Received 7 March 2023; accepted 5 April 2023. Date of publication 14 April 2023; date of current version 1 May 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3266581

On the Coexistence of eMBB and URLLC in Multi-Cell Massive MIMO

GIOVANNI INTERDONATO[1,2] (Member, IEEE), STEFANO BUZZI[1,2,3] (Senior Member, IEEE), CARMEN D'ANDREA[1,2] (Member, IEEE), LUCA VENTURINO[1,2] (Senior Member, IEEE), CIRO D'ELIA[1,2], AND PAOLO VENDITTELLI[4]

[1]Department of Electrical and Information Engineering, University of Cassino and Southern Latium, 03043 Cassino, Italy

[2]Consorzio Nazionale Interuniversitario per le Telecomunicazioni, 43124 Parma, Italy

[3]Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy

[4]TIM S.p.A., 20133 Milan, Italy

CORRESPONDING AUTHOR: G. INTERDONATO (e-mail: giovanni.interdonato@unicas.it)

This work was supported by the Ministero delle Imprese e del Made in Italy (former MISE) through the project "Smart Urban Mobility Management" (5G-SUMMA), Asse II, Supporto alle Tecnologie Emergenti.

ABSTRACT The non-orthogonal coexistence between the enhanced mobile broadband (eMBB) and the ultra-reliable low-latency communication (URLLC) in the downlink of a multi-cell massive MIMO system is rigorously analyzed in this work. We provide a unified information-theoretic framework blending an infinite-blocklength analysis of the eMBB spectral efficiency (SE) in the ergodic regime with a finite-blocklength analysis of the URLLC error probability relying on the use of mismatched decoding, and of the so-called saddlepoint approximation. Puncturing (PUNC) and superposition coding (SPC) are considered as alternative downlink coexistence strategies to deal with the inter-service interference, under the assumption of only statistical channel state information (CSI) knowledge at the users. eMBB and URLLC performances are then evaluated over different precoding techniques and power control schemes, by accounting for imperfect CSI knowledge at the base stations, pilot-based estimation overhead, pilot contamination, spatially correlated channels, the structure of the radio frame, and the characteristics of the URLLC activation pattern. Simulation results reveal that SPC is, in many operating regimes, superior to PUNC in providing higher SE for the eMBB yet achieving the target reliability for the URLLC with high probability. Moreover, PUNC might cause eMBB service outage in presence of high URLLC traffic loads. However, PUNC turns to be necessary to preserve the URLLC performance in scenarios where the multi-user interference cannot be satisfactorily alleviated.

INDEX TERMS Enhanced mobile broadband, error probability, massive MIMO, mismatched decoding, network availability, non-orthogonal multiple access, puncturing, saddlepoint approximation, spectral efficiency, superposition coding, ultra-reliable low-latency communications.

I. INTRODUCTION

WITH the advent of the mobile application ecosystem and the resulting increase of the data-processing and storage capabilities of the smart devices, several heterogeneous services have emerged setting various stringent communication requirements in terms of data rates, latency, reliability and massive connectivity. These requirements and related use cases have been summarized by the 3rd Generation Partnership Project (3GPP) into three macro services, namely *enhanced mobile broadband* (eMBB), ultra-reliable low-latency communications (URLLC) and *massive machine-type communications* (mMTC) [1]. eMBB services require high-peak data-rate and stable connectivity, and include most of the everyday usage applications: entertainment, multimedia, communication, collaboration, mapping, Web-surfing, etc. URLLC services demand an one-way radio

TABLE 1. Features of the 5G use cases.

	eMBB	URLLC	mMTC
characteristics	high rate, moderate reliability	low latency, ultra reliability, low rate	low rate, large connectivity
traffic	large payload, several devices	small payload, few devices	small payload, massive devices
activation pattern	stable	intermittent	intermittent
time span	long, multiple resources	short, slot	long, multiple resources
frequency span	single/multiple resources	multiple resources, diversity	single resource
scheduling	to prevent access collision	for high reliability	infeaseable
random access	if needed	to support intermittency	fundamental
target	maximize data rate	meet latency and reliability requirements	maximize supported arrival rate
reliability requirement	$\sim 10^{-3}$	$\sim 10^{-5}$	$\sim 10^{-1}$
applications	video streaming, augmented reality, entertainment	connected factories, traffic safety, autonomous vehicles, telemedicine	internet of things, low-power sensors, smart cities

latency down to 1 ms with a general reliability requirement of 99.999%, and include real-time and time-critical applications, such as autonomous driving, automation control, augmented reality, video and image processing, etc. mMTC services enable connectivity between a vast number of miscellaneous devices, and include applications such as smart grids, traffic management systems, environmental monitoring, etc.

5G started to roll out variously as an eMBB service, essentially like a faster version of LTE, whereas mMTC and URLLC requirements continue to be refined and will materialize within the next decade, although some experimental activities are already taking place in many parts of the world.[1] Academic research and industrial standardization is currently interested at different coexistence mechanisms for such heterogeneous services, apparently moving apart from the initial vision of a *sliced* network [2]. Slicing the network basically means allocating orthogonal resources (storage, computing, radio communications, etc.) to heterogeneous services so that to guarantee their mutual isolation. This approach is, in broad sense, generally known as *orthogonal multiple access* (OMA). As an interesting alternative to orthogonal resource allocation, non-orthogonal OMA (NOMA) is gaining increasing importance especially with respect to the allocation of the *radio access network* (RAN) communication resources. The conventional approach to slice the RAN is to separate eMBB, mMTC, and URLLC services in time and/or frequency domains, whereas NOMA relies on efficient coexistence strategies wherein heterogeneous services share the same time-frequency resources, being separated in the power and spatial domain. In this regard, the terminology *Heterogeneous* OMA (H-OMA) is often adopted [2] to distinguish the orthogonal resource allocation of heterogeneous services from that of the same type, referred to as OMA. (The same distinction applies to H-NOMA with respect to NOMA.)

Massive MIMO [3], [4], [5] is a technology that uses a very large number of co-located antennas at the base stations (BSs) to coherently and simultaneously serve multiple users over the same radio resources. The users are multiplexed in the spatial domain by using beamforming techniques that enable high-directivity transmission and reception. The use of many antennas also triggers the *favorable propagation* which further reduces the multi-user interference and the *channel hardening* which reduces the random fluctuations of the effective channel gain. As a consequence, there is no need to adopt intricate signal processing techniques to deal with the multi-user interference. Such an aggressive spatial multiplexing along with the intrinsic practicality and scalability of the massive MIMO technology leads to high levels of energy and spectral efficiency, spatial diversity, link reliability and connectivity.

The primary focus of the massive MIMO research has been on increasing the user data rates, thereby targeting the eMBB requirements. Lately, some studies have highlighted the significant benefits that massive MIMO is able to provide to URLLC [6], [7], [8] by reducing the outage and error probability, and therefore increasing the link reliability. Higher reliability results to less retransmissions which, in turn, translates to a lower latency. mMTC also benefits from massive MIMO technology [7], [9] by capitalizing on the high energy efficiency to increase devices' battery lifetime. Besides, favorable propagation enables an aggressive spatial multiplexing of the mMTC devices, facilitating the detection and the random access procedures.

A. RELATED WORKS

Coexistence between heterogeneous services has been initially studied in systems wherein a single-antenna BS serves multiple heterogeneous users. In [2], Popovski et al. proposed a first tractable communication-theoretic model that captures the key features of eMBB, URLLC and mMTC traffic. (These features are summarized in Table 1.) Specifically, [2] analyzes two scenarios for a *single-cell* model: (*i*) slicing for URLLC and eMBB, and (*ii*) slicing for mMTC and eMBB. The downlink multiplexing of URLLC and eMBB is studied in [10] by abstracting the operation at the physical layer. Coexistence mechanisms between URLLC and eMBB traffic, based on the *puncturing* technique, have been proposed

1. See, e.g., the funding programs from the Italian former Ministry of Economic Development, as well as those of other European Countries, the EU, USA, China and Japan.

in [11] for the uplink of a *multi-cell* network wherein a simplified Wyner channel model with no fading was assumed. As for multi-user MIMO systems, in [12] a null-space-based spatial preemptive scheduler for joint URLLC and eMBB traffic is proposed for cross-objective optimization. A similar study but for a distributed setup was conducted in [13] where a joint user association and resource allocation problem is formulated for the downlink of a fog network, considering the coexistence of URLLC and eMBB services for *Internet-of-Things* (IoT) applications.

The coexistence between eMBB and URLLC is of most interest [14], [15], [16], [17], [18], and is mainly handled with three alternative techniques, herein listed in descending order of complexity:

- *successive interference cancellation* (SIC), with which the receiver iteratively decode and remove the contributions of a specific service from the cumulative received signal. This approach requires that the receiver has access to the *channel state information* (CSI) to be able to perform the multi-stage decoding, with decreasing levels of interference, to the required successful decoding probability.
- *puncturing* (PUNC), consisting in preventing the inter-service interference. In the downlink, whenever the transmitter has to transmit a URLLC signal, then the eMBB signals are dropped over the channel uses involved by the URLLC transmission. In the uplink, the receiver uses an erasure decoder to discard the eMBB signals, provided that it is able to detect the presence of URLLC transmissions, e.g., via energy detection.
- *superposition coding* (SPC), with which the transmitter simply sends a linear combination of eMBB and URLLC signals. At the receiver, both for the uplink and the downlink, the inter-service interference is treated as uncorrelated noise (TIN). Again, this approach requires the receiver to be able to detect the presence of the undesired transmissions.

In [14] the coexistence of URLLC and eMBB services in the uplink of a cloud radio access network (C-RAN) architecture with shared analog fronthaul links is analyzed, accounting for SIC, puncturing, and TIN. This work provides an information-theoretic study in the performance of URLLC and eMBB traffic under both H-OMA and H-NOMA, by considering standard cellular models with additive Gaussian noise links and a finite inter-cell interference. A similar analysis is conducted in [19] including both uplink and downlink of C-RAN without analog fronthaul but considering practical aspects, such as fading, the lack of CSI for URLLC transmitters, rate adaptation for eMBB transmitters and finite fronthaul capacity. Abreu et al. in [16] analyzes both the H-OMA and H-NOMA options for eMBB traffic, and grant-free URLLC in the uplink accounting for minimum mean square error (MMSE) receivers with and without SIC, and under the assumption of Rayleigh fading channels. Recently, [17] proposed an approach to improve the supported loads for URLLC in the uplink, for both H-OMA and H-NOMA in presence of eMBB traffic, showing the superiority of H-NOMA in ensuring the reliability requirements of both the services. A similar analysis but for the downlink is conducted in [18], [20] where optimal resource allocation strategies and H-NOMA are combined to satisfy the eMBB and URLLC QoS constraints, under the assumption of perfect eMBB CSI and statistical URLLC CSI knowledge.

The information-theoretic framework used by the aforementioned works to characterize the performance achieved by eMBB and URLLC users cannot be applied to massive MIMO scenarios, for different reasons. Establishing the rate (or the spectral efficiency) of the eMBB users in the *ergodic (infinite-blocklength) regime*, upon the block-fading channel model, is sound as the eMBB codewords span an infinite number of independent fading realizations. Nevertheless, as per the performance of the URLLC users in a quasi-static fading scenario, the use of the outage capacity, whose analysis includes infinite-blocklength assumptions, leads to an inaccurate evaluation of the error probability, as demonstrated in [8]. In addition, outage capacity analyses do not capture the effects of the CSI acquisition overhead when pilots are used to estimate the uplink channel. As an alternative, finite-blocklength analyses have been proposed for URLLC in conventional cellular networks [18], [20], co-located massive MIMO networks [21], [22] and cell-free massive MIMO networks [23], and rely on the information-theoretic bounds and tools developed in [24], e.g., the well known *normal approximation*. However, the work in [8] proved that the normal approximation is not accurate in the region of low error probabilities of interest in URLLC ($< 10^{-4}$), especially as the number of antennas at the BS increases, and in presence of imperfect CSI. Importantly, Östman et al. in [8] provided a more rigorous finite-blocklength information-theoretic framework relying on the use of a mismatched decoding [25], and of the *saddlepoint approximation* [26] for evaluating the error probability of the URLLC users in co-located massive MIMO systems. This framework, priory developed for wireless fading channels in [27], [28], [29], accounts for linear signal processing, imperfect CSI and instantaneous channel estimation error, and additive uncorrelated noise including multi-user interference. However, the analysis of [8] is limited to the URLLC regime, and the coexistence with the eMBB is yet to be investigated under a unified information-theoretic framework.

B. CONTRIBUTIONS

Our contributions can be summarized as follows.

- We investigate the non-orthogonal multiplexing of the eMBB and the URLLC, in the downlink of a multi-cell massive MIMO system, by providing a novel unified information-theoretic framework that combines an infinite-blocklength analysis to assess the SE of the eMBB and a finite-blocklength analysis to assess the error probability of the URLLC.

- Unlike prior works wherein the URLLC performance is inappropriately evaluated by the use of the outage capacity analysis or the error probability obtained via the normal approximation, in this work the finite-blocklength information-theoretic analysis relies on the results and tools established in [8], where mismatched receivers and saddlepoint approximation are assumed, but the coexistence between URLLC and eMBB was not investigated. Moreover, in contrast to [8], our analysis realistically characterizes the URLLCs with random activation patterns.
- The proposed unified framework accommodates two alternative coexistence strategies: PUNC and SPC. The former prevents the inter-service interference to protect the URLLC reliability, whereas the latter accepts it to maintain the eMBB service. To the best of the authors' knowledge, this is the first work investigating the coexistence between eMBB and URLLC, whose analytical framework simultaneously accounts for imperfect CSI acquisition via uplink pilot transmissions, pilot contamination and pilot overhead, spatially correlated channels and the lack of CSI at the users.
- We numerically evaluate the performance achieved by PUNC and SPC under different precoding schemes and, in contrast to [8] which assumes fixed equal power allocation, under advanced power allocation strategies, such as weighted fractional power allocation and optimal power allocation maximizing the product SINR throughout the network. The coexistence between eMBB and URLLC is explored in various scenarios, including different configurations of the radio frame, and different URLLC random activation patterns.
- Pilot contamination among URLLC users is particularly destructive. This led us to devise a pilot assignment policy that prioritizes the URLLC users. In our approach, we primarily assign unique orthogonal pilots to the URLLC users, admitting pilot reuse only among eMBB users. If doable, orthogonal pilots are assigned within cells to prevent the intra-cell pilot contamination, and if the uplink training length is sufficiently large, then mutually orthogonal pilots are guaranteed to everyone. To the best of the authors' knowledge, there are no works in the related literature that customize the pilot assignment and the pilot reuse policies to enable an efficient coexistence of eMBB and URLLC.

C. PAPER OUTLINE

The remainder of this paper is organized as follows. In Section II, we introduce the system model of the multi-cell massive MIMO system, including the description of the uplink training and a unified framework for the data transmission stage accounting for both puncturing and superposition coding techniques. In Section III we present the information-theoretic analyses in the infinite-blocklength regime and finite-blocklength regime for the eMBB and the URLLC performance evaluation, respectively. Section IV details the precoding techniques and power allocation strategies to deal with the coexistence of eMBB and URLLC users. Simulation results and discussions are provided in Section V. Section VI sheds lights on candidate low-complexity decoding schemes for URLLC, while the main findings of this work are discussed in Section VII.

D. NOTATION

Vectors and matrices are denoted by boldface lowercase and boldface uppercase letters, respectively. Calligraphy uppercase letters denote sets, while $\mathbb{C}$ and $\mathbb{R}$ represent the sets of complex and real numbers, respectively. $\mathsf{E}\{\cdot\}$ indicates the expectation operator, while $\mathsf{Pr}\{\cdot\}$ denotes the probability of a set. x^+ represents the *positive part* function, namely $x^+ = \max\{x, 0\}$, and $\lfloor\cdot\rfloor$ denotes the *floor* function. The natural logarithm is indicated by $\log(\cdot)$ and $Q(\cdot)$ describes the Gaussian Q-function. $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ describes a circularly symmetric complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The superscripts $(\cdot)^{\mathsf{T}}$, $(\cdot)^*$ and $(\cdot)^{\mathsf{H}}$ denote the transpose, the conjugate and the conjugate transpose (Hermitian) operators, respectively. $\mathrm{tr}(\mathbf{A})$ indicates the trace of the matrix $\mathbf{A}$, while $\|\mathbf{a}\|$ denotes the ℓ_2-norm of the vector $\mathbf{a}$. The notation $[\mathbf{A}]_{:,i}$ indicates the ith column of the matrix $\mathbf{A}$. $\mathbf{I}_N$ represents the identity matrix of size $N \times N$. Table 2 introduces the notation definition used in the system model of this paper.

II. SYSTEM MODEL

Let us consider a multi-cell massive MIMO system with L cells, each one served by a BS that is placed at the cell-center. Each cell covers a square area of $D \times D$ km^2. An arbitrary BS j, $j = 1, \ldots, L$, is equipped with M_j co-located antennas, and provides service to K_j users, with $M_j \gg K_j$ so that interference suppression can be efficiently carried out by exploiting the spatial degrees of freedom. A fraction $0 \leq \alpha_j \leq 1$ of the K_j users requests a URLLC service, e.g., a vehicle in *cellular vehicle-to-everything* (C-V2X) use cases for intelligent transportation systems, or a machine in factory automation use cases for "Industry 4.0". Letting $K_j^{\mathrm{u}} = \alpha_j K_j$ be the number of URLLC users in cell j, then $K_j^{\mathrm{e}} = K_j - K_j^{\mathrm{u}}$ is the number of eMBB users in cell j. The set including the indices of the eMBB and URLLC users in cell j is denoted as $\mathcal{K}_j^{\mathrm{e}}$ and $\mathcal{K}_j^{\mathrm{u}}$, respectively.

A. TDD PROTOCOL AND FRAME STRUCTURE

The considered system operates in time-division duplex (TDD) mode to facilitate CSI acquisition and limit the estimation overhead. In addition, we assume that the channel is reciprocal as a result of a perfect calibration of the RF chains. By leveraging the channel reciprocity, the channel estimates acquired by the BS in the uplink are then utilized in the downlink to design the transmit precoding vectors. As channel hardening holds for co-located massive MIMO systems with sufficiently large antenna arrays in most of the propagation environments, we assume that the users do not

TABLE 2. System model notation.

Symbol	Description	Symbol	Description
L	n. of cells	K_j	n. of users in cell j
M_j	n. of antennas at BS j	K_j^{u}	n. of URLLC users/cell
α_j	$K_j^{\mathrm{u}}/K_j \in (0,1)$	K_j^{e}	n. of eMBB users/cell
τ_{c}	TDD frame length	$\mathcal{K}_j^{\mathrm{u}}$	URLLC users set in cell j
τ_{p}	UL training length	$\mathcal{K}_j^{\mathrm{e}}$	eMBB users set in cell j
τ_{d}	DL data trans. length	T	n. of slots in a TDD frame
f	pilot reuse factor	n_{d}	URLLC codeword length
$\boldsymbol{\phi}_{jk}$	UL pilot sequence of user k in cell j	$\mathbf{w}_{jk}$	precoding vector used by BS j to its user k
$\mathbf{h}_{lk}^{j}$	UL channel from user k in cell l to BS j	g_{jk}^{li}	precoded DL channel from BS l using $\mathbf{w}_{li}$ to user k in cell j
$\widehat{\mathbf{h}}_{lk}^{j}$	estimate of $\mathbf{h}_{lk}^{j}$	$\widehat{g}_{jk}^{li}$	estimate of g_{jk}^{li}
$\mathbf{R}_{lk}^{j}$	correl. matrix of $\mathbf{h}_{lk}^{j}$	$\widetilde{\mathbf{h}}_{lk}^{j}$	estimation error $\mathbf{h}_{lk}^{j}-\widehat{\mathbf{h}}_{lk}^{j}$
$\mathbf{C}_{lk}^{j}$	correl. matrix of $\widetilde{\mathbf{h}}_{lk}^{j}$	β_{lk}^{j}	average channel gain of $\mathbf{h}_{lk}^{j}$
p_{jk}^{p}	UL pilot power	$\rho_j^{\max}$	max transmit power at BS j
$\epsilon_{jk}^{\mathrm{dl}}$	DL error probability	η^{dl}	DL network availability
σ_{u}^2	UL noise variance	ρ_{ji}^{u}	DL power to URLLC user i
σ_{d}^2	DL noise variance	ρ_{jk}^{e}	DL power to eMBB user k

Symbol	Description
$\mathcal{P}_{jk}$	set of all the users using the same pilot as user k in cell j
A_{jk}^{t}	1 if URLLC user k in cell j is active in slot t, 0 otherwise
a_{u}	parameter of the Bernoulli distribution that draws A_{jk}^{t}
$\varsigma_{jk}^{\mathrm{e}}[n]$	data transmitted by BS j to eMBB user k in channel use n
$\varsigma_{ji}^{\mathrm{u}}[n]$	data transmitted by BS j to URLLC user i in channel use n
ν	exponent characterizing the fractional power allocation (FPA)
ω	FPA weight tuning the power allocated to the URLLC users

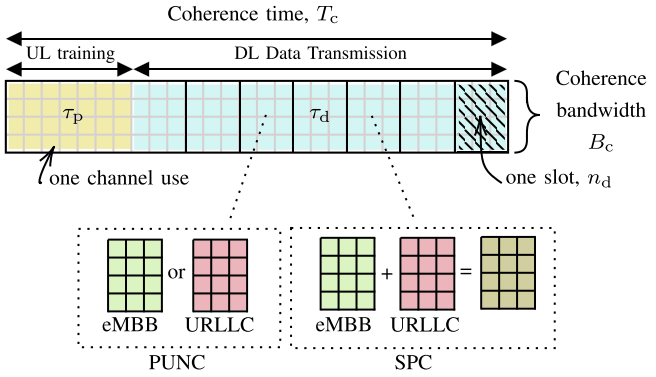

FIGURE 1. An illustration of the TDD frame assuming no uplink data transmission phase, and representing the resource allocation in case of puncturing (PUNC) and superposition coding (SPC) operation.

estimate the downlink channels, and reliably decode downlink data solely relying on the knowledge of the statistical CSI. Hence, the TDD protocol consists of three phases: (*i*) pilot-based uplink training, (*ii*) uplink data transmission, and (*iii*) downlink data transmission.

The time-frequency resources are structured in TDD frames, each one grouping a set of subcarriers and time samples over which the channel response is assumed being frequency-flat and time-invariant. The TDD frame must accommodate the aforementioned protocol phases and supporting all the users, thus its size is designed to match that of the smallest user's coherence block in the network. As shown in Fig. 1, the TDD frame consists of $\tau_{\mathrm{c}} = T_{\mathrm{c}}B_{\mathrm{c}}$ samples (or *channel uses*) where T_{c} is the coherence time and B_{c} is the coherence bandwidth. τ_{p} channel uses out of τ_{c} are spent for the uplink CSI acquisition, whereas the remaining channel uses are devoted to the uplink and downlink data transmission. Since, in this paper, we only focus on the downlink operation, we assume that $\tau_{\mathrm{d}} = \tau_{\mathrm{c}} - \tau_{\mathrm{p}}$ is the length of the downlink data transmission phase, without loss of generality. The latter is divided in T slots of equal length. As conventionally assumed in the *ergodic regime*, an eMBB transmission spans multiple (theoretically an infinite number of) TDD frames, wherein the channel realizations evolve independently according to the block-fading model. To evaluate the spectral efficiency achieved by the eMBB users, we look at a single TDD frame and resort to the information-theoretic bounds and tools in the infinite-blocklength regime [4], [5]. Whereas, URLLC transmissions are confined in time to meet the very strict latency requirements and are allowed to span only one slot. Hence, the number of channel uses in a slot equals the URLLC codeword length. We assume a random activation pattern of the URLLC users. Within a TDD frame, a URLLC user may be active in multiple slots. To characterize the error probability of the URLLC transmissions, we look separately at each single slot of a TDD frame and resort to the finite-blocklength information-theoretic bounds and tools presented in [8].

B. CHANNEL MODEL AND UPLINK TRAINING

The channel response between the k-th user in cell l and the BS in cell j is denoted by the M_j-dimensional complex-valued vector $\mathbf{h}_{lk}^{j}$. We assume correlated Rayleigh fading, that is $\mathbf{h}_{lk}^{j} \sim \mathcal{CN}\left(\mathbf{0}_{M_j}, \mathbf{R}_{lk}^{j}\right)$, where $\mathbf{R}_{lk}^{j} \in \mathbb{C}^{M_j \times M_j}$ is the positive semi-definite spatial correlation matrix. The corresponding average channel gain (or large-scale fading coefficient) is given by $\beta_{lk}^{j} = \mathrm{tr}(\mathbf{R}_{lk}^{j})/M_j$. Large-scale fading quantities are assumed to be known at the BS.

In the uplink training phase, each user transmits a pilot sequence that spans τ_{p} channel uses. The pilot sequence of user k in cell j is denoted by $\boldsymbol{\phi}_{jk} \in \mathbb{C}^{\tau_{\mathrm{p}}}$. All the pilot sequences are drawn from a set of τ_{p} mutually orthogonal pilots, thereby the inner product between two pilots equals either τ_{p} if the sequences are identical or 0 if they are mutually orthogonal. Notice that re-using the pilots throughout the network might be unavoidable as the share of the TDD frame reserved to the training is limited and, importantly, as the CSI acquisition overhead significantly degrades the spectral efficiency. Pilot reuse gives rise to additional interference, known as *pilot contamination* [3], that degrades the quality of the acquired CSI and correlates the channel estimates. The cumulative uplink signal received at BS j, denoted by

$\mathbf{Y}_j^{\mathsf{p}} \in \mathbb{C}^{M_j \times \tau_{\mathrm{p}}}$, reads

$$\mathbf{Y}_j^{\mathsf{p}} = \sum_{k=1}^{K_j} \sqrt{p_{jk}^{\mathsf{p}}} \mathbf{h}_{jk}^j \boldsymbol{\phi}_{jk}^{\mathsf{T}} + \sum_{\substack{l=1 \\ l \neq j}}^{L} \sum_{i=1}^{K_l} \sqrt{p_{li}^{\mathsf{p}}} \mathbf{h}_{li}^j \boldsymbol{\phi}_{li}^{\mathsf{T}} + \mathbf{N}_j^{\mathsf{p}}, \quad (1)$$

where p_{jk}^{p} is the transmit pilot power, and $\mathbf{N}_j^{\mathsf{p}}$ is the additive receiver noise with i.i.d. elements distributed as $\mathcal{CN}(0, \sigma_{\mathrm{u}}^2)$, with σ_{u}^2 being the receiver noise variance in the uplink. To estimate the channel of user k in its own cell, $\mathbf{h}_{jk}^j$, BS j correlates $\mathbf{Y}_j^{\mathsf{p}}$ with the known pilot sequence $\boldsymbol{\phi}_{jk}$ as

$$\begin{aligned} \mathbf{y}_{jjk}^{\mathsf{p}} &= \mathbf{Y}_j^{\mathsf{p}} \boldsymbol{\phi}_{jk}^* \\ &= \sqrt{p_{jk}^{\mathsf{p}}} \tau_{\mathrm{p}} \mathbf{h}_{jk}^j + \sum_{\substack{i=1 \\ i \neq k}}^{K_j} \sqrt{p_{ji}^{\mathsf{p}}} \mathbf{h}_{ji}^j \boldsymbol{\phi}_{ji}^{\mathsf{T}} \boldsymbol{\phi}_{jk}^* \\ &\quad + \sum_{\substack{l=1 \\ l \neq j}}^{L} \sum_{i=1}^{K_l} \sqrt{p_{li}^{\mathsf{p}}} \mathbf{h}_{li}^j \boldsymbol{\phi}_{li}^{\mathsf{T}} \boldsymbol{\phi}_{jk}^* + \mathbf{N}_j^{\mathsf{p}} \boldsymbol{\phi}_{jk}^*. \end{aligned} \quad (2)$$

In (2), the second term of the rightmost right-hand side represents the intra-cell pilot contamination term, while the third term quantifies the inter-cell pilot contamination. A conventional pilot allocation strategy consists in assigning mutually orthogonal pilots to users within the same cell, and re-using the pilot sequences over different cells [5]. This is a reasonable choice as intra-cell pilot contamination is presumably stronger than inter-cell pilot contamination. We let $\tau_{\mathrm{p}} = f \cdot \max_j K_j$ where f is referred to as *pilot reuse factor*. Importantly, in order not to jeopardize the ultra-reliability of the URLLC transmissions, we assume that unique orthogonal pilot sequences are assigned to all the URLLC users in the network, if doable (namely when $\tau_{\mathrm{p}} > \sum_{j=1}^{L} K_j^{\mathsf{e}}$). Summarizing, the pilot allocation strategy we propose primarily aims to prevent URLLC users from being affected of pilot contamination, and secondarily to prevent intra-cell pilot contamination. Finally, if τ_{p} is sufficiently large, that is $\tau_{\mathrm{p}} \geq \sum_{j=1}^{L} K_j$, then mutually orthogonal pilots can be guaranteed to everyone. Let us define the set

$$\mathcal{P}_{jk} = \left\{ (l, i) : \boldsymbol{\phi}_{li} = \boldsymbol{\phi}_{jk},\ l = 1, \ldots, L,\ i = 1, \ldots, K_l \right\}, \quad (3)$$

including the indices of all the users (and of the corresponding cells) that use the same pilot as user k in cell j. Hence, we can rewrite (2) as

$$\mathbf{y}_{jjk}^{\mathsf{p}} = \sqrt{p_{jk}^{\mathsf{p}}} \tau_{\mathrm{p}} \mathbf{h}_{jk}^j + \tau_{\mathrm{p}} \sum_{(l,i) \in \mathcal{P}_{jk} \setminus (j,k)} \sqrt{p_{li}^{\mathsf{p}}} \mathbf{h}_{li}^j + \mathbf{N}_j^{\mathsf{p}} \boldsymbol{\phi}_{jk}^*. \quad (4)$$

The processed uplink signal, $\mathbf{y}_{jjk}^{\mathsf{p}}$, is a *sufficient statistic* for the estimation of $\mathbf{h}_{jk}^j$. Upon the knowledge of the spatial correlation matrices, BS j can compute the minimum mean-squared error (MMSE) estimate of $\mathbf{h}_{jk}^j$, denoted by $\widehat{\mathbf{h}}_{jk}^j$, based on the observation $\mathbf{y}_{jjk}^{\mathsf{p}}$ as [5]

$$\widehat{\mathbf{h}}_{jk}^j = \sqrt{p_{jk}^{\mathsf{p}}} \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{y}_{jjk}^{\mathsf{p}}, \quad (5)$$

where

$$\boldsymbol{\Psi}_{jk}^j = \left(\sum_{(l,i) \in \mathcal{P}_{jk}} p_{li}^{\mathsf{p}} \tau_{\mathrm{p}} \mathbf{R}_{li}^j + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{M_j} \right)^{-1}. \quad (6)$$

The estimation error is given by $\widetilde{\mathbf{h}}_{jk}^j = \mathbf{h}_{jk}^j - \widehat{\mathbf{h}}_{jk}^j$, and has correlation matrix

$$\mathbf{C}_{jk}^j = \mathsf{E}\left\{ \widetilde{\mathbf{h}}_{jk}^j \left(\widetilde{\mathbf{h}}_{jk}^j \right)^{\mathsf{H}} \right\} = \mathbf{R}_{jk}^j - p_{jk}^{\mathsf{p}} \tau_{\mathrm{p}} \mathbf{R}_{jk}^j \boldsymbol{\Psi}_{jk}^j \mathbf{R}_{jk}^j.$$

It follows that $\widetilde{\mathbf{h}}_{jk}^j$ and $\widehat{\mathbf{h}}_{jk}^j$ are independent random variables distributed as

$$\begin{aligned} \widetilde{\mathbf{h}}_{jk}^j &\sim \mathcal{CN}\left(\mathbf{0}_M, \mathbf{C}_{jk}^j \right), \\ \widehat{\mathbf{h}}_{jk}^j &\sim \mathcal{CN}\left(\mathbf{0}_M, \mathbf{R}_{jk}^j - \mathbf{C}_{jk}^j \right). \end{aligned}$$

C. DOWNLINK TRANSMISSION

In the downlink transmission phase, each BS transmits payload data to all the active users of its cell. Let A_{jk}^t be a coefficient that equals 1 if a URLLC transmission takes place at the t-th slot for URLLC user k in cell j, and 0 otherwise. This coefficient models the random activation pattern of the URLLC users which follows a Bernoulli distribution with parameter a_{u}, $A_{jk}^t \sim \mathcal{B}ern(a_{\mathrm{u}})$. To handle the coexistence of eMBB and URLLC users in the downlink, we consider two transmission techniques: (*i*) puncturing, and (*ii*) superposition coding. Under puncturing, whenever a URLLC transmission is triggered by a BS in a certain slot, all the eMBB transmissions therein are dropped. However, the eMBB service can be still guaranteed in the remaining slots of the frame where no URLLC users are active. Under superposition coding, eMBB transmissions occur in all the slots and each BS linearly combines eMBB and URLLC signals whenever URLLC transmissions are triggered.

The analytical framework detailed next is generalized, namely holds for both the aforementioned transmission techniques upon setting, for an arbitrary BS j and slot t, the coefficient

$$\tilde{A}_j^t = \begin{cases} \left(1 - \sum_{i \in \mathcal{K}_j^{\mathsf{u}}} A_{ji}^t \right)^+, & \text{for puncturing,} \\ 1, & \text{for superposition coding.} \end{cases}$$

Let $\varsigma_{jk}^{\mathsf{e}}[n]$ or $\varsigma_{jk}^{\mathsf{u}}[n]$ be the data symbol transmitted by BS j to user k over an arbitrary channel use n, if k is an eMBB user or a URLLC user, respectively. We assume that $\varsigma_{jk}^{\mathsf{s}}[n] \sim \mathcal{CN}(0, 1)$, with $\mathsf{s} = \{\mathsf{e}, \mathsf{u}\}$. A slot consists of n_{d} channel uses, with $n_{\mathrm{d}} = \lfloor \tau_{\mathrm{d}} / T \rfloor$, and equals the length of the URLLC codeword. The data symbol is precoded by using the M_j-dimensional precoding vector $\mathbf{w}_{jk}$, which is function of the CSI acquired at the BS during the uplink training. It also holds $\mathsf{E}\left\{ \left\| \mathbf{w}_{jk} \right\|^2 \right\} = 1$. The data signal transmitted by BS j over an arbitrary channel use n of slot t is given by

$$\mathbf{x}_j^t[n] = \tilde{A}_j^t \sum_{k \in \mathcal{K}_j^{\mathsf{e}}} \sqrt{\rho_{jk}^{\mathsf{e}}} \mathbf{w}_{jk} \varsigma_{jk}^{\mathsf{e}}[n] + \sum_{i \in \mathcal{K}_j^{\mathsf{u}}} A_{ji}^t \sqrt{\rho_{ji}^{\mathsf{u}}} \mathbf{w}_{ji} \varsigma_{ji}^{\mathsf{u}}[n], \quad (7)$$

with $n = 1, \ldots, n_{\mathrm{d}}$, and where ρ_{jk}^{e} and ρ_{ji}^{u} are the downlink transmit powers used by BS j to its eMBB user k and URLLC user i, respectively, satisfying the following per-BS power constraint

$$\mathsf{E}\left\{\left\|\mathbf{x}_j^t[n]\right\|^2\right\} = \tilde{A}_j^t \sum_{k\in\mathcal{K}_j^{\mathsf{e}}} \rho_{jk}^{\mathsf{e}} + \sum_{i\in\mathcal{K}_j^{\mathsf{u}}} A_{ji}^t \rho_{ji}^{\mathsf{u}} \leq \rho_j^{\max}, \quad (8)$$

with $j = 1, \ldots, L$, and where $\rho_j^{\max}$ is the maximum transmit power at BS j. The data signal received at user k in cell j over an arbitrary channel use n of slot t is denoted as $y_{jk}^{t,\mathsf{s}}[n]$, with $\mathsf{s} = \{\mathsf{e}, \mathsf{u}\}$. In line with the conventional massive MIMO operation, we assume that the users do not acquire the instantaneous downlink CSI, but rather rely on a mean value approximation of their downlink precoded channels. Such approximation is accurate if channel hardening occurs. If user k in cell j is an eMBB user, namely $k \in \mathcal{K}_j^{\mathsf{e}}$, then its received data signal over an arbitrary channel use n of slot t can be written as in (9) at the bottom of the page, where $w_{jk}[n] \sim \mathcal{CN}(0, \sigma_{\mathrm{d}}^2)$ is the i.i.d. receiver noise with variance σ_{d}^2, and we have defined $g_{jk}^{li} = (\mathbf{h}_{jk}^l)^{\mathsf{H}}\mathbf{w}_{li}$, namely the precoded downlink (scalar) channel between the BS in cell l, using the precoding vector intended for its user i, and the k-th user in cell j. If user k in cell j is a URLLC user, its received data signal over an arbitrary channel use n in slot t can be written as in (10) at the bottom of the page. Equation (9) emphasizes the fact that user k in cell j solely knows the statistical CSI of the downlink channel, that is $\mathsf{E}\left\{g_{jk}^{jk}\right\}$. The second term in (9) represents the self-interference due to this lack of instantaneous CSI, referred to as *beamforming gain uncertainty*. Going forward, the intra-cell inter-service interference and intra-cell intra-service interference terms represent the interference caused by the URLLC and eMBB users of cell j, respectively. This is presumably stronger than the inter-cell interference caused by the eMBB users (i.e., intra-service) and the URLLC users (i.e., inter-service) in the other cells. A similar distinction of the various signal contributions is reported in (10) for URLLC user k in cell j. In this case, the lack of instantaneous CSI at the user will be highlighted in the next section.

III. PERFORMANCE ANALYSIS

In this section, we evaluate the downlink performance of eMBB and URLLC users. As per the eMBB users, we consider the spectral efficiency (SE) by applying the infinite-blocklength information-theoretic results established in the *ergodic regime* [4], [5], [30]. An achievable downlink SE, namely a lower-bound on the ergodic downlink capacity, can be obtained by applying the popular *hardening bound* technique [4], [5] on the signal model in (9), by treating all the interference sources as uncorrelated noise. Specifically, an achievable downlink SE of an arbitrary eMBB user k in cell j, is given by

$$\mathsf{SE}_{jk}^{\mathsf{e}} = \frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{c}}} \frac{1}{T} \sum_{t=1}^{T} \log_2\left(1 + \mathsf{SINR}_{jk}^{t,\mathsf{e}}\right), \left[\text{bits/s/Hz}\right], \quad (11)$$

where $\tau_{\mathrm{d}}/\tau_{\mathrm{c}}$ accounts for the estimation overhead,

$$\mathsf{SINR}_{jk}^{t,\mathsf{e}} = \frac{\tilde{A}_j^t \rho_{jk}^{\mathsf{e}} \left|\mathsf{E}\left\{g_{jk}^{jk}\right\}\right|^2}{\sum_{l=1}^{L}\sum_{i=1}^{K_l} \varrho_{li}^t \mathsf{E}\left\{|g_{jk}^{li}|^2\right\} - \tilde{A}_j^t \rho_{jk}^{\mathsf{e}} \left|\mathsf{E}\left\{g_{jk}^{jk}\right\}\right|^2 + \sigma_{\mathrm{d}}^2}, \quad (12)$$

is the effective SINR of user $k \in \mathcal{K}_j^{\mathsf{e}}$, where the expectations are taken with respect to the random channel realizations, and

$$\varrho_{li}^t = \begin{cases} A_{li}^t \rho_{li}^{\mathsf{u}}, & \text{if } i \in \mathcal{K}_l^{\mathsf{u}}, \\ \tilde{A}_l^t \rho_{li}^{\mathsf{e}}, & \text{if } i \in \mathcal{K}_l^{\mathsf{e}}. \end{cases} \quad (13)$$

The expression of the achievable SE shown in (11) holds for any choice of precoding scheme, any channel estimator and any channel distributions. Importantly, it accounts for any choice of coexistence technique between heterogeneous services, namely puncturing or superposition coding. The

$$y_{jk}^{t,\mathsf{e}}[n] = \underbrace{\mathsf{E}\left\{g_{jk}^{jk}\right\}\tilde{A}_j^t\sqrt{\rho_{jk}^{\mathsf{e}}}\varsigma_{jk}^{\mathsf{e}}[n]}_{\text{desired signal}} + \underbrace{\left(g_{jk}^{jk} - \mathsf{E}\left\{g_{jk}^{jk}\right\}\right)\tilde{A}_j^t\sqrt{\rho_{jk}^{\mathsf{e}}}\varsigma_{jk}^{\mathsf{e}}[n]}_{\text{self-interference}} + \underbrace{\sum_{i\in\mathcal{K}_j^{\mathsf{u}}} g_{jk}^{ji} A_{ji}^t \sqrt{\rho_{ji}^{\mathsf{u}}}\varsigma_{ji}^{\mathsf{u}}[n]}_{\text{intra-cell inter-service interference}}$$
$$+ \underbrace{\sum_{i\in\mathcal{K}_j^{\mathsf{e}}\setminus\{k\}} g_{jk}^{ji}\tilde{A}_j^t\sqrt{\rho_{ji}^{\mathsf{e}}}\varsigma_{ji}^{\mathsf{e}}[n]}_{\text{intra-cell intra-service interference}} + \underbrace{\sum_{\substack{l=1\\ l\neq j}}^{L}\sum_{i\in\mathcal{K}_l^{\mathsf{e}}} g_{jk}^{li}\tilde{A}_l^t\sqrt{\rho_{li}^{\mathsf{e}}}\varsigma_{li}^{\mathsf{e}}[n]}_{\text{inter-cell intra-service interference}} + \underbrace{\sum_{\substack{l=1\\ l\neq j}}^{L}\sum_{i\in\mathcal{K}_l^{\mathsf{u}}} g_{jk}^{li} A_{li}^t\sqrt{\rho_{li}^{\mathsf{u}}}\varsigma_{li}^{\mathsf{u}}[n]}_{\text{inter-cell inter-service interference}} + \underbrace{w_{jk}[n]}_{\text{noise}} \quad (9)$$

$$y_{jk}^{t,\mathsf{u}}[n] = \underbrace{g_{jk}^{jk}A_{jk}^t\sqrt{\rho_{jk}^{\mathsf{u}}}\varsigma_{jk}^{\mathsf{u}}[n]}_{\text{desired signal}} + \underbrace{\sum_{i\in\mathcal{K}_j^{\mathsf{u}}\setminus\{k\}} g_{jk}^{ji}A_{ji}^t\sqrt{\rho_{ji}^{\mathsf{u}}}\varsigma_{ji}^{\mathsf{u}}[n]}_{\text{intra-cell intra-service interference}} + \underbrace{\sum_{\substack{l=1\\ l\neq j}}^{L}\sum_{i\in\mathcal{K}_l^{\mathsf{u}}} g_{jk}^{li}A_{li}^t\sqrt{\rho_{li}^{\mathsf{u}}}\varsigma_{li}^{\mathsf{u}}[n]}_{\text{inter-cell intra-service interference}} + \underbrace{\sum_{l=1}^{L}\sum_{i\in\mathcal{K}_l^{\mathsf{e}}} g_{jk}^{li}\tilde{A}_l^t\sqrt{\rho_{li}^{\mathsf{e}}}\varsigma_{li}^{\mathsf{e}}[n]}_{\text{inter-service interference}} + \underbrace{w_{jk}[n]}_{\text{noise}} \quad (10)$$

infinite-blocklength analysis above is established upon the assumption of block-fading channel model, entailing that each eMBB codeword has infinite length that spans a large number of independent fading realizations. This assumption cannot be applied to the URLLC case. As per the URLLC user, we consider a nonasymptotic analysis of the downlink error probability on a slot basis by applying the finite-blocklength information-theoretic results established in [8]. Firstly, we rewrite (10) as

$$y_{jk}^{t,\mathsf{u}}[n] = g_{jk}^{jk} q_{jk}[n] + z_{jk}[n], \quad n = 1, \ldots, n_{\mathrm{d}}, \tag{14}$$

where $q_{jk}[n] = A_{jk}^{t}\sqrt{\rho_{jk}^{\mathsf{u}}}\varsigma_{jk}^{\mathsf{u}}[n]$, and

$$\begin{aligned} z_{jk}[n] = &\sum_{i\in\mathcal{K}_j^{\mathsf{u}}\setminus\{k\}} g_{jk}^{ji} q_{ji}[n] + \sum_{i\in\mathcal{K}_j^{\mathsf{e}}} g_{jk}^{ji}\tilde{A}_j^{t}\sqrt{\rho_{ji}^{\mathsf{e}}}\varsigma_{ji}^{\mathsf{e}}[n] \\ &+ \sum_{\substack{l=1\\ l\neq j}}^{L}\left(\sum_{i\in\mathcal{K}_l^{\mathsf{u}}} g_{jk}^{li} q_{li}[n] + \sum_{i\in\mathcal{K}_l^{\mathsf{e}}} g_{jk}^{li}\tilde{A}_l^{t}\sqrt{\rho_{li}^{\mathsf{e}}}\varsigma_{li}^{\mathsf{e}}[n]\right) \\ &+ w_{jk}[n]. \end{aligned} \tag{15}$$

However, URLLC user k in cell j has not access to g_{jk}^{jk}, but performs data decoding by only leveraging its mean value, $\widehat{g}_{jk}^{jk} = \mathsf{E}\left\{(\mathbf{h}_{jk}^{j})^{\mathsf{H}}\mathbf{w}_{jk}\right\}$, which is treated as perfect. This estimate is accurate if channel hardening holds. Notice that, the precoded channel g_{jk}^{jk} is frequency-flat and time-invariant over the transmission of the n_{d}-length URLLC codeword in slot t. Moreover, g_{jk}^{jk} remains constant for any other transmission from BS j to user k over slots in the same TDD frame. Given all channels and precoding vectors, the effective noise terms $\{z_{jk}[n] \in \mathbb{C}; n = 1, \ldots, n_{\mathrm{d}}\}$ are random variables conditionally i.i.d. with variance σ_{jk}^2, i.e., $\mathcal{CN}\left(0, \sigma_{jk}^2\right)$, given by

$$\begin{aligned} \sigma_{jk}^2 = &\sum_{i\in\mathcal{K}_j^{\mathsf{u}}\setminus\{k\}} A_{ji}^{t}\rho_{ji}^{\mathsf{u}}|g_{jk}^{ji}|^2 + \sum_{i\in\mathcal{K}_j^{\mathsf{e}}}\tilde{A}_j^{t}\rho_{ji}^{\mathsf{e}}|g_{jk}^{ji}|^2 \\ &+ \sum_{\substack{l=1\\ l\neq j}}^{L}\left(\sum_{i\in\mathcal{K}_l^{\mathsf{u}}} A_{li}^{t}\rho_{li}^{\mathsf{u}}|g_{jk}^{li}|^2 + \sum_{i\in\mathcal{K}_l^{\mathsf{e}}}\tilde{A}_l^{t}\rho_{li}^{\mathsf{e}}|g_{jk}^{li}|^2\right) + \sigma_{\mathrm{d}}^2. \end{aligned} \tag{16}$$

To determine the transmitted codeword

$$\mathbf{q}_{jk} = \left[q_{jk}[1], \ldots, q_{jk}[n_{\mathrm{d}}]\right]^{\mathsf{T}},$$

user k in cell j employs a *mismatched scaled nearest-neighbor* (SNN) decoder [31], with which selects the codeword $\widetilde{\mathbf{q}}_{jk}$ from the codebook $\mathcal{C}$ by applying the rule

$$\widehat{\mathbf{q}}_{jk} = \arg\min_{\widetilde{\mathbf{q}}_{jk}\in\mathcal{C}} \left\|\mathbf{y}_{jk}^{t,\mathsf{u}} - \widehat{g}_{jk}^{jk}\widetilde{\mathbf{q}}_{jk}\right\|^2, \tag{17}$$

where $\mathbf{y}_{jk}^{t,\mathsf{u}} = \left[y_{jk}^{t,\mathsf{u}}[1], \ldots, y_{jk}^{t,\mathsf{u}}[n_{\mathrm{d}}]\right]^{\mathsf{T}} \in \mathbb{C}^{n_{\mathrm{d}}}$ is the received data vector.

Let $\epsilon_{jk}^{\mathsf{dl}} = \mathsf{Pr}\{\widehat{\mathbf{q}}_{jk} \neq \mathbf{q}_{jk}\}$ be the downlink error probability experienced by the URLLC user k in cell j achieved by the SNN decoding. An upper bound on $\epsilon_{jk}^{\mathsf{dl}}$ is obtained by using the standard *random-coding* approach [32],

$$\epsilon_{jk}^{\mathsf{dl}} \leq \mathsf{E}_{g_{jk}^{jk}}\left\{\mathsf{Pr}\left\{\sum_{n=1}^{n_{\mathrm{d}}} \iota_s\left(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n]\right) \leq \log\frac{m-1}{r}\middle| g_{jk}^{jk}\right\}\right\}, \tag{18}$$

where $m = 2^b$ is the number of codewords with length n_{d} that convey b information bits, r is a random variable uniformly distributed in the interval $[0, 1]$ and $\iota_s(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n])$ is the *generalized information density*, given by

$$\begin{aligned} &\iota_s\left(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n]\right) \\ &\quad = -s\left|\mathbf{y}_{jk}^{t,\mathsf{u}}[n] - \widehat{g}_{jk}^{jk}q_{jk}[n]\right|^2 + \frac{s|\mathbf{y}_{jk}^{t,\mathsf{u}}[n]|^2}{1 + s\rho_{jk}^{\mathsf{u}}|\widehat{g}_{jk}^{jk}|^2} \\ &\qquad + \log\left(1 + s\rho_{jk}^{\mathsf{u}}|\widehat{g}_{jk}^{jk}|^2\right), \end{aligned} \tag{19}$$

for all $s > 0$. In (18) the expectation is taken over the distribution of g_{jk}^{jk}, and the probability is computed with respect to the downlink data symbol $\{q_{jk}[n]\}_{n=1}^{n_{\mathrm{d}}}$, the effective additive noise $\{z_{jk}[n]\}_{n=1}^{n_{\mathrm{d}}}$, and the random variable r. The evaluation of the upper bound in (18) entails a very demanding numerical computation to firstly obtain the probability, and then to numerically tighten the upper bound value to the low error probability target of the URLLC use case by optimizing with respect to s.

Luckily, we can reliably approximate the right-hand side of (18) in closed form, hence with a significant relief of the computational burden, by using the *saddlepoint* approximation provided in [8, Th. 2]. The existence of a saddlepoint approximation is guaranteed by the fact that the third derivative of the *moment-generating* function of $-\iota_s(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n])$ exists in a neighborhood of zero delimited by the values $\underline{\varepsilon} < 0 < \overline{\varepsilon}$ given by [8, Appendix B]

$$\underline{\varepsilon} = -\frac{\sqrt{(\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}})^2 + 4\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)} + \zeta_{\mathsf{a}} - \zeta_{\mathsf{b}}}{2\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)}, \tag{20}$$

$$\overline{\varepsilon} = \frac{\sqrt{(\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}})^2 + 4\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)} - \zeta_{\mathsf{a}} + \zeta_{\mathsf{b}}}{2\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)}, \tag{21}$$

where

$$\zeta_{\mathsf{a}} = s\left(\rho_{jk}^{\mathsf{u}}|g_{jk}^{jk} - \widehat{g}_{jk}^{jk}|^2 + \sigma^2\right), \tag{22}$$

$$\zeta_{\mathsf{b}} = \frac{s}{1 + s\rho_{jk}^{\mathsf{u}}|\widehat{g}_{jk}^{jk}|^2}\left(\rho_{jk}^{\mathsf{u}}|g_{jk}^{jk}|^2 + \sigma^2\right), \tag{23}$$

$$\mu = \frac{s^2\left|\rho_{jk}^{\mathsf{u}}\left|g_{jk}^{jk}\right|^2 + \sigma^2 - \left(g_{jk}^{jk}\right)^{*}\widehat{g}_{jk}^{jk}\rho_{jk}^{\mathsf{u}}\right|^2}{\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}\left(1 + s\rho_{jk}^{\mathsf{u}}|\widehat{g}_{jk}^{jk}|^2\right)}. \tag{24}$$

The saddlepoint approximation hinges on the *cumulant-generating* function of $-\iota_s(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n])$ given by

$$\upsilon(\varepsilon) = \log\mathsf{E}\left\{e^{-\varepsilon\iota_s\left(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n]\right)}\right\}, \tag{25}$$

on its first derivative $\upsilon'(\zeta)$, and second derivative $\upsilon''(\zeta)$, for all $\varepsilon \in (\underline{\varepsilon}, \overline{\varepsilon})$

$$\upsilon(\varepsilon) = -\varepsilon \log\left(1 + s\rho_{jk}^{\mathsf{u}} |\widehat{g}_{jk}^{jk}|^2\right) - \log\left(1 + (\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}})\varepsilon - \zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)\varepsilon^2\right) \quad (26)$$

$$\upsilon'(\varepsilon) = -\log\left(1 + s\rho_{jk}^{\mathsf{u}} |\widehat{g}_{jk}^{jk}|^2\right) - \frac{(\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}}) - 2\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)\varepsilon}{1 + (\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}})\varepsilon - \zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)\varepsilon^2} \quad (27)$$

$$\upsilon''(\varepsilon) = \left[\frac{(\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}}) - 2\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)\varepsilon}{1 + (\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}})\varepsilon - \zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)\varepsilon^2}\right]^2 + \frac{2\zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)}{1 + (\zeta_{\mathsf{b}} - \zeta_{\mathsf{a}})\varepsilon - \zeta_{\mathsf{a}}\zeta_{\mathsf{b}}(1-\mu)\varepsilon^2}. \quad (28)$$

Let $m = e^{n_{\mathsf{d}}R}$ for some strictly positive transmission rate $R = (\log m)/n_{\mathsf{d}}$, and let $\varepsilon \in (\underline{\varepsilon}, \overline{\varepsilon})$ be the solution to the equation $R = -\upsilon'(\varepsilon)$. Let I_{s} be the *generalized mutual information* [31] defined as $I_{\mathsf{s}} = \mathsf{E}\{\imath_s(q_{jk}[1], v_{jk}[1])\} = -\upsilon'(0)$. Lastly, consider the *critical rate* [32, eq. (5.6.30)] given by $R_{\mathsf{s}}^{\mathsf{cr}} = -\upsilon'(1)$. Then, we have three possible saddlepoint approximations for the error probability upper bound [8].

If $\varepsilon \in [0, 1]$, then $R_{\mathsf{s}}^{\mathsf{cr}} \leq R \leq I_{\mathsf{s}}$ and

$$\mathsf{Pr}\left\{\sum_{n=1}^{n_{\mathsf{d}}} \imath_s\left(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n]\right) \leq \log \frac{e^{n_{\mathsf{d}}R} - 1}{r}\right\} \approx e^{n_{\mathsf{d}}[\upsilon(\varepsilon) + \varepsilon R]}\left[\Psi_{n_{\mathsf{d}},\varepsilon}(\varepsilon) + \Psi_{n_{\mathsf{d}},\varepsilon}(1-\varepsilon)\right], \quad (29)$$

where

$$\Psi_{n_{\mathsf{d}},\varepsilon}(\ell) \triangleq e^{\frac{1}{2}n_{\mathsf{d}}\ell^2\upsilon''(\varepsilon)} Q\left(\ell\sqrt{n_{\mathsf{d}}\upsilon''(\varepsilon)}\right). \quad (30)$$

If $\varepsilon > 1$, then $R < R_{\mathsf{s}}^{\mathsf{cr}}$ and

$$\mathsf{Pr}\left\{\sum_{n=1}^{n_{\mathsf{d}}} \imath_s\left(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n]\right) \leq \log \frac{e^{n_{\mathsf{d}}R} - 1}{r}\right\} \approx e^{n_{\mathsf{d}}[\upsilon(1) + R]}\left[\widetilde{\Psi}_{n_{\mathsf{d}}}(1, 1) + \widetilde{\Psi}_{n_{\mathsf{d}}}(0, -1)\right], \quad (31)$$

where

$$\widetilde{\Psi}_{n_{\mathsf{d}}}(\ell_1, \ell_2) \triangleq e^{n_{\mathsf{d}}\ell_1\left[R_{\mathsf{s}}^{\mathsf{cr}} - R + \frac{1}{2}\upsilon''(1)\right]} \times Q\left(\ell_1\sqrt{n_{\mathsf{d}}\upsilon''(1)} + \ell_2 \frac{n_{\mathsf{d}}\left(R_{\mathsf{s}}^{\mathsf{cr}} - R\right)}{\sqrt{n_{\mathsf{d}}\upsilon''(1)}}\right). \quad (32)$$

If $\varepsilon < 0$, then $R > I_{\mathsf{s}}$ and

$$\mathsf{Pr}\left\{\sum_{n=1}^{n_{\mathsf{d}}} \imath_s\left(q_{jk}[n], \mathbf{y}_{jk}^{t,\mathsf{u}}[n]\right) \leq \log \frac{e^{n_{\mathsf{d}}R} - 1}{r}\right\} \approx 1 - e^{n_{\mathsf{d}}[\upsilon(\varepsilon) + \varepsilon R]}\left[\Psi_{n_{\mathsf{d}},\varepsilon}(-\varepsilon) - \Psi_{n_{\mathsf{d}},\varepsilon}(1-\varepsilon)\right]. \quad (33)$$

The saddlepoint approximation is more accurate in the URLLC massive MIMO regime than the conventionally-used *normal approximation* [24] as the former characterizes the exponential decay of the error probability, i.e., the error-exponent, as a function of the URLLC codeword length, and the transmission rate requirement R, while uses the *Berry-Esseen central-limit theorem* (used in the normal approximation) to only characterize the multiplicative factor following the error-exponent term. The normal approximation, whose formulation directly involves the generalized mutual information, I_{s}, but does not R, is accurate only when I_{s} is close to R. This operating regime does not hold for URLLC wherein R is typically lower than I_{s} to accomplish the very low error probability targets. Once that the approximate upper bounds on the downlink error probability are obtained via saddlepoint approximation, we compute the *downlink network availability* [8], η^{dl}, as

$$\eta^{\mathsf{dl}} = \mathsf{Pr}\left\{\epsilon_{jk}^{\mathsf{dl}} \leq \epsilon_{\mathrm{target}}^{\mathsf{dl}}\right\} \quad (34)$$

which measures the probability that the target error probability $\epsilon_{\mathrm{target}}^{\mathsf{dl}}$ is satisfied by an arbitrary user k in cell j, in presence of interfering users. While the expectation in the error probability definition is taken with respect to the small-scale fading and the effective additive noise, given a large-scale fading realization, the probability in the network availability definition is computed with respect to the large-scale fading (i.e., path loss, shadowing etc.). The expression of the network availability shown in (34) holds for any choice of precoding scheme, any channel estimator and any channel distributions. Importantly, it accounts for any choice of coexistence technique between heterogeneous services, namely puncturing or superposition coding.

IV. PRECODING AND POWER CONTROL

The choice of the precoding scheme and of the downlink power allocation deeply affects the SE of the eMBB users and the network availability for the URLLC users. A popular heuristic precoding design results from the *uplink-downlink* duality principle [5, Sec. 4.3.2], which consists in selecting the precoding vector as a function of its dual combining vector. Let the precoding vector for user k in cell j be

$$\mathbf{w}_{jk} = \frac{\mathbf{v}_{jk}}{\|\mathbf{v}_{jk}\|}, \quad (35)$$

where the denominator serves to make the average power of the precoding vector unitary, and $\mathbf{v}_{jk}$ is the dual combining vector. Importantly, we realistically assume that each BS designs the combining vectors (and in turn the precoding vectors) on a frame basis rather than on a slot basis, and they stay constant over the TDD radio frame. Hence, the combining (precoding) scheme design is agnostic with respect to the random activation pattern of the URLLC users and the coexistence strategy (i.e., either SPC or PUNC). Assuming that the BSs are unaware of the random activation pattern of the URLLC users in the uplink, then an achievable uplink SE of eMBB user k in cell j is given by

$$\mathsf{SE}_{jk}^{\mathsf{ul}} = \frac{\tau_{\mathsf{u}}}{\tau_{\mathsf{c}}} \mathsf{E}\left\{\log_2\left(1 + \mathsf{SINR}_{jk}^{\mathsf{ul}}\right)\right\}, \quad (36)$$

where the uplink effective SINR is given by

$$\mathsf{SINR}_{jk}^{\mathsf{ul}} = \frac{p_{jk}\left|\mathbf{v}_{jk}^{\mathsf{H}}\widehat{\mathbf{h}}_{jk}^{j}\right|^2}{\sum\limits_{l=1}^{L}\sum\limits_{\substack{i=1\\(l,i)\neq(j,k)}}^{K_l} p_{li}\left|\mathbf{v}_{jk}^{\mathsf{H}}\widehat{\mathbf{h}}_{li}^{j}\right|^2 + \mathbf{v}_{jk}^{\mathsf{H}}\left(\boldsymbol{\Upsilon}^{j}+\sigma_{\mathsf{u}}^2\mathbf{I}_{M_j}\right)\mathbf{v}_{jk}},$$

and $\boldsymbol{\Upsilon}^{j} = \sum_{l=1}^{L}\sum_{i=1}^{K_l} p_{li}\mathbf{C}_{li}^{j}$. Equation (36) reflects the worst case in which all the URLLC users are active in all the slots, hence it is a pessimistic achievable uplink SE. For the sake of comparison, we herein consider three precoding schemes, obtained from the following combining vectors via uplink-downlink duality.

Multi-cell MMSE (M-MMSE):

$$\mathbf{v}_{jk}^{\mathsf{M-MMSE}} = \left[\left(\sum_{l=1}^{L}\widehat{\mathbf{H}}_{l}^{j}\mathbf{P}_{l}\left(\widehat{\mathbf{H}}_{l}^{j}\right)^{\mathsf{H}} + \boldsymbol{\Upsilon}^{j} + \sigma_{\mathsf{u}}^2\mathbf{I}_{M_j}\right)^{-1}\widehat{\mathbf{H}}_{j}^{j}\mathbf{P}_{j}\right]_{:,k}$$

where $\mathbf{P}_l = \mathrm{diag}(p_{li},\ldots,p_{lK_l}) \in \mathbb{R}^{K_l\times K_l}$ is the matrix with the uplink transmit powers of all the users in cell l as diagonal elements and $\widehat{\mathbf{H}}_{l}^{j} = [\widehat{\mathbf{h}}_{l1}^{j}\ldots\widehat{\mathbf{h}}_{lK_l}^{j}]$. The M-MMSE combining vector, $\mathbf{v}_{jk}^{\mathsf{M-MMSE}}$, maximizes the achievable SE in (36) [5, Appendix C.3.2], while its corresponding M-MMSE precoding vector provides a *suboptimal* value of the achievable downlink SE in (11). Importantly, M-MMSE requires each BS to acquire CSI and statistical CSI of all the users of the multi-cell system. Moreover, the computation of the precoding vector, which entails inverting a matrix $M_j \times M_j$, may be demanding for large BS arrays. Although impractical, M-MMSE precoding will serve as benchmark.

Regularized zero-forcing (RZF):

$$\mathbf{v}_{jk}^{\mathsf{RZF}} = \left[\widehat{\mathbf{H}}_{j}^{j}\left(\left(\widehat{\mathbf{H}}_{j}^{j}\right)^{\mathsf{H}}\widehat{\mathbf{H}}_{j}^{j} + \sigma_{\mathsf{u}}^2\mathbf{P}_{j}^{-1}\right)^{-1}\right]_{:,k}.$$

Compared to M-MMSE, RZF precoding requires each BS to estimate the channels of only its users. Moreover, computing the RZF precoding vector is computationally cheaper since the size of the matrix to be inverted is $K_j \times K_j$. However, RZF does only suppress the intra-cell interference while, unlike M-MMSE, does not provide to the users any protection mechanism against inter-cell interference and channel estimation error.

Maximum Ratio (MR): $\mathbf{v}_{jk}^{\mathsf{MR}} = \widehat{\mathbf{h}}_{jk}^{j}$. It is computationally the cheapest but performance-wise the worst precoding scheme. MR only aims at maximizing the power of the desired signal, providing no interference-suppression mechanism. MR will serve as lower bound on the performance.

Properly allocating the downlink power can make all the difference to meet the strict reliability requirements of the URLLC and to improve the SE of the eMBB users. Next, we provide three power allocation schemes that take into account the power budget at the BSs, the adopted eMBB-URLLC coexistence strategy and the URLLC activation pattern, which is known at the BS in the downlink operation.

Equal power allocation (EPA): It consists in setting

$$\rho_{ji}^{\mathsf{u}} = \rho_{j}^{\max}\frac{A_{ji}^{t}}{\tilde{A}_{j}^{t}K_{j}^{\mathsf{e}} + \sum\limits_{k\in\mathcal{K}_{j}^{\mathsf{u}}}A_{jk}^{t}},\ i\in\mathcal{K}_{j}^{\mathsf{u}} \tag{37}$$

$$\rho_{jk}^{\mathsf{e}} = \rho_{j}^{\max}\frac{\tilde{A}_{j}^{t}}{\tilde{A}_{j}^{t}K_{j}^{\mathsf{e}} + \sum\limits_{i\in\mathcal{K}_{j}^{\mathsf{u}}}A_{ji}^{t}},\ k\in\mathcal{K}_{j}^{\mathsf{e}} \tag{38}$$

to satisfy the per-BS power constraint in (8) with equality and allocate the same share of power to each user, regardless of its channel conditions and its service requirements.

Weighted fractional power allocation (FPA): it consists in setting the powers as

$$\rho_{ji}^{\mathsf{u}} = \frac{\omega\rho_{j}^{\max}A_{ji}^{t}\left(\beta_{ji}^{j}\right)^{\nu}}{(1-\omega)\tilde{A}_{j}^{t}\sum\limits_{k\in\mathcal{K}_{j}^{\mathsf{e}}}\left(\beta_{jk}^{j}\right)^{\nu} + \omega\sum\limits_{u\in\mathcal{K}_{j}^{\mathsf{u}}}A_{ju}^{t}\left(\beta_{ju}^{j}\right)^{\nu}},\ i\in\mathcal{K}_{j}^{\mathsf{u}} \tag{39}$$

$$\rho_{jk}^{\mathsf{e}} = \frac{(1-\omega)\rho_{j}^{\max}\tilde{A}_{j}^{t}\left(\beta_{jk}^{j}\right)^{\nu}}{(1-\omega)\tilde{A}_{j}^{t}\sum\limits_{e\in\mathcal{K}_{j}^{\mathsf{e}}}\left(\beta_{je}^{j}\right)^{\nu} + \omega\sum\limits_{i\in\mathcal{K}_{j}^{\mathsf{u}}}A_{ji}^{t}\left(\beta_{ji}^{j}\right)^{\nu}},\ k\in\mathcal{K}_{j}^{\mathsf{e}} \tag{40}$$

where the weight $\omega \in (0,1)$ adjusts the amount of downlink power to be allocated to the URLLC users, while ν establishes the power control policy as a function of the average channel gain. An opportunistic power allocation is attained by setting $\nu > 0$, with which more power is allocated to the users with better channel conditions. Conversely, fairness is supported by setting $\nu < 0$, with which more power is allocated to the users with worse channel conditions. If $\omega \in (0.5,1)$ a larger share of power is allocated to the URLLC users rather than to the eMBB users, whereas it is the other way around if $\omega \in (0,0.5)$. Notice that, if $\nu = 0$ and $\omega = 0.5$, then the FPA reduces to the EPA.

Optimal power allocation (OPA) for max product SINR: The powers are the solution of the optimization problem

$$\underset{\left\{\rho_{jk}^{\mathsf{s}}\right\}}{\text{maximize}} \quad \prod_{j=1}^{L}\prod_{k=1}^{K_j}\mathsf{SINR}_{jk}^{t,\mathsf{s}} \tag{41a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K_j}\varrho_{jk}^{t} \le \rho_{j}^{\max},\ \forall j, \tag{41b}$$

where the superscript $\mathsf{s} = \mathsf{e}$ if user $k \in \mathcal{K}_j^{\mathsf{e}}$, $\mathsf{s} = \mathsf{u}$ otherwise, and ϱ_{jk}^{t} is given in (13). Without further entangling the notation in (41), we remark that the SINR of inactive users is fictitiously set to 1 to preserve the optimization problem formulation. This power allocation strategy treats all the users as eMBB users, hence it would be optimal if there would be no URLLC users active in a given slot, by maximizing a lower bound on the sum SE of the multi-cell system. Although the SINR expression in (12) is meaningless when applied to a

URLLC user, we can still heuristically plug the URLLC powers resulting from (41) into the error probability analysis and motivate this approach by looking at the performance. All the considered power allocation schemes, in principle, run on a slot-basis in order to adapt the power coefficients to the URLLC activation pattern. Fortunately, these schemes only rely on the knowledge of the statistical CSI which allows to pre-compute some power coefficients or to keep the power allocation for multiple slots/frames in case of no macroscopic changes in the propagation environment. Unlike the EPA and the FPA schemes, the OPA scheme requires a certain degree of cooperation among the BSs which must send statistical CSI to let a central processing unit (e.g., a *master* BS) compute the SINR of all the users and solve the optimization problem, and feed them back with the power coefficients to use. This would introduce intolerable delay for the URLLC users. Moreover, solving problem (41), although efficiently as a geometric program [5, Th. 7.2], is unlikely to be doable within a time-slot, especially for crowded networks. Hence, the OPA scheme is of limited practical use, but will serve for benchmarking purposes.

V. SIMULATION RESULTS

In this section, we present and discuss the results of our simulations in which the coexistence of eMBB and URLLC is deeply analyzed under different setups. Specifically, we shed light on the impact of different factors on the performance, such as the transmission technique and the precoding scheme, the power control strategy, the imperfect CSI and estimation overhead, the pilot contamination, the length and number of slots in a TDD frame, and the characteristics of the URLLC activation pattern.

Our simulation scenario consists of a multi-cell massive MIMO system with $L = 4$ cells. Each cell covers a nominal area of 500×500 squared meters, and is served by a BS, placed at the cell center, equipped with a uniform linear array (ULA) with equispaced half-wavelength antenna elements. Without loss of generality, we assume an equal number $M = 100$ of antennas at each BS, and an equal number K of users in each cell. A wrap-around topology is implemented as in [5, Sec. 4.1.3]. The users are dropped uniformly at random over the coverage area but at a minimum distance of 25 m from the BS. In addition, we assume that the URLLC users are distributed uniformly at random in an area of 125×125 squared meters that surrounds the BS. A random realization of the user locations determines a set of large-scale fading coefficients and constitutes a snapshot of the network. For a given network snapshot the achievable downlink SEs of the active eMBB users are computed according to (11), while the downlink error probabilities of the URLLC users are obtained according to the approximations (29)-(33). The cumulative distribution function (CDF) of the SE and the network availability are then drawn over many network snapshots. The channel correlation matrices are generated according to the popular *local scattering* spatial correlation model [5, Sec. 2.6], and we assume that the scattering is only localized around the users and uniformly distributed at random with delay spread 25° degrees [8]. The average channel gain is obtained according to the non-line-of-sight macro cell 3GPP model for 2 GHz carriers [33], and given in dB by

$$\beta_{lk}^{j} = -35.3 - 37.6 \log_{10}\left(\frac{d_{lk}^{j}}{1 \text{ m}}\right) + F_{lk}^{j}$$

for an arbitrary user k in cell l placed at a distance d_{lk}^{j} from BS j, and where $F_{lk}^{j} \sim \mathcal{N}(0, \sigma_{\mathsf{sh}}^2)$ models the log-normal shadowing as an i.i.d. random variable with standard deviation $\sigma_{\mathsf{sh}} = 4$ dB. The transmission bandwidth is 20 MHz, and the receiver noise power equals −94 dBm both for the uplink and the downlink. Moreover, we let $\rho_j^{\max} = 46$ dBm, $j = 1, \ldots, L$, and the uplink transmit power, both for pilot and payload data, be 23 dBm for all the users. We assume that the URLLC packet consists of $b = 160$ bits, yielding a transmission rate $R = b/n_{\mathrm{d}}$. Lastly, without loss of generality, we set $\tau_{\mathrm{u}} = 0$ as we only focus on the downlink performance. Unless otherwise stated, we consider TDD frames with length $\tau_{\mathrm{c}} = 580$ channel uses, given by $T_{\mathrm{c}} = 2$ ms and $B_{\mathrm{c}} = 290$ kHz, which supports user mobility up to 67.50 km/h. The 3GPP specifications in [34] reports reliability and latency requirements associated to specific URLLC packet sizes and service areas[2] for different URLLC use cases. With these specific simulation settings, we mainly targeted URLLC applications such as motion control for factory automation and closed-loop control for process automation, which require a minimum target error probability of 10^{-5} and a minimum latency of 2 ms, and are characterized by service areas with typical size 50 m×10 m, and 100 m×100 m [34], respectively.

In the first set of simulations we consider the following setup: $K = 20$, $\alpha = 0.2$, $a_{\mathrm{u}} = 10^{-0.5}$, $\tau_{\mathrm{p}} = 80$ (no pilot contamination), $T = 5$ slots of length $n_{\mathrm{d}} = 100$ channel uses. In Fig. 2 we plot the CDFs of the achievable downlink SE per "active" eMBB user obtained for different precoding and power allocation strategies, both for superposition coding (top subfigure) and puncturing technique (bottom subfigure). Under these assumptions, SPC is greatly superior than PUNC, precoding and power allocation strategies being equal. M-MMSE with OPA gives, as expected, the best SE but EPA performs almost equally well, regardless of the precoding scheme. RZF provides a practical excellent trade-off between M-MMSE and MR. These results suggest that we are approximately operating in an interference-free scenario, thanks to the full and partial interference-suppression mechanism provided by M-MMSE and RZF, respectively. As per the FPA strategy, in these simulations we have selected $\nu = 0.5$ to promote an opportunistic

2. The service area is defined as the geographic region where a 3GPP communication service is accessible. Typically, the stricter the requirements are, the smaller the service area is. In general, eMBB users have larger service areas than URLLC users due to their looser reliability and latency requirements. This justifies our assumption to drop the URLLC users in a smaller area around the BS.

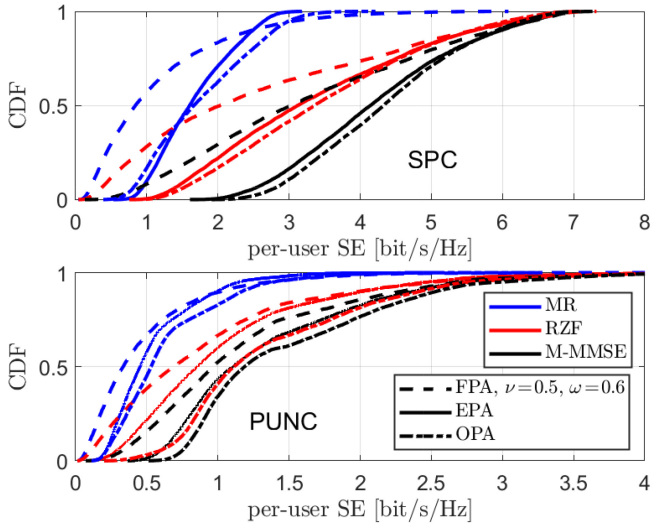

FIGURE 2. CDFs of the achievable downlink SE per active eMBB user, for different transmission, precoding and power allocation strategies. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

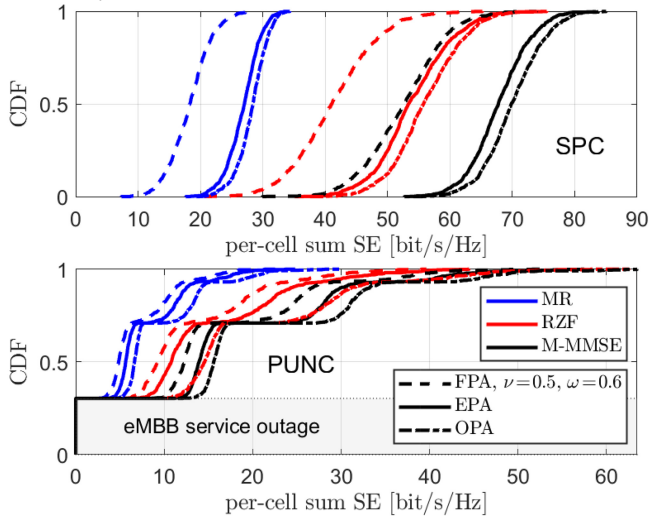

FIGURE 3. CDFs of the achievable downlink sum SE per cell, for different transmission, precoding and power allocation strategies. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

power allocation and $\omega = 0.6$ to prioritize the URLLC users. Such a choice does not favor the eMBB users and justify the worst performance of FPA among the considered strategies when SPC is applied.

Same conclusions hold for the results shown in Fig. 3 where the CDFs of the corresponding sum SE per cell are illustrated. In these figures, we mainly emphasize the eMBB *service outage* likely occurring when PUNC is adopted. We define the eMBB service outage, under PUNC operation, as

$$\varsigma_{\text{out}} = \Pr\left\{ \sum_{k \in \mathcal{K}_j^{\text{e}}} \text{SE}_{jk}^{\text{e}} = 0 \right\}, \quad j = 1, \ldots, L,$$

where the probability is computed with respect to the large-scale fading. This probability for a BS to provide no service in a TDD frame to its eMBB users depends on the activation

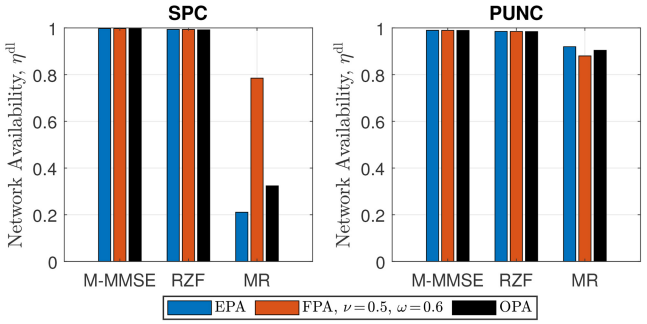

FIGURE 4. Network availability for different transmission, precoding and power allocation strategies. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

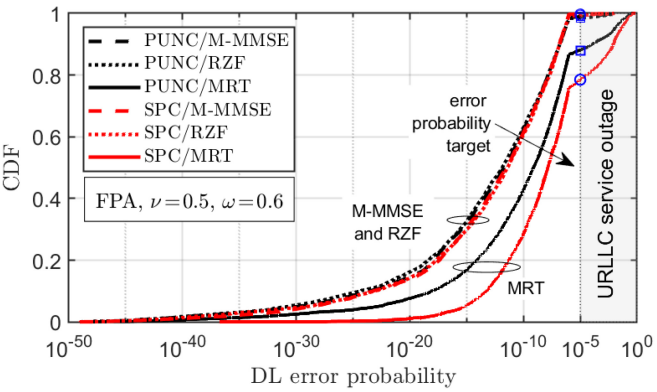

FIGURE 5. Downlink per-user error probability for different transmission and precoding strategies. Settings: EPA, $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

pattern of the URLLC users and the number of slots per frame. We will discuss this aspect in detail later. Under the settings considered in Fig. 3, the eMBB service outage is quite significant as amounts to about 30%.

In Fig. 4 we move to the URLLC performance by showing the downlink network availability achieved when $\epsilon_{\text{target}}^{\text{dl}} = 10^{-5}$. Despite the interference caused by the eMBB users when SPC is performed, both M-MMSE and RZF are able to provide levels of network availability close to one, in line with PUNC, revealing a great ability of suppressing the interference and supporting high reliability. Conversely, MR provides poor performance in SPC when EPA or OPA (which is optimal for the eMBB users) schemes are used. Notice that, our choice for the parameters of the FPA scheme pays off for the combination SPC/MR. The network availability values shown in Fig. 4 are obtained by the error probabilities whose CDFs are illustrated in Fig. 5. To better understand its meaning, the network availability is given by the cross-point between the CDF of the per-user error probability and the vertical line representing the error probability target value, as Fig. 5 highlights (blue circle markers). From this set of simulations, we conclude that SPC is clearly superior to PUNC in terms of SE yet providing very high network availability, when M-MMSE or RZF are carried out. If MR is the only viable option (for instance due to strict complexity or hardware constraints), then SPC with

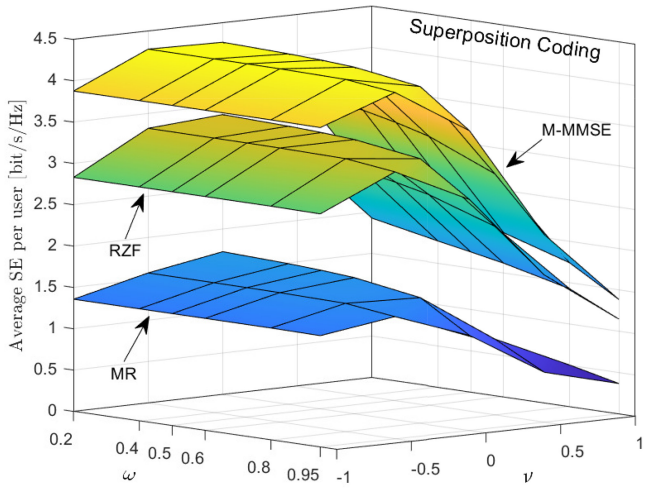

FIGURE 6. Average per-user SE achieved by SPC with FPA, for different precoding schemes and values of ν, ω. The average is taken over 200 network snapshots. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

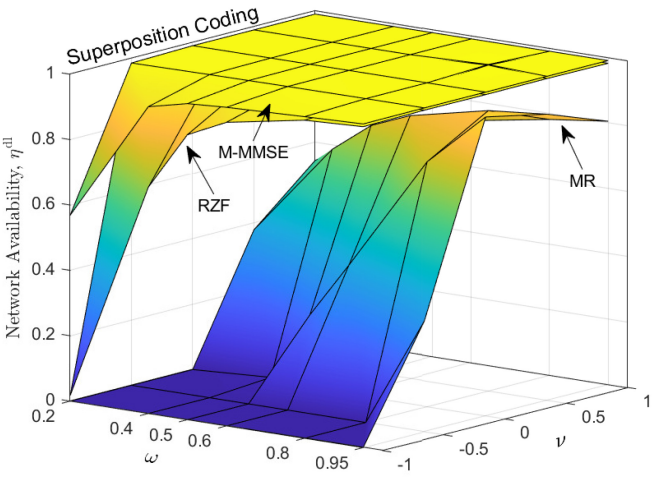

FIGURE 7. Network availability achieved by SPC with FPA, for different precoding schemes and values of ν, ω. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

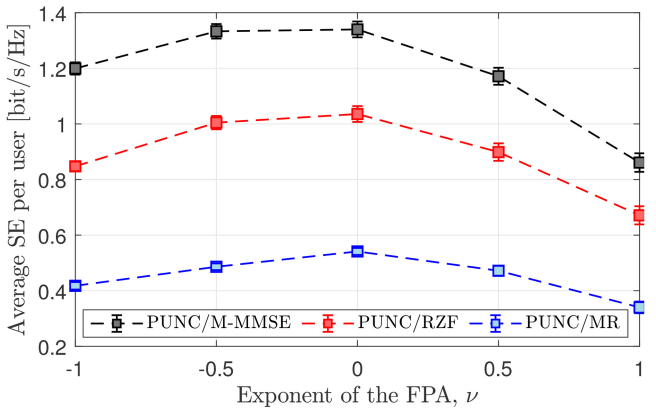

FIGURE 8. Average per-user SE (with 95% confidence interval) achieved by PUNC with FPA, for different precoding schemes and values of ν. The average is taken over 200 network snapshots. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

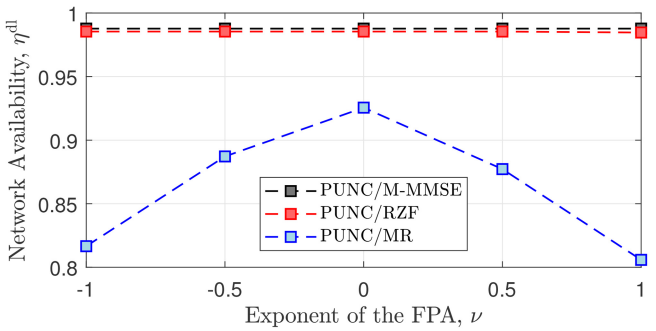

FIGURE 9. Network availability achieved by PUNC with FPA, for different precoding schemes and values of ν. Settings: $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$, $T = 5$, $n_d = 100$.

FPA, upon properly setting the design parameters ν and ω, is an effective choice to keep the network availability high while preventing any eMBB service outage.

In this regard, we now focus on how to select ν and ω appropriately. By using the same settings as in the first set of simulations, in Fig. 6 we plot the average per-user SE assuming SPC and different precoding schemes with FPA as ν and ω vary.

From the eMBB user perspective, it is preferable setting a small value for ω, and ν in the interval $[-0.5, 0]$. While the former is trivial, the latter needs further discussions. Indeed, recall that positive values for ν enable allocating more power to users with better channel conditions. Since we assume the URLLC users are uniformly distributed in a smaller area surrounding the BSs, it is very likely that they are closer to the BS than most of the eMBB users. Therefore, negative values for ν increase the fairness and improve eMBB users performance. Large values for both ω and ν excessively unbalance the power distribution in favor of the URLLC users, degrading the SE of the eMBB users.

Conversely, small values for both ω and ν break down the network availability of the URLLC users in SPC operation, as clearly seen in Fig. 7. Nevertheless, both M-MMSE and RZF are able to provide levels of network availability close to 1 except when $\nu = -1$, while MR is quite sensitive to this parameters tuning. Suppressing the multi-user interference is of a vital importance when SPC is adopted, and RZF, although not dealing with the inter-cell interference, is an excellent trade-off between performance and practicality. Fine-tuning the parameters of the FPA scheme yields satisfying performance when using MR. FPA becomes a valid, heuristic alternative to combat the multi-user interference whenever the latter cannot be removed by the precoding technique.

Setting ω becomes pointless when using PUNC with FPA as only URLLC transmissions take place in the considered slot. Hence, in Fig. 8 and Fig. 9 we focus on the average SE per user and the network availability as only ν varies. For both cases we notice that an equal power allocation, i.e., $\nu = 0$, is desirable. As per the SE of the eMBB users, negative values of ν support lower SEs (e.g., the 95%-likely SE per user), hence the fairness among the users, while

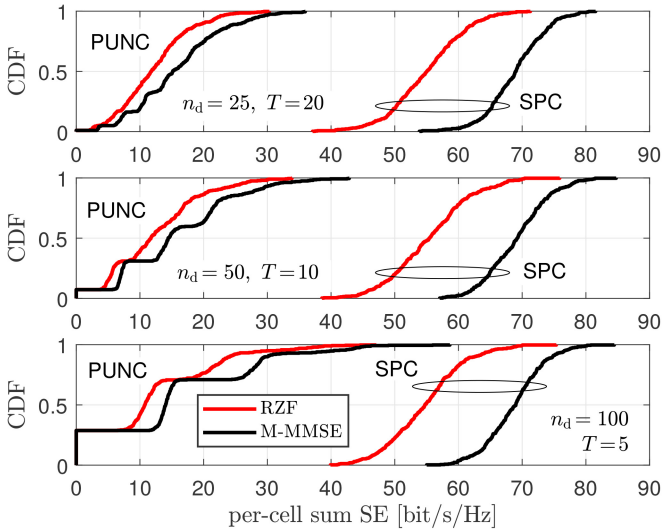

FIGURE 10. CDFs of the achievable downlink sum SE per cell, for different transmission and precoding strategies, as the number of slots per frame varies. Settings: FPA with $\nu = 0$ and $\omega = 0.2$, $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$.

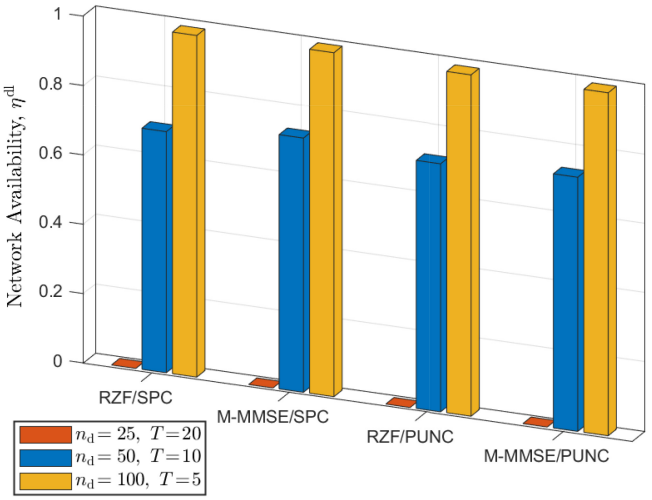

FIGURE 11. Network availability, for different transmission and precoding strategies, as the length of the slot varies. Settings: FPA with $\nu = 0$ and $\omega = 0.2$, $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_p = 80$.

large positive values of ν support the peak SE in a greedy fashion, neglecting lower SEs. Therefore, $\nu = 0$ is sound if the average SE is targeted, especially when the multi-user interference is partially or fully canceled.

As per the network availability of the URLLC users, any choice of $\nu \in [-1, 1]$ is solid as long as M-MMSE or RZF are employed, while the performance of MR is relatively penalized whenever a non-neutral choice for ν is taken. Presumably, the number of URLLC users simultaneously active in the same slot (resulting from the chosen values of α and a_u) is such that the multi-user interference is not significant.

Next, we evaluate the performance as a function of the number of the slots in a TDD frame, T, and the size of the slot, n_d, which in turn determines the URLLC codeword length. In this set of simulations and hereafter, we omit the results achieved by MR and only consider FPA with $\nu = 0$ and $\omega = \alpha$ motivated by the previous results. Fig. 10 shows the CDFs of the sum SE per cell, for three different setups: (*i*) $n_d = 25$, $T = 20$, (*ii*) $n_d = 50$, $T = 10$, and (*iii*) $n_d = 100$, $T = 5$.

The structure of the TDD frame has not a significant impact on the SE of the eMBB users when SPC is used. Conversely, that deeply affects the per-cell sum SE in case of PUNC. Indeed, increasing the number of slots per frame makes the probability of having eMBB service outage smaller as it increases the opportunities for an eMBB user to find slots with no active URLLC users. This argument is supported by the results in Fig. 10 in which the eMBB service outage equals 0.01, 0.0725 and 0.2875 when $T = 20$, $T = 10$ and $T = 5$, respectively. On the other hand, with fewer slots, eMBB users might be active for longer time, thereby experiencing higher SE. This explains the larger variations of the per-cell sum SE as T is decreased.

The length of the slot directly affects the performance of the URLLC users. As we can see in Fig. 11, the network availability increases drastically with the length of the slot (i.e., the URLLC codeword length).

In fact, the length of the URLLC codeword determines the transmission rate of the URLLC users as $R = b/n_d$, thus the shorter the codeword the higher the rate requirement to be reliably achieved and, in turn, the larger the error probability.[3] Again, SPC is the technique that overall guarantees the best performance to both the eMBB and URLLC users as its main limitation, namely the caused multi-user interference, is overcome by using interference-suppression-based precoding schemes. Lastly, although letting the URLLC transmissions span many channel uses is beneficial in terms of network availability, the latency requirements impose to localize the transmissions in time.

Now, we move our focus on the impact of the pilot contamination and estimation overhead on the performance. By fixing the TDD frame length and the number of slots per frame, we vary the length of the uplink training, hence the number of available orthogonal pilots, and the length of each slot accordingly. In Fig. 12 we show how the average sum SE per cell evolves in different operating regimes with respect to the uplink training length.

In these simulations, we assume $K = 20$, $\alpha = 0.2$, $\tau_c = 580$ and $T = 5$. Small values of τ_p entails low channel estimation overhead but high levels of pilot contamination which reduces the effectiveness of the precoding. Our pilot assignment scheme preserves the performance of the URLLC users by assigning them unique pilots if available, otherwise pilots are assigned randomly and contamination hits any user indiscriminately. The maximum number of URLLC users potentially active in this scenario is, according to the chosen

3. The random-coding union bound (RCU) in (18) defines the error probability as the probability that the average generalized information density is smaller than the transmission rate requirement.

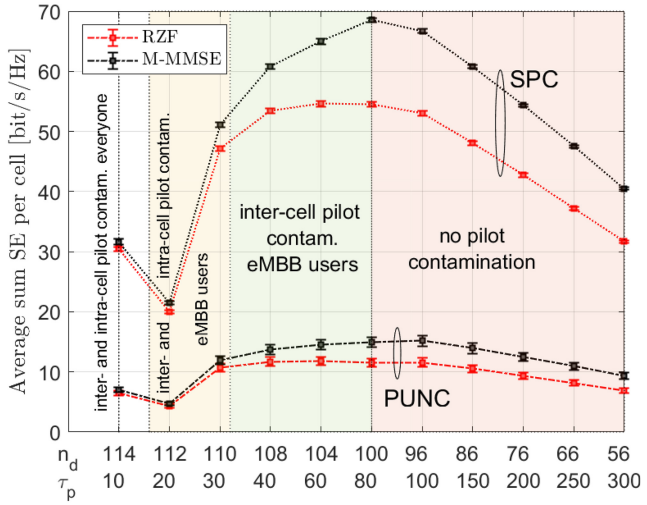

FIGURE 12. Average SE per cell (with 95% confidence interval), for different transmission and precoding strategies, as τ_p (and n_d) varies. The average is taken over 200 network snapshots. Settings: FPA with $\nu = 0$ and $\omega = 0.2$, $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_c = 580$, $T = 5$.

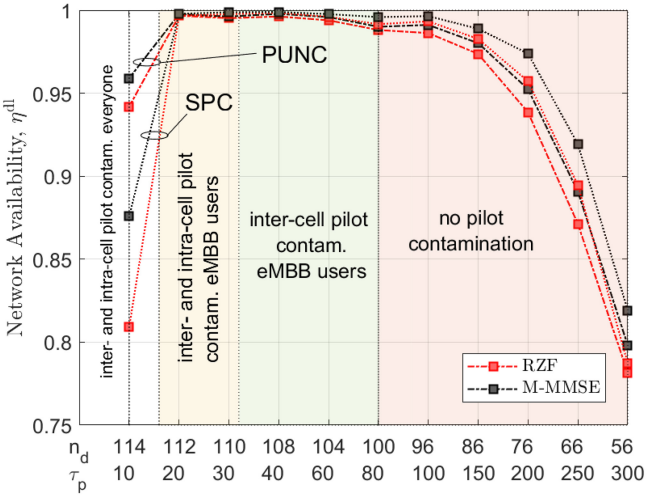

FIGURE 13. Network availability, for different transmission and precoding strategies, as τ_p (and n_d) varies. Settings: FPA with $\nu = 0$ and $\omega = 0.2$, $K = 20$, $\alpha = 0.2$, $a_u = 10^{-0.5}$, $\tau_c = 580$, $T = 5$.

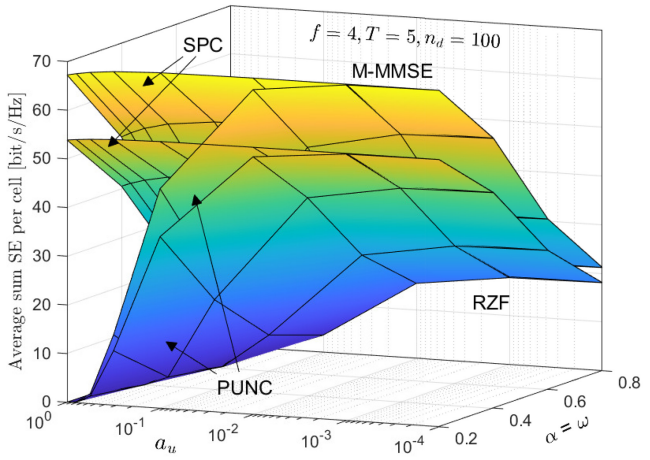

FIGURE 14. Average SE per cell, for different transmission and precoding strategies, as a_u and α vary. The average is taken over 200 network snapshots. Settings: FPA with $\nu = 0$ and $\omega = \alpha$, $K = 20$, $\tau_c = 580$, $f = 4$, $T = 5$, $n_d = 100$.

parameter, 16. Hence, pilots are assigned randomly when $\tau_p = 10$ causing both intra- and inter-cell pilot contamination and providing a low sum SE per cell, namely about 30 bit/s/Hz with SPC and less than 10 bit/s/Hz with PUNC. The performance worsens when $\tau_p = 20$ as the eMBB users have to share only 4 orthogonal pilots since the protection mechanism of the URLLC users is now triggered. As we increase the value of τ_p, the intra-cell pilot contamination is primarily reduced by assigning orthogonal pilots to eMBB users of the same cell. If $\tau_p \geq 32$ then intra-cell pilot contamination is prevented and the inter-cell interference among the eMBB users remains the only impairment. The sum SE per cell keep growing up to $\tau_p = 80$, when all the users in the network are assigned mutual orthogonal pilots and the benefits of having no pilot contamination at all overcome the penalty from increasing the estimation overhead. Trivially, there are no benefits in the channel estimation when further increasing τ_p, while the estimation overhead turns to be expensive and drastically lowers the sum SE per cell. Finally, notice that RZF and M-MMSE provide essentially the same performance when both the intra- and inter-cell pilot contamination occur, because the ability of suppressing the multi-user interference is poor for both the schemes.

As per the URLLC users, pilot contamination heavily affects the network availability when $\tau_p < 16$, especially when SPC is employed and despite a long slot lowers the rate requirements, as we can observe in Fig. 13.

Pilot contamination among URLLC users is destructive mainly because they are likely to be close to the BS and to each other, experiencing strong interference that cannot be resolved when their channel estimates are correlated. Hence, our approach aiming at prioritizing the URLLC users in the pilot assignment is technically sound. In addition, increasing the estimation overhead deeply penalizes the network availability since more resources are subtracted to the data transmission, namely the slot length reduces and, as already explained earlier, the rate requirements of the URLLC users increase.

Next we study how the performance are affected by the random activation pattern and the number of potentially active URLLC users per frame. Fig. 14 shows the average sum SE per cell as a_u and α vary, assuming different transmission and precoding schemes, and FPA with $\nu = 0$ and $\omega = \alpha$. Notice that, proportionally increasing ω to α is a reasonable approach for SPC as more power is allocated to an increasing number of potentially active URLLC users, especially for large values of a_u.

In these simulations, we assume two TDD frame configurations: (*i*) $f = 4$, $T = 5$, $n_d = 100$, and (*ii*)

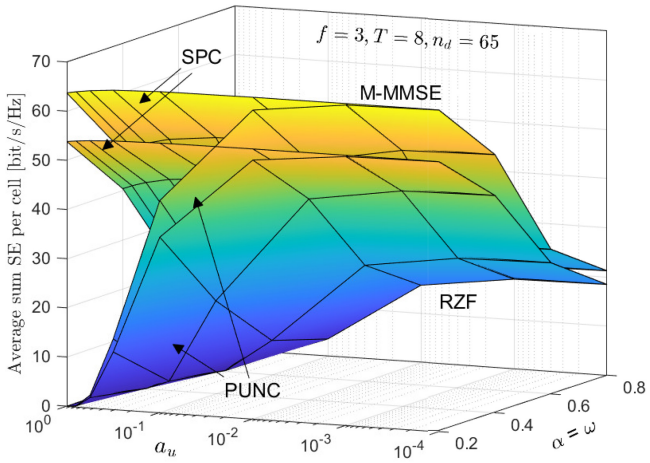

FIGURE 15. Average SE per cell, for different transmission and precoding strategies, as a_u and α vary. The average is taken over 200 network snapshots. Settings: FPA with $\nu = 0$ and $\omega = \alpha$, $K = 20$, $\tau_c = 580$, $f = 3$, $T = 8$, $n_d = 65$.

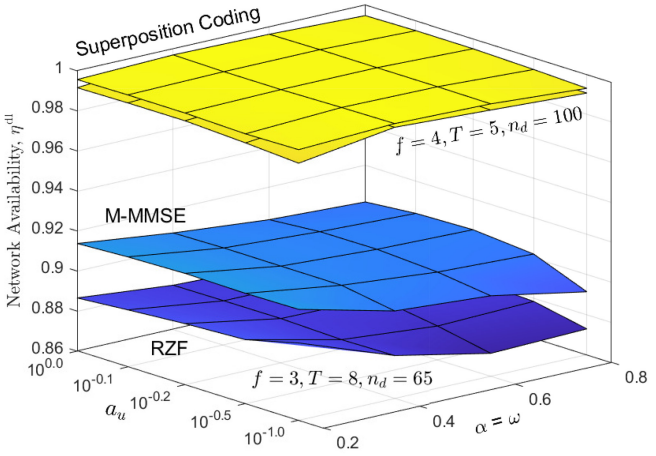

FIGURE 16. Network availability, for different precoding strategies, as a_u and α vary. The average is taken over 200 network snapshots. Settings: SPC and FPA with $\nu = 0$ and $\omega = \alpha$, $K = 20$, $\tau_c = 580$. Two TDD frame configurations are considered.

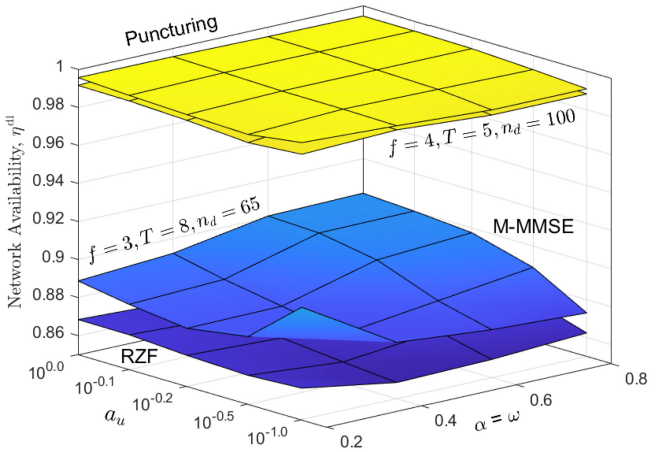

FIGURE 17. Network availability, for different precoding strategies, as a_u and α vary. Settings: PUNC and FPA with $\nu = 0$ and $\omega = \alpha$, $K = 20$, $\tau_c = 580$. Two TDD frame configurations are considered.

$f = 3$, $T = 8$, $n_d = 65$ (whose results are instead shown in Fig. 15). First, we observe that similar average sum SE per cell can be achieved by adopting the considered TDD frame configurations: pilot contamination is what slightly degrades the performance of the eMBB users when using the second frame configuration. The performance of PUNC converges to that of SPC when $a_u \geq 10^{-2}$, hence for sparse activation patterns, as expected. Again, the performance gap between RZF and M-MMSE reduces in the second scenario (Fig. 15) as the inter-cell pilot contamination decreases the ability of M-MMSE in suppressing the multi-user interference. PUNC provides eMBB service outage for large values of a_u, whereas SPC is still able to cancel the URLLC user interference and to provide excellent SEs. Lastly, we observe that if the 80% of the users requests URLLC, then the performance of the eMBB users is reduced of almost one third with respect to the case $\alpha = 0.2$. This result is mainly due to the chosen value of ω in the FPA scheme that aims to favor the URLLC performance as the number of URLLC users increases.

The performance achieved by the two considered TDD frame configurations appreciably differ in terms of network availability as shown in Fig. 16 for SPC and Fig. 17 for PUNC. In both cases, reducing the length of the slot leads to about a 10% performance loss, while the pilot contamination only concerns the eMBB users. This performance gap is slightly more pronounced when using PUNC because the entire BS power is distributed among the URLLC users causing stronger mutual interference. Overall, the first TDD frame configuration turns to be quite robust to any of the considered transmission and precoding strategies, considered random URLLC activation pattern and URLLC user load.

A final aspect to be analyzed for this set of simulations is how the probability of eMBB service outage varies with a_u and α when PUNC is adopted. This would complete the picture on which operating points PUNC is an effective choice for the eMBB users too, and importantly, further remark the relevance of properly structuring the TDD frame.

As we can see in Fig. 18, the advantage of adopting the TDD frame configuration with $T = 8$ slots, when using PUNC, consists in better preventing the eMBB service outage than the configuration with $T = 5$. For instance, when $a_u = 10^{-1}$ and $\alpha = 0.8$ or $\alpha = 0.6$, partitioning the share of the frame devoted to the data transmission in 8 slots enables to halve the eMBB outage service compared to the case where 5 slots are adopted. Overall, PUNC can compete with SPC only in scenarios with low URLLC traffic loads, upon properly structuring the TDD frame, as long as a moderate eMBB performance loss is tolerated, either in terms of sum SE per cell or of eMBB service outage. On the other hand, SPC hinges on precoding schemes able to suppress the multi-user interference which, in turn, leverages the spatial degrees of freedom available at the BS and the high accuracy of the acquired CSI.

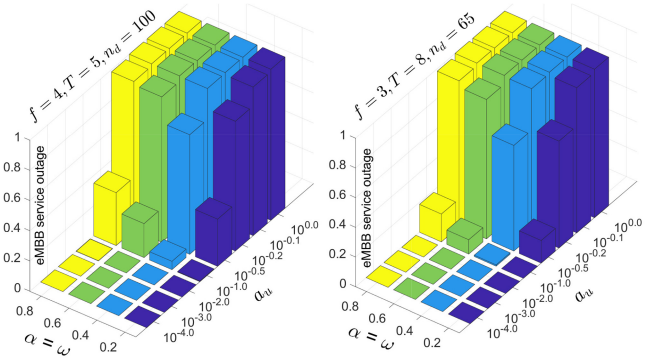

FIGURE 18. eMBB service outage, for different precoding strategies, as a_u and α vary. Settings: PUNC and FPA with $\nu = 0$ and $\omega = \alpha$, $K = 20$, $\tau_c = 580$. Two TDD frame configurations are considered.

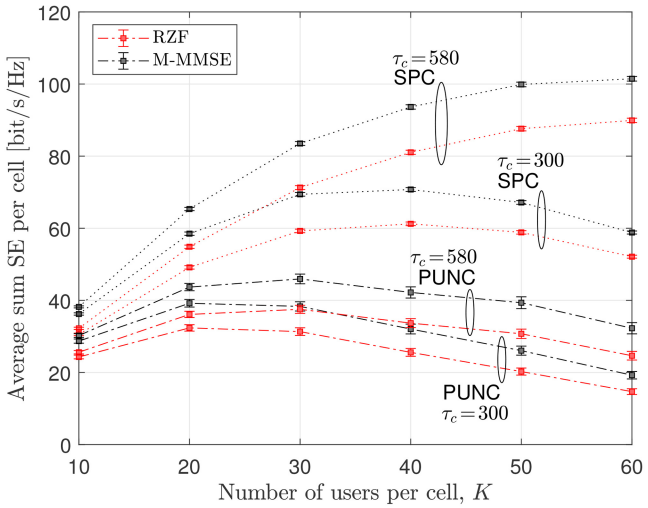

FIGURE 19. Average SE per cell (with 95% confidence interval), for different transmission and precoding strategies, as K and τ_c vary. The average is taken over 200 network snapshots. Settings: FPA with $\nu = 0$ and $\omega = 0.2$, $\alpha = 0.2$, $a_u = 10^{-1}$, $f = 3$, $T = 5$.

Finally, we evaluate the performance varying the total number of users and the TDD frame length. Fig. 19 shows the average sum SE per cell, for different transmission and precoding strategies, as the number of users per cell, K, grows from 10 to 60, and considering two different TDD frame lengths, namely 580 and 300 channel uses. The latter may support a shorter coherence time and a narrower coherence bandwidth as well as a higher user mobility compared to the case with 580 channel uses. However, a shorter frame entails less resources that can be allocated to the data transmission and uplink training.

In these simulations we assume FPA with $\nu = 0$ and $\omega = 0.2$, $\alpha = 0.2$, $a_u = 10^{-1}$, $T = 5$ and pilot reuse factor $f = 3$. Moreover, as $\tau_p = fK$ and τ_c is fixed, for each value of K we have different configurations of uplink training and slot length, i.e., τ_p and n_d, respectively. From Fig. 19 we observe the average sum SE per cell increasing with K, which demonstrates the great ability of SPC with M-MMSE and RZF to spatially multiplex the users. The average sum SE per cell saturates for values of K larger than 60 for $\tau_c = 580$, and around 40 for $\tau_c = 300$ wherein the channel estimation overhead heavily burden the SE. PUNC is far inferior to SPC because allocates less resources to the eMBB users and the performance gap increases with K as the number of URLLC users per cell grows proportionally. Therefore, letting K increase makes punctured slots more likely, which not only subtracts resources to the eMBB user reducing its SE but also increases the eMBB service outage, as shown in Table 3. Notice that, the eMBB service outage does not change when varying τ_c as long as T is fixed.

TABLE 3. Network availability and eMBB service outage, $\tau_c = 580$.

K	τ_p	n_d	η^{dl}				Sout
			SPC		PUNC		PUNC
			M-MMSE	RZF	M-MMSE	RZF	
10	30	110	0.9989	0.9966	1	0.9989	0.0012
20	60	104	0.9988	0.9957	0.9944	0.9906	0.0038
30	90	98	0.9988	0.9950	0.9934	0.9893	0.0225
40	120	92	0.9969	0.9885	0.9881	0.9819	0.0625
50	150	86	0.9864	0.9787	0.9790	0.9672	0.1050
60	180	80	0.9807	0.9697	0.9728	0.9601	0.1737

TABLE 4. Network availability and eMBB service outage, $\tau_c = 300$.

K	τ_p	n_d	η^{dl}				Sout
			SPC		PUNC		PUNC
			M-MMSE	RZF	M-MMSE	RZF	
10	30	54	0.7936	0.7683	0.7844	0.7534	0.0012
20	60	48	0.6786	0.6353	0.6905	0.6685	0.0038
30	90	42	0.4796	0.4296	0.5646	0.5435	0.0225
40	120	36	0.1813	0.1457	0.3192	0.3192	0.0625
50	150	30	0.0021	0	0.0250	0.0250	0.1050
60	180	24	0	0	0	0	0.1737

Table 3 and Table 4 show the network availability for different transmission and precoding strategies, and different values of K, also emphasizing how τ_p and n_d vary accordingly to meet the TDD frame length. In particular, Table 3 shows the performance achieved by considering $\tau_c = 580$, while Table 4 shows the performance achieved with $\tau_c = 300$. The TDD frame with $\tau_c = 580$ allows to achieve a network availability above 96% up to 60 users per cell (of which 12 are URLLC users) with any of the considered transmission and precoding techniques, meaning that such an amount of resources are sufficient to excellently support the considered URLLC user loads and their activation pattern. Conversely, the network availability supported by the TDD frame with $\tau_c = 300$, reported in Table 4, is considerably lower, even close (or equal) to zero for $K \geq 50$, emphasizing how sensitive the network availability is to the length of the TDD frame, hence to the amount of available resources. Importantly, we observe the decreasing trend of the network availability as K increases, which for PUNC is milder and mainly due to the shorter URLLC codeword length, but for SPC is severe and mainly due to the increase

of the multi-user interference. Indeed, the results in Table 4 clearly confirms that PUNC is more robust than SPC when $K \geq 20$.

VI. PRACTICAL URLLC DECODERS

As per the URLLC performance, this work provides an information-theoretic study on the error probability in the nonasymptotic finite-blocklength regime. Our analysis hinges on the random-coding union (RCU) achievability bound with parameter s, mathematically expressed in (18) and introduced in [26, Th. 1] to the case of massive MIMO systems. Importantly, the RCU bound is the upper bound on the best error probability achieved by $(\mathtt{n}, \mathtt{k})$ codes [35, Th. 2], where $\mathtt{n}$ is the blocklength and $\mathtt{k}$ is the information blocklength. Therefore, the obtained results do not reflect a specific decoding implementation, but rather relies on a mathematically-tractable information-theoretic framework.

In this section, we shed lights on candidate low-complexity decoding techniques that are able to attain the URLLC requirements for short blocklength codes, and potentially to achieve, with high accuracy, the performance predicted by the RCU bound based on the SNN decoding rule analyzed in this paper.

A comprehensive overview on efficient URLLC decoders for short blocklength codes has been recently given in [36]. Candidate decoding schemes for URLLC applications are the ordered-statistics decoding (OSD) [37], guessing random additive noise decoding (GRAND) [38], and successive cancellation list (SCL) algorithm [39]. As demonstrated in [36], both OSD and GRAND combined with short blocklength codes, such as the Bose-Chaudhuri-Hocquengham (eBCH), the cyclic-redundancy-check-aided (CRC) polar and the polarization-adjusted convolutional (PAC) codes, have shown a great ability to achieve the near maximum-likelihood (ML) decoding performance,[4] but the SCL-based sequential decoder offers a better performance-complexity trade-off. As the complexity of the decoders decreases with the SNR, some specific considerations are needed. OSD decoders are the least complex in the low SNR regime, whereas GRAND and SCL-based decoders are more efficient at high SNRs. However, GRAND suffers of high *worst-case decoding* complexity, namely the complexity required to perform decoding over code blocks with low reliability [40]. The worst-case decoding complexity is thus a crucial factor for URLLCs. The decoding complexity also scale with the blocklength $\mathtt{n}$. SCL-based decoding complexity approaches $\mathcal{O}(\mathtt{n}\log\mathtt{n})$ at high SNRs, whereas the complexity of OSD and GRAND scales more rapidly with $\mathtt{n}$. Although OSD and GRAND are universal decoders capable of flexibly decoding unstructured codes with varying blocklength and rate, their use is preferable with relatively short codes. Importantly, simulations results in [36] reveal almost a perfect matching between the RCU bound and the performances achieved by OSD, and SCL-based decoders with CRC-polar codes for the high reliability requirements of interest in URLLC applications. Similarly, the authors in [29] demonstrate the accuracy and validity of the RCU-SNN achievability bound by comparing that with the performance achieved by a practical OSD-SNN decoder operating on short quasi-cyclic (QC) binary block codes and showing a gap within 1 dB for any blocklength.

In URLLC applications the decoding delay and the number of retransmissions dominate the end-to-end latency. Our information-theoretic framework considers communications with fixed-blocklength and no feedback (FBL-NF), thereby not taking retransmissions into account. Indeed, the excellent levels of network availability shown in the simulation results, especially for SPC, are achieved with only one transmission. The authors in [41] derive an achievability bound for low-latency short-packet communications including hybrid automatic repeat request (HARQ) and based on mismatched SNN decoding. The varying communication latency of HARQ is then compared with its fixed counterpart attained by the FBL-NF scheme. Fast HARQ protocols for URLLC are proposed in [42], along with a general characterization of the decoding delay in the system model. The decoding delay is certainly proportional to the decoding complexity. OSD and GRAND are capable to reduce the decoding time by parallelizing their operations [36]. Conversely, SCL-based decoders are sequential by nature, hence decoding time and complexity go hand in hand. Low decoding latency and complexity are guaranteed by OSDs and SCL-based decoders at low and high SNRs, respectively.

VII. CONCLUSION

We investigated the non-orthogonal multiplexing of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) in the downlink of a multi-cell massive MIMO system. Such heterogeneous services calls for effective resource allocation strategies to let eMBB and URLLC peacefully coexist.

Firstly, we provided a unified information-theoretic framework to assess the spectral efficiency (SE) of the eMBB in the infinite-blocklength ergodic regime, and the error probability of the URLLC in the nonasymptotic finite-blocklength regime, whose analysis relies on mismatched receivers and on the so-called saddlepoint approximation.

Secondly, we generalized the proposed framework to accommodate two alternative coexistence strategies: puncturing (PUNC) and superposition coding (SPC). The former prevents the inter-service interference aiming to protect the URLLC reliability, while the latter accepts it aiming to maintain the eMBB service.

Thirdly, we numerically evaluated the performance achieved by PUNC and SPC under different precoding and power allocation schemes, and subject to different configurations of the time-division duplex radio frame and URLLC random activation pattern.

4. The SNN decoding rule coincides with the optimal ML decoding rule under the assumption of perfect CSI knowledge.

Simulation results revealed that the spatial degrees of freedom available at the BSs, when fully exploited by interference-suppression-based precoding schemes, and upon a high-quality CSI acquisition, enable to resolve the interference caused by SPC, providing way higher eMBB SE than PUNC, yet ensuring great levels of error probability to the URLLC. However, PUNC is necessary to preserve the URLLC performance at the expense of some eMBB service outage, whenever the interference is not properly suppressed by the precoding technique (e.g., due to a severe pilot contamination, or because of limited degrees of freedom).

Extensions to this work may concern the study of coexistence strategies for the uplink, also including massive machine-type communications (mMTCs). Investigating the non-orthogonal multiplexing of heterogeneous services in distributed user-centric systems, such as cell-free massive MIMO [43], [44], [45], [46], is certainly another appealing future research direction. Last but not least, future works may concern a further generalization of the achievability bound for low-latency short-packet communications to include HARQ mechanisms.

REFERENCES

[1] *IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, ITU, Geneva, Switzerland, Rec. M.2083-0, 2015.

[2] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[5] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[6] P. Popovski et al., "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

[7] A.-S. Bana et al., "Massive MIMO for Internet of Things (IoT) connectivity," *Phys. Commun.*, vol. 37, Dec. 2019, Art. no. 100859. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874490719303891

[8] J. Östman, A. Lancho, G. Durisi, and L. Sanguinetti, "URLLC with massive MIMO: Analysis and design at finite blocklength," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6387–6401, Oct. 2021.

[9] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.

[10] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 1970–1978.

[11] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[12] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.

[13] S. F. Abedin, M. G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 489–502, Jan. 2019.

[14] A. Matera, R. Kassab, O. Simeone, and U. Spagnolini, "Non-orthogonal eMBB-URLLC radio access for cloud radio access networks with analog fronthauling," *Entropy*, vol. 20, no. 9, p. 661, 2018.

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

[16] R. Abreu et al., "On the multiplexing of broadband traffic and grant-free ultra-reliable communication in uplink," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Apr. 2019, pp. 1–6.

[17] E. N. Tominaga, H. Alves, R. D. Souza, J. L. Rebelatto, and M. Latva-aho, "Non-orthogonal multiple access and network slicing: Scalable coexistence of eMBB and URLLC," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–6.

[18] F. Saggese, M. Moretti, and P. Popovski, "Power minimization of downlink spectrum slicing for eMBB and URLLC users," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 11051–11065, Dec. 2022.

[19] R. Kassab, O. Simeone, P. Popovski, and T. Islam, "Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures," *IEEE Access*, vol. 7, pp. 13035–13049, 2019.

[20] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[21] J. Zeng, T. Lv, R. P. Liu, X. Su, Y. J. Guo, and N. C. Beaulieu, "Enabling ultrareliable and low-latency communications under shadow fading by massive MU-MIMO," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 234–246, Jan. 2020.

[22] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[23] A. A. Nasir, H. D. Tuan, H. Q. Ngo, T. Q. Duong, and H. V. Poor, "Cell-free massive MIMO in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5861–5871, Sep. 2021.

[24] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[25] J. Scarlett, A. Martinez, and A. Guillén I Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[26] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random-coding bounds," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, Feb. 2011, pp. 1–6.

[27] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[28] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[29] J. Östman, G. Durisi, E. G. Ström, M. C. Coçkun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1521–1536, Feb. 2019.

[30] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[31] A. Lapidoth and S. Shamai, "Fading channels: How perfect need 'perfect side information' be?" *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[32] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[33] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, 3GPP Standard TS 36.814, Mar. 2017.

[34] *Service Requirements for Cyber-Physical Control Applications in Vertical Domains (Release 17)*, 3GPP Standard TS 22.104, V.17.2.0, Dec. 2019.

[35] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854–6883, Dec. 2016.

[36] C. Yue, V. Miloslavskaya, M. Shirvanimoghaddam, B. Vucetic, and Y. Li, "Efficient decoders for short block length codes in 6G URLLC," 2022, *arXiv:2206.09572*.
[37] M. P. C. Fossorier and S. Lin, "Soft-decision decoding of linear block codes based on ordered statistics," *IEEE Trans. Inf. Theory*, vol. 41, no. 5, pp. 1379–1396, Sep. 1995.
[38] K. R. Duffy, J. Li, and M. Médard, "Capacity-achieving guessing random additive noise decoding," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4023–4040, Jul. 2019.
[39] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015.
[40] K. R. Duffy, M. Médard, and W. An, "Guessing random additive noise decoding with symbol reliability information (SRGRAND)," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 3–18, Jan. 2022.
[41] J. Östman, R. Devassy, G. C. Ferrante, and G. Durisi, "Low-latency short-packet transmissions: Fixed length or HARQ?" in *Proc. IEEE Global Commun. Conf. Workshops (GLOBECOM Wkshps)*, Dec. 2018, pp. 1–6.
[42] B. Makki, T. Svensson, G. Caire, and M. Zorzi, "Fast HARQ over finite blocklength codes: A technique for low-latency reliable communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 194–209, Jan. 2019.
[43] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
[44] S. Buzzi and C. D'Andrea, "User-centric communications versus cell-free massive MIMO for 5G cellular networks," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, Jun. 2017, pp. 1–6.
[45] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, p. 197, 2019.
[46] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.

Open Access funding provided by 'Università degli Studi di Cassino e del Lazio Meridionale' within the CRUI CARE Agreement