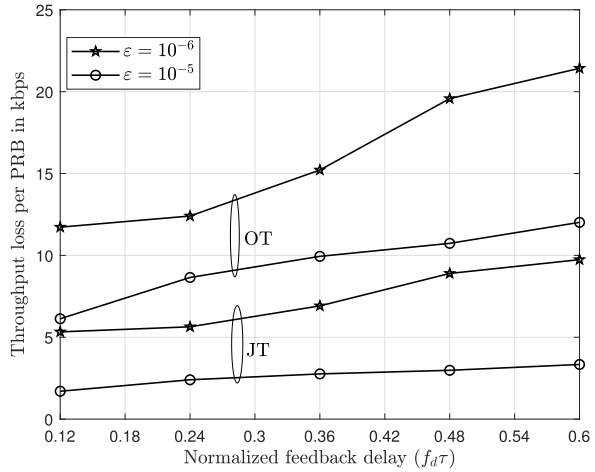


Fig. 7. eMBB throughput loss per PRB as a function of normalized feedback delay for JT and OT ($\Delta = 0.05$ and $P_T = 20$ dBm).

is more and the SNR is more sensitive to multi-path fading. This increases the odds that the MCS selected has a BLER greater than ε at the time of transmission. Fig. 6b shows the corresponding heat map when the MCS is selected as per the proposed approach in Section IV-B, which accounts for the feedback delays. We see that 93% of the time, the BLER is below the target value at any location in the grid for $\Delta = 0.05$. Thus, the proposed approach has a much higher probability of meeting the BLER target.

Fig. 7 plots the eMBB throughput loss per PRB of MCMSA as a function of the normalized feed back delay $f_d\tau$ for JT and OT for two error targets. The throughput loss increases as $f_d\tau$ increases because the CSI becomes less accurate. Consequently, the BSs select lower rate MCSs, which require more resource elements, more often. The throughput loss for OT is more than that of JT, as was the case with negligible feedback delays.

VI. Conclusion

We addressed the problem of optimizing the set of BSs that cooperate to transmit the URLLC data and their MCSs. The goal was to minimize the eMBB throughput loss due to multi-connectivity while meeting the URLLC reliability requirements. We saw that the eMBB throughput loss when the set of cooperating BSs and their MCSs were jointly selected was lower than with FSS. MCMSA enabled the above selection with significantly fewer computations compared to exhaustive search. Using results about EESM, we derived a tractable lower bound for the achievability of OT and a closed-form expression for the achievability of JT for frequency-selective fading channels. For the scenario with non-negligible feedback delays, we introduced a stochastic reliability constraint and derived a tractable expression for the conditional probability that the instantaneous BLER remained below a pre-specified threshold. Selecting the set of cooperating BSs and their MCSs led to a markedly higher probability that the BLER was below the target value compared to the conventional approach that neglected the feedback delays.

JT had a higher achievability than OT over a wide range of URLLC error targets and BS transmit powers, but required more feedback. Similarly, the low interference setting had a higher achievability, but incurred a larger eMBB throughput loss. Therefore, when the uplink feedback bandwidth is very limited, OT and the low interference setting are preferable. When reducing the eMBB throughput loss is the main concern, JT and the high interference setting are preferable. When ensuring high achievability is the main concern, JT is preferable.

We focused on the single URLLC user scenario. For heavier URLLC traffic, the multiple user scenario is of interest. Since a BS employs at most R subbands to serve a URLLC user, it can support $\lfloor N/R \rfloor$ users simultaneously, where $\lfloor \cdot \rfloor$ denotes the floor function. Furthermore, the definition and expressions for achievability for JT and OT apply to the multiple user scenario as well. The same holds for the stochastic reliability constraint when the feedback delays are not negligible. Extending the model and analysis further is an interesting avenue for future work.

Appendix

A. Proof of Result 1

From (2), we know that $\mathsf{BLER}_1\left(\gamma_1^{(j)}\right) = 1$ when $\gamma_1^{(j)} \leq \lambda_1$, and $\mathsf{BLER}_1\left(\gamma_1^{(j)}\right) = c_1 e^{-d_1\gamma_1^{(j)}}$ when $\gamma_1^{(j)} > \lambda_1$. Let $\mathcal{V}$ denote a subset of BSs whose effective SNR of MCS 1 exceeds λ_1. From the law of total probability, by summing over all possible $\mathcal{V}$, A in (9) can be expressed as

$$A = \sum_{\mathcal{V}\subseteq\mathcal{B}} \Pr\left(\prod_{j\in\mathcal{V}} c_1 \exp\left(-d_1\gamma_1^{(j)}\right) \leq \varepsilon | \mathcal{V}\right) \Pr\left(\mathcal{V}\right), \quad (27)$$

where

$$\Pr\left(\mathcal{V}\right) = \Pr\left(\gamma_1^{(j)} > \lambda_1, \forall j \in \mathcal{V}; \gamma_1^{(j)} \leq \lambda_1, \forall j \in \mathcal{B}\setminus\mathcal{V}\right).$$

Since the effective SNRs $\gamma_1^{(1)}, \ldots, \gamma_1^{(K)}$ are mutually independent, we get

$$\Pr\left(\mathcal{V}\right) = \left[\prod_{j\in\mathcal{V}} \left(1 - F_{\gamma_1^{(j)}}\left(\lambda_1\right)\right)\right] \prod_{j\in\mathcal{B}\setminus\mathcal{V}} F_{\gamma_1^{(j)}}\left(\lambda_1\right), \quad (28)$$

where $F_{\gamma_1^{(j)}}\left(.\right)$ denotes the CDF of $\gamma_1^{(j)}$.

Upon rearranging terms and simplifying, the first probability term in (27) reduces to

$$\begin{aligned}
&\Pr\left(\prod_{j\in\mathcal{V}} c_1 \exp\left(-d_1\gamma_1^{(j)}\right) \leq \varepsilon | \mathcal{V}\right) \\
&= \Pr\left(\sum_{j\in\mathcal{V}} \gamma_1^{(j)} \geq \frac{|\mathcal{V}|\log(c_1) - \log(\varepsilon)}{d_1} | \gamma_1^{(j)} \geq \lambda_1, \forall j \in \mathcal{V}\right), \\
&= \frac{\Pr\left(\sum_{j\in\mathcal{V}} \gamma_1^{(j)} \geq \frac{|\mathcal{V}|\log(c_1)-\log(\varepsilon)}{d_1}; \gamma_1^{(j)} \geq \lambda_1, \forall j \in \mathcal{V}\right)}{\Pr\left(\gamma_1^{(j)} \geq \lambda_1, \forall j \in \mathcal{V}\right)}.
\end{aligned} \quad (29)$$

Let $\mathcal{R}_1 = \left\{\gamma_1^{(j)}, \forall j \in \mathcal{V} : \sum_{j\in\mathcal{V}} \gamma_1^{(j)} \geq \frac{|\mathcal{V}|\log(c_1)-\log(\varepsilon)}{d_1}\right\}$, $\mathcal{R}_2 = \left\{\gamma_1^{(j)}, \forall j \in \mathcal{V} : \gamma_1^{(j)} > \lambda_1, \forall j \in \mathcal{V}\right\}$, and

$$\mathcal{R}_3 = \left\{\gamma_1^{(j)}, \forall j \in \mathcal{V} : \lambda_1 \leq \gamma_1^{(j)} \leq \frac{|\mathcal{V}|\log(c_1) - \log(\varepsilon)}{d_1} - (|\mathcal{V}|-1)\lambda_1, \forall j \in \mathcal{V}\right\}.$$

We can show that $\mathcal{R}_2 \backslash \mathcal{R}_3 \subseteq \mathcal{R}_1 \cap \mathcal{R}_2$. Hence, $\Pr(\mathcal{R}_1 \cap \mathcal{R}_2) \geq \Pr(\mathcal{R}_2 \setminus \mathcal{R}_3) \geq \Pr(\mathcal{R}_2) - \Pr(\mathcal{R}_3)$. Therefore,

$$\begin{aligned}
&\Pr\left(\sum_{j\in\mathcal{V}} \gamma_1^{(j)} \geq \frac{|\mathcal{V}|\log(c_1)-\log(\varepsilon)}{d_1}; \gamma_1^{(j)} \geq \lambda_1, \forall j \in \mathcal{V}\right) \\
&\geq \Pr\left(\gamma_1^{(j)} \geq \lambda_1, \forall j \in \mathcal{V}\right) \\
&\quad - \Pr\Big(\lambda_1 \leq \gamma_1^{(j)} \leq \frac{|\mathcal{V}|\log(c_1)-\log(\varepsilon)}{d_1} \\
&\quad - (|\mathcal{V}|-1)\lambda_1, \forall j \in \mathcal{V}\Big), \\
&= \left[\prod_{j\in\mathcal{V}} \left(1 - F_{\gamma_1^{(j)}}(\lambda_1)\right)\right] \\
&\quad - \prod_{j\in\mathcal{V}} \left[F_{\gamma_1^{(j)}}\left(\frac{1}{d_1}\log\left(\frac{c_1^{|\mathcal{V}|}}{\varepsilon}\right) - (|\mathcal{V}|-1)\lambda_1\right)\right. \\
&\quad \left. - F_{\gamma_1^{(j)}}(\lambda_1)\right]. \qquad (30)
\end{aligned}$$

Thus, all the probability terms can be written in terms of the CDF $F_{\gamma_1^{(j)}}(.)$.

Expression for $F_{\gamma_1^{(j)}}(.)$*:* From (3), $\gamma_1^{(j)} = -\beta_1 \log\left(Y_1^{(j)}\right)$, where

$$Y_1^{(j)} = \frac{1}{N_1} \sum_{i\in\mathcal{D}_j} \exp\left(-\frac{\Gamma_{ij}}{\beta_1}\right). \qquad (31)$$

$Y_1^{(j)}$ is a sum of N_1 positive RVs with a finite support of $[0,1]$. It can be approximated by a beta RV, as per Papoulis' central limit approximation [33]. The beta parameters $a_1^{(j)}$ and $b_1^{(j)}$ of $Y_1^{(j)}$ can be expressed as follows [34, Ch. 25]:

$$a_1^{(j)} = \frac{\mathbb{E}\left[Y_1^{(j)}\right]\left(\mathbb{E}\left[Y_1^{(j)}\right] - \mathbb{E}\left[\left(Y_1^{(j)}\right)^2\right]\right)}{\mathbb{E}\left[\left(Y_1^{(j)}\right)^2\right] - \left(\mathbb{E}\left[Y_1^{(j)}\right]\right)^2}, \qquad (32)$$

$$b_1^{(j)} = \frac{\left(1 - \mathbb{E}\left[Y_1^{(j)}\right]\right)\left(\mathbb{E}\left[Y_1^{(j)}\right] - \mathbb{E}\left[\left(Y_1^{(j)}\right)^2\right]\right)}{\mathbb{E}\left[\left(Y_1^{(j)}\right)^2\right] - \left(\mathbb{E}\left[Y_1^{(j)}\right]\right)^2}. \qquad (33)$$

From (31), $\mathbb{E}\left[Y_1^{(j)}\right] = (1/N_1)\sum_{i\in\mathcal{D}_j} \Psi_{\Gamma_{ij}}\left(-\beta_1^{-1}\right)$. Since Γ_{ij} is an exponential RV with parameter $\omega^2/(P_T\sigma_j^2)$, its MGF can be shown to be $\Psi_{\Gamma_{ij}}(s) = \omega^2/(\omega^2 - sP_T\sigma_j^2)$, for $\Re\{s\} < \omega^2/(P_T\sigma_j^2)$. Hence, $\Psi_{\Gamma_{ij}}\left(-\beta_1^{-1}\right) = \omega^2\beta_1/(\omega^2\beta_1 + P_T\sigma_j^2) \triangleq g_j$. It follows that $\mathbb{E}\left[Y_1^{(j)}\right] = g_j$. Similarly,

$$\mathbb{E}\left[\left(Y_1^{(j)}\right)^2\right] = \frac{1}{N_1^2}\sum_{i\in\mathcal{D}_j} \mathbb{E}\left[e^{-\frac{2\Gamma_{ij}}{\beta_1}}\right] + \frac{1}{N_1^2}\sum_{i\in\mathcal{D}_j}\sum_{l\in\mathcal{D}_j, l\neq i} \mathbb{E}\left[e^{-\frac{\Gamma_{ij}+\Gamma_{lj}}{\beta_1}}\right]. \qquad (34)$$

From above, $\mathbb{E}\left[e^{-2\Gamma_{ij}/\beta_1}\right] = \Psi_{\Gamma_{ij}}\left(-2\beta_1^{-1}\right)$ and $\Psi_{\Gamma_{ij}}\left(-2\beta_1^{-1}\right) = \omega^2\beta_1/(\omega^2\beta_1 + 2P_T\sigma_j^2) \triangleq s_j$. As Γ_{ij} and Γ_{lj} are independent for $i \neq l$, $\mathbb{E}\left[e^{-(\Gamma_{ij}+\Gamma_{lj})/\beta_1}\right] = \Psi_{\Gamma_{ij}}\left(-\beta^{-1}\right)\Psi_{\Gamma_{lj}}\left(-\beta^{-1}\right) = g_j^2$. Substituting these moment expressions in (32) and (33) yields (11) and (12). Lastly,

$$F_{\gamma_1^{(j)}}(x) = \Pr\left(Y_1^{(j)} \geq e^{-\frac{x}{\beta_1}}\right) = 1 - B\left(e^{-\frac{x}{\beta_1}}, a_1^{(j)}, b_1^{(j)}\right). \qquad (35)$$

B. Brief Derivation of Result 2

The achievability is given by $A = \Pr\left(c_1 \exp\left(-d_1\gamma_1^{\mathcal{B}}\right) \leq \varepsilon\right) = \Pr\left(\gamma_1^{\mathcal{B}} \geq (\log(c_1\varepsilon^{-1})/d_1)\right)$. The effective SNR $\gamma_1^{\mathcal{B}}$ can be written as $\gamma_1^{\mathcal{B}} = -\beta_1\log\left(Y_1^{\mathcal{B}}\right)$, where $Y_1^{\mathcal{B}} = (1/N_1)\sum_{i=1}^{N_1} \exp\left(-\left(\sum_{j\in\mathcal{B}}\Gamma_{ij}\right)/\beta_1\right)$. Thus,

$$A = \Pr\left(Y_1^{\mathcal{B}} < \left(c_1\varepsilon^{-1}\right)^{\frac{1}{d_1\beta_1}}\right) = F_{Y_1^{\mathcal{B}}}\left(\left(c_1\varepsilon^{-1}\right)^{\frac{1}{d_1\beta_1}}\right), \qquad (36)$$

where $F_{Y_1^{\mathcal{B}}}(.)$ is the CDF of $Y_1^{\mathcal{B}}$. As in Appendix A, $Y_1^{\mathcal{B}}$ is a beta RV with parameters $a_1^{\mathcal{B}}$ and $b_1^{\mathcal{B}}$. Hence, $A = B\left(\left(c_1\varepsilon^{-1}\right)^{\frac{1}{d_1\beta_1}}, a_1^{\mathcal{B}}, b_1^{\mathcal{B}}\right)$.

The beta parameters $a_1^{\mathcal{B}}$ and $b_1^{\mathcal{B}}$ can be computed from the first two moments of $Y_1^{\mathcal{B}}$ using formulae similar to (32) and (33). The moments of $Y_1^{(\mathcal{B})}$ are given as follows:

$$\mathbb{E}\left[Y_1^{\mathcal{B}}\right] = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\prod_{j\in\mathcal{B}} \mathbb{E}\left[\exp\left(-\frac{\Gamma_{ij}}{\beta_1}\right)\right]\right) = \prod_{j\in\mathcal{B}} g_j, \qquad (37)$$

$$\begin{aligned}
\mathbb{E}\left[\left(Y_1^{\mathcal{B}}\right)^2\right] &= \frac{1}{N_1^2}\sum_{i=1}^{N_1}\left(\prod_{j\in\mathcal{B}} \mathbb{E}\left[e^{-\frac{2\Gamma_{ij}}{\beta_1}}\right]\right) \\
&\quad + \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{l=1,l\neq i}^{N_1} \mathbb{E}\left[e^{-\frac{\sum_{j\in\mathcal{B}}\Gamma_{ij}+\sum_{j\in\mathcal{B}}\Gamma_{lj}}{\beta_1}}\right], \\
&= \frac{1}{N_1}\left(\prod_{j\in\mathcal{B}} s_j\right) + \frac{N_1-1}{N_1}\left(\prod_{j\in\mathcal{B}} g_j\right)^2. \qquad (38)
\end{aligned}$$

C. Proof of Lemma 1

a) Since the error target is satisfied by m_j^*, using an MCS with a lower rate than m_j^* will require more resource elements and, thus, increase the number of PRBs that are pre-empted for the eMBB user. Hence, it is suboptimal.

b) For the same SNR, a larger MCS index implies a higher rate and a higher BLER. Hence, if $(m_j, \forall j \in \mathcal{S})$ cannot meet the BLER target, neither can $(m_j + \nu_j, \forall j \in \mathcal{S})$, where $\nu_j \geq 0$.

D. Brief Proof of Result 3

Let $\mathcal{V}$ denote the subset of BSs whose effective SNR $\gamma_{m_j}^{(j)}(t_2)$ of MCS m_j is greater than λ_{m_j}. Along lines similar to Appendix A, we have

$$\Pr\left(\prod_{j\in\mathcal{S}} \mathsf{BLER}_{m_j}\left(\gamma_{m_j}^{(j)}(t_2)\right) \le \varepsilon|\mathbf{Q}_{t_1}\right) = \sum_{\mathcal{V}\subseteq\mathcal{S}} \Pr\left(\sum_{j\in\mathcal{V}} d_{m_j}\gamma_{m_j}^{(j)}(t_2) \ge \theta_{\mathcal{V}}|\mathcal{V},\mathbf{Q}_{t_1}\right)\Pr\left(\mathcal{V}|\mathbf{Q}_{t_1}\right), \tag{39}$$

where $\theta_{\mathcal{V}} = \log\left(\varepsilon^{-1}\prod_{j\in\mathcal{V}} c_{m_j}\right)$. The first probability term in (39) can be written as

$$\Pr\left(\sum_{j\in\mathcal{V}} d_{m_j}\gamma_{m_j}^{(j)}(t_2) \ge \theta_{\mathcal{V}}|\mathcal{V},\mathbf{Q}_{t_1}\right) = \frac{\Pr\left(\sum_{j\in\mathcal{V}} d_{m_j}\gamma_{m_j}^{(j)}(t_2) \ge \theta_{\mathcal{V}};\gamma_{m_j}^{(j)}(t_2) > \lambda_{m_j},\forall j\in\mathcal{V}|\mathbf{Q}_{t_1}\right)}{\Pr\left(\gamma_{m_j}^{(j)}(t_2) > \lambda_{m_j},\forall j\in\mathcal{V}\mid\mathbf{Q}_{t_1}\right)}. \tag{40}$$

In a manner similar to Appendix A, we can show that

$$\begin{aligned}\Pr&\left(\sum_{j\in\mathcal{V}} d_{m_j}\gamma_{m_j}^{(j)}(t_2) \ge \theta_{\mathcal{V}};\gamma_{m_j}^{(j)}(t_2) > \lambda_{m_j},\forall j\in\mathcal{V}|\mathbf{Q}_{t_1}\right)\\ &\ge \Pr\left(\gamma_{m_j}^{(j)}(t_2) > \lambda_{m_j},\forall j\in\mathcal{V}\mid\mathbf{Q}_{t_1}\right) - \Pr\left(\lambda_{m_j} \le \gamma_{m_j}^{(j)}(t_2) \le \frac{\theta_{\mathcal{V}}-\zeta_j}{d_{m_j}},\forall j\in\mathcal{V}|\mathbf{Q}_{t_1}\right),\\ &= \left[\prod_{j\in\mathcal{V}}\left(1 - F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\lambda_{m_j}\right)\right)\right] - \prod_{j\in\mathcal{V}}\left[F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\frac{\theta_{\mathcal{V}}-\zeta_j}{d_{m_j}}\right) - F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\lambda_{m_j}\right)\right],\end{aligned} \tag{41}$$

where $F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}(.)$ is the CDF of $\gamma_{m_j}^{(j)}(t_2)$ conditioned on $\mathbf{Q}_{t_1}$. And, $\Pr\left(\mathcal{V}|\mathbf{Q}_{t_1}\right)$ simplifies to

$$\Pr\left(\mathcal{V}\mid\mathbf{Q}_{t_1}\right) = \Pr\left(\gamma_{m_j}^{(j)}(t_2) > \lambda_{m_j},\forall j\in\mathcal{V}\mid\mathbf{Q}_{t_1}\right) \times \prod_{j\in\mathcal{S}\setminus\mathcal{V}} F_{\gamma_{m_j}^{(j)}(t_2)|\mathbf{Q}_{t_1}}\left(\lambda_{m_j}\right). \tag{42}$$

Multiplying (41) and (42), and substituting in (40) and then in (39) yields (16).

E. Proof of Result 4

As before, $\gamma_{m_j}^{(j)}(t_2) = -\beta_{m_j}\log\left(Y_{m_j}(t_2)\right)$, where $Y_{m_j}(t_2) = (1/N_{m_j})\sum_{i\in\mathcal{D}_j}\exp\left(-\Gamma_{ij}(t_2)/\beta_{m_j}\right)$. Along lines similar to Appendix A, $Y_{m_j}(t_2)$ conditioned on $\mathbf{Q}_{t_1}$ is a beta RV with parameters a_{m_j} and b_{m_j}. Hence, the conditional CDF of $\gamma_{m_j}^{(j)}(t_2)$ can be rewritten as

$$F_{\gamma_1^{(j)}}(x) = \Pr\left(Y_{m_j}(t_2) \le e^{-\frac{x}{\beta_{m_j}}}\right) = B\left(e^{-\frac{x}{\beta_1}}, a_{m_j}, b_{m_j}\right). \tag{43}$$

Expressions for Beta Parameters: a_{m_j} and b_{m_j} can be written in terms of the conditional moments of $Y_{m_j}(t_2)$ using formulae similar to (32) and (33) as follows. The conditional mean $\mathbb{E}\left[Y_{m_j}(t_2)\mid\mathbf{Q}_{t_1}\right]$ is equal to $(1/N_{m_j})\sum_{i\in\mathcal{D}_j}\Psi_{\Gamma_{ij}(t_2)|\mathbf{Q}_{t_1}}(-\beta_{m_j}^{-1})$. Since $H_{ij}(t_1)$ and $H_{ij}(t_2)$ are jointly Gaussian with correlation coefficient $\rho(\tau)$, $\Gamma_{ij}(t_2)$ conditioned on $\Gamma_{ij}(t_1)$ is a weighted non-central chi-square RV with weight $\alpha_j = d_{m_j}P_T\left(1-\rho^2(\tau)\right)\sigma_j^2/(2\omega^2)$ [31, Ch. 9]. Therefore,

$$\Psi_{\Gamma_{ij}(t_2)|\mathbf{Q}_{t_1}}(s) = \frac{1}{1-2\alpha_j s}\exp\left(\frac{\alpha_j\delta_{ij}s}{1-2\alpha_j s}\right), \quad \text{for } \Re\{s\} < \frac{1}{2\alpha_j}. \tag{44}$$

Hence, $\Psi_{\Gamma_{ij}(t_2)|\mathbf{Q}_{t_1}}\left(-\beta_{m_j}^{-1}\right) = (\beta_{m_j}/(\beta_{m_j}+2\alpha_j))\exp\left(-\delta_{ij}\alpha_j/(\beta_{m_j}+2\alpha_j)\right) \triangleq g_{ij}$. Similarly,

$$\mathbb{E}\left[(Y_{m_j}(t_2))^2\mid\mathbf{Q}_{t_1}\right] = \frac{1}{N_{m_j}^2}\sum_{i\in\mathcal{D}_j} s_{ij} + \frac{1}{N_{m_j}^2}\sum_{i\in\mathcal{D}_j}\sum_{l\in\mathcal{D}_j,l\ne i} g_{ij}g_{lj}, \tag{45}$$

where $s_{ij} \triangleq \Psi_{\Gamma_{ij}(t_2)|\mathbf{Q}_{t_1}}\left(-2\beta_{m_j}^{-1}\right)$.

Substituting these in (32) and (33) and simplifying further yields (18) and (19).

F. Brief Proof of Result 5

From (2), the constraint can be written as

$$\begin{aligned}\Pr&\left(\mathsf{BLER}_m\left(\gamma_m^{\mathcal{B}}(t_2)\right) \le \varepsilon|\mathbf{Q}_{t_1}\right)\\ &= \Pr\left(\gamma_m^{\mathcal{B}}(t_2) > \frac{\log\left(\frac{c_m}{\varepsilon}\right)}{d_m}|\mathbf{Q}_{t_1}\right),\\ &= F_{Y'_m|\mathbf{Q}_{t_1}}\left(\left(\frac{c_m}{\varepsilon}\right)^{\frac{1}{\beta_m d_m}}\right),\end{aligned} \tag{46}$$

where $\gamma_m^{\mathcal{B}}(t_2) = -\beta_m\log(Y'_m)$, $Y'_m = (1/N_m)\sum_{i=1}^{N_m}\exp\left(-\Gamma_i^{\mathcal{B}}(t_2)/\beta_m\right)$, and $F_{Y'_m|\mathbf{Q}_{t_1}}(.)$ is the CDF of Y'_m conditioned on $\mathbf{Q}_{t_1}$. Since $H_{ij}(t_1)$ and $H_{ij}(t_2)$ are jointly complex Gaussian, $\left|\sum_{j\in\mathcal{B}} H_{ij}^*(t_1)H_{ij}(t_2)\right|$ conditioned on $\mathbf{Q}_{t_1}$ is a Rician RV with non-centrality parameter $\rho(\tau)(w_i(t_1))^2$ and scale parameter $\sqrt{\left(1-\rho^2(\tau)\right)\left(\sum_{j\in\mathcal{B}}\sigma_j^2|H_{ij}(t_1)|^2\right)/2}$. As a result, the SNR $\Gamma_i^{\mathcal{B}}(t_2) = (P_T/(\omega^2(w_i(t_1))^2))\left|\sum_{j\in\mathcal{B}} H_{ij}^*(t_1)H_{ij}(t_2)\right|^2$ of subband i, which is the square of a Rician RV, is a weighted non-central chi-square RV with weight α'_i and non-centrality parameter δ'_i when conditioned on $\mathbf{Q}_{t_1}$. Hence, $F_{Y'_m|\mathbf{Q}_{t_1}}\left((c_m/\varepsilon)^{\frac{1}{\beta_m d_m}}\right) = B\left((c_m/\varepsilon)^{\frac{1}{\beta_m d_m}}, a'_m, b'_m\right)$.

The moments of Y'_m can be expressed in terms of the MGF of $\Gamma_i^{\mathcal{B}}(t_2)$ conditioned on $\mathbf{Q}_{t_1}$, which is given in Appendix D.

REFERENCES

[1] G. Saikesava and N. B. Mehta, "MCS selection for multi-connectivity and eMBB-URLLC coexistence in time-varying frequency-selective fading channels," in *Proc. IEEE ICC*, May 2022, pp. 1–6.

[2] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2018.

[3] P. Popovski et al., "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[4] M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-connectivity as an enabler for reliable low latency communications—An overview," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 156–169, 1st Quart., 2020.

[5] T. Hößler, P. Schulz, E. A. Jorswieck, M. Simsek, and G. P. Fettweis, "Stable matching for wireless URLLC in multi-cellular, multi-user systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5228–5241, Aug. 2020.

[6] D. Lee et al., "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[7] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. ISWCS*, Aug. 2019, pp. 607–612.

[8] A. Wolf, P. Schulz, M. Dörpinghaus, J. C. S. S. Filho, and G. Fettweis, "How reliable and capable is multi-connectivity?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1506–1520, Feb. 2019.

[9] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[10] A. H. Mahdi, T. Hößler, N. Franchi, and G. Fettweis, "Multi-connectivity for reliable wireless industrial communications: Gains and limitations," in *Proc. IEEE WCNC*, May 2020, pp. 1–7.

[11] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio," in *Proc. IEEE WCNC*, Apr. 2019, pp. 1–6.

[12] D. Öehmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Modeling and analysis of intra-frequency multi-connectivity for high availability in 5G," in *Proc. IEEE VTC Spring*, Jun. 2018, pp. 1–7.

[13] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE VTC-Fall*, Sep. 2017, pp. 1–6.

[14] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Feb. 2020.

[15] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Resource allocation of URLLC and eMBB mixed traffic in 5G networks: A deep learning approach," in *Proc. GLOBECOM*, Dec. 2020, pp. 1–6.

[16] Y. Huang, Y. T. Hou, and W. Lou, "A deep-learning-based link adaptation design for eMBB/URLLC multiplexing in 5G NR," in *Proc. IEEE INFOCOM*, May 2021, pp. 1–10.

[17] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1080–1093, Feb. 2021.

[18] Q. Shang, F. Liu, C. Feng, R. Zhang, and S. Zhao, "A BP neural network based punctured scheduling scheme within mini-slots for joint URLLC and eMBB traffic," in *Proc. GlobalSIP*, Nov. 2019, pp. 1–5.

[19] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "EMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.

[20] M. Alsenwi, S. R. Pandey, Y. K. Tun, K. T. Kim, and C. S. Hong, "A chance constrained based formulation for dynamic multiplexing of eMBB-URLLC traffics in 5G new radio," in *Proc. IEEE ICOIN*, Jan. 2019, pp. 108–113.

[21] R. Liu, G. Yu, J. Yuan, and G. Y. Li, "Resource management for millimeter-wave ultra-reliable and low-latency communications," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1094–1108, Feb. 2021.

[22] A. H. Mahdi, T. Hößler, L. Scheuvens, N. Franchi, and G. Fettweis, "Multi-connectivity management for mobile ultra-reliable low-latency communications," in *Proc. IEEE WSA*, Feb. 2020, pp. 1–6.

[23] G. L. Stüber, *Principles of Mobile Communication*, 4th ed. Atlanta, GA, USA: Springer, 2017.

[24] V. Kumar and N. B. Mehta, "Modeling and analysis of differential CQI feedback in 4G/5G OFDM cellular systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2361–2373, Apr. 2019.

[25] Q. Liu, S. Zhou, and G. B. 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.

[26] *NR; Physical Layer Procedures for Data*, document TR 38.214, Version 15.3.0, 3rd Generation Partnership Project (3GPP), Oct. 2018.

[27] S. Lagen, K. Wanuga, H. Elkotby, S. Goyal, N. Patriciello, and L. Giupponi, "New radio physical layer abstraction for system-level simulations of 5G networks," in *Proc. IEEE ICC*, Jun. 2020, pp. 1–7.

[28] W. Anwar, K. Kulkarni, N. Franchi, and G. Fettweis, "Physical layer abstraction for ultra-reliable communications in 5G multi-connectivity networks," in *Proc. IEEE PIMRC*, Sep. 2018, pp. 1–6.

[29] J. Fan, Q. Yin, G. Y. Li, B. Peng, and X. Zhu, "MCS selection for throughput improvement in downlink LTE systems," in *Proc. ICCCN*, Jul. 2011, pp. 1–5.

[30] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, 9th ed. New York, NY, USA: Dover, 1972.

[31] A. Goldsmith, *Wireless Communications*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[32] *Study on New Radio (NR) Access Technology*, document 38.912, Version 15.0.0, 3rd Generation Partnership Project (3GPP), Jul. 2018.

[33] J. Francis and N. B. Mehta, "EESM-based link adaptation in point-to-point and multi-cell OFDM systems: Modeling and analysis," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 407–417, Jan. 2014.

[34] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. New York, NY, USA: Wiley, 1995.

Govindu Sai Kesava received the B.Tech. degree from the Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India, in 2017. He is currently pursuing the M.Tech. (Research) degree with the Department of Electrical Communication Engineering, Indian Institute of Science (IISc), Bengaluru. His research interests include design and analysis of 5G wireless networks.

Neelesh B. Mehta (Fellow, IEEE) received the B.Tech. degree from the Indian Institute of Technology (IIT) Madras in 1996 and the M.S. and Ph.D. degrees from the California Institute of Technology, Pasadena, CA, USA, in 1997 and 2001, respectively.

He worked in USA as a Research Scientist at AT&T Research Labs, Broadcom Corp., and Mitsubishi Electric Research Labs, until 2007. He is currently a Professor with the Department of Electrical Communication Engineering, Indian Institute of Science (IISc), Bengaluru. His research interests include design, modeling, analysis, and optimization of 5G and beyond wireless systems. He is a fellow of the Indian National Science Academy (INSA), the Indian National Academy of Engineering (INAE), and the National Academy of Sciences India (NASI). He is also a Member Secretary of the IEEE ComSoc Nominations and Elections Committee. He was a recipient of the IIT Roorkee's Khosla National Award, the Shanti Swarup Bhatnagar Award, the Vikram Sarabhai Research Award, the DST-Swarnajayanti Fellowship, and the NASI-Scopus Young Scientist Award. He currently serves as the Chair for the Steering Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has served on the Executive Editorial Committee for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2014 to 2017 and the Chair from 2017 to 2018. He also served on the Board of Governors for IEEE ComSoc from 2012 to 2015 and the IEEE ComSoc Awards Committee from 2018 to 2020.