Robust Downlink Multi-Antenna Beamforming With Heterogenous CSI: Enabling eMBB and URLLC Coexistence

Dian Echevarría Pérez [ID], *Graduate Student Member, IEEE*, Onel L. Alcaraz López [ID], *Member, IEEE*, and Hirley Alves [ID], *Member, IEEE*

Abstract— Two of the main problems to achieve ultra-reliable low-latency communications (URLLC) are related to instantaneous channel state information (I-CSI) acquisition and the coexistence with other service modes such as enhanced mobile broadband (eMBB). The former comes from the non-negligible time required for accurate I-CSI acquisition, while the latter, from the heterogeneous and conflicting requirements of different nodes sharing the same network resources. In this paper, we leverage the I-CSI of multiple eMBB links and the channel measurement's history of a URLLC user for multi-antenna beamforming design. Specifically, we propose a precoding design that minimizes the transmit power of a base station (BS) providing eMBB and URLLC services with signal-to-interference-plus-noise ratio (SINR) and outage constraints, respectively, by modifying existing I-CSI-based precoding schemes to account for URLLC channel history information. Moreover, we illustrate and validate the proposed method by adopting zero-forcing (ZF) and the transmit power minimization (TPM) precoding with SINR constraints. We show that the ZF implementation outperforms TPM in adverse channel conditions as in Rayleigh fading, while the situation is rapidly reversed as the channel experiences some line-of-sight (LOS). Finally, we determine the confidence levels at which the target outage probabilities are reached. For instance, we show that outage probabilities below 10^{-3} are achievable with more than 99% confidence for both precoding schemes under favorable LOS conditions with 16 transmit antennas and 500 samples of URLLC channel history.

Index Terms— Multi-antenna beamforming, channel history, CSI, eMBB, URLLC.

I. INTRODUCTION

ULTRA-RELIABLE low-latency communication (URLLC) is a key operation mode in current and future wireless communication networks. Many envisioned applications require reliability levels close to those offered by wired networks, *i.e.*, error probabilities below 10^{-5} almost 100 % of the time, and latency levels below 10 ms [1]. For instance, automotive communications require a user plane reliability of 10^{-5} and end-to-end (E2E) latencies below 5, 10, and 20 ms for assisted, cooperative, and tele-operated driving, respectively. Motion control in industrial processes requires also reliability levels of 10^{-5} and E2E latencies up to 1 ms [2].

In practice, it is difficult/challenging to meet both reliability and latency requirements simultaneously. Communication latency can be reduced by shortening the transmission time interval, using non-slot or mini-slot scheduling policies, and/or employing uplink (UL) grant-free transmissions [3]. On the other hand, diversity techniques, *e.g.*, time, frequency, or spatial diversity, are key to achieving high reliability levels. Notice that message retransmission may be seen as a time diversity technique to improve reliability, but at the cost of higher latencies [4], while frequency diversity may not always be available given spectrum-sharing/slicing constraints. Instead, spatial diversity, whose availability is increasing due to the rise of multiple-input multiple-output (MIMO) communications and node densification, is often the most appealing to support ultra-reliable services [1]. Specifically, precoding/combining allows increasing the signal power and/or suppressing the interference from other users or base stations (BS). This, in turn, improves the signal-to-interference-plus-noise ratio (SINR) statistics, the data decoding performance, and consequently leads to reliability enhancements. For both, precoding and combining procedures, channel state information (CSI) is commonly needed. However, instantaneous CSI (I-CSI) of URLLC links might be too costly to acquire if the latency constraints are too tight [5]. The energy consumption is also a critical aspect for the CSI acquisition. Low-energy devices may not afford to participate frequently in CSI acquisition procedures due to the corresponding non-negligible energy expenditure. Hence, other approaches must be considered, e.g., exploiting channel statistics (*e.g.*, mean and covariance matrix) instead of instantaneous channel realizations. These statistics do not change regularly, especially in slow fading scenarios where the coherence time is larger than the delay requirements of the application. In general, the use of channel statistics may be a suitable option when delay and/or energy constraints are strict, while I-CSI acquisition procedures may be carried out when the conditions are more favorable.

Several works have been conducted exploiting the channel statistics for multi-antenna precoding design. For instance, the authors in [6] considered the problem of transmit power minimization with channel covariance-based beamforming in multicast scenarios. The work in [7] focused on the beamforming design for weighted sum-rate maximization using combined channel mean and covariance information. The authors in [8] addressed the problem of downlink (DL) precoding design with mixed statistical and imperfect I-CSI in massive MIMO systems. They proposed extended zero-forcing (ZF) and maximal ratio transmission (MRT) methods to minimize the total transmit power. However, none of these works focused on supporting URLLC services. In this regard, the work in [9] exploited the sparsity of the propagation channel and relied on the estimation of a small number of channel coefficients for the beamforming design. Meanwhile, authors in [10] proposed a beamformer that maximizes the users' minimum rate in an interference limited multi-user system with short packet transmissions and URLLC constraints.

Notice that URLLC services will necessarily coexist with other operation modes such as enhanced mobile broadband (eMMB). Such coexistence is of paramount importance in current and future wireless communication networks and has attracted a lot of attention in recent years [11], [12], [13], [14], [15], [16], [17]. On the one hand, URLLC users transmit usually at low data rates, are characterized by intermittent activation patterns mainly associated with external events, such as alarms, and require transmitting short messages, in the order of a few tens of bytes. On the other hand, eMBB users require high data rates with steady activation patterns [18]. These fundamental differences make the network design to meet all the requirements, a challenging task. Therefore, to cope with the quality-of-service (QoS) requirements of all nodes, it is necessary to develop efficient multiplexing techniques for URRLC and eMBB users. DL channel preemptive scheduling, where the data of the URLLC user is transmitted immediately and overwrites the current transmission intended for the eMBB user, is an example technique. The advantage of this method is that the transmission intended for the URLLC user does not have to wait for the scheduled slot, and the drawback is a performance degradation of the eMBB user [11]. To tackle this issue, the authors in [13] were the first to explore resource allocation for joint scheduling of URLLC and eMBB traffic using puncturing/superposition based methods. Therein, they investigated various models, i.e., linear, convex, and threshold-based, for describing the impact of the URLLC traffic load on the rate loss of the eMBB users. In [16], the authors focused on the co-scheduling of URLLC and eMBB traffic based on puncturing, and aimed to maximize the minimum expected achieved rate of eMBB users while fulfilling the URLLC traffic demands. Meanwhile, the multiplexing of eMBB and URLLC traffic in the DL channel was analyzed in [15]. Specifically, a resource allocation problem in each mini-slot was formulated as an integer programming problem to maximize an eMBB utility function while satisfying URLLC constraints. The work in [14] presented a risk-sensitive based formulation that allocates relatively more URLLC traffic to the network resources reserved for eMBB users with higher

data rates. In [17], a resource slicing optimization problem was formulated to maximize the eMBB data rate while satisfying the performance requirements of the URLLC traffic. Therein, a deep reinforcement learning framework, including eMBB resource allocation followed by URLLC scheduling, was proposed to solve the problem.

A. Motivation

There are still some open challenges for efficiently enabling URLLC-eMBB coexistence, which has not been considered in the previous works. For instance, how to efficiently multiplex the correspondingly heterogeneous services in time, frequency, and space. Even more challenging is how to design the spatial precoding if the latency constraint is too tight such that I-CSI cannot be acquired. The CSI history of URLLC links may be leveraged to address such an issue. Interestingly, relying on a limited number of past channel measurements to design URLLC-supporting precoders, although appealing, has not been considered in the literature to the best of authors' knowledge. Still, exploiting channel history has already proven valuable to enable URLLC [19], [20], [21], [22]. For instance, the authors in [19] proposed an interference prediction algorithm for supporting URLLC. Specifically, the interference dynamics were modeled as a discrete-time Markov chain with state transition probability matrices being estimated using past interference measurements. Meanwhile, machine-learning (ML) mechanisms were proposed in [20], [21], and [22] to support URLLC in different scenarios. In [20], the authors studied the coexistence design challenges of scheduled and non-scheduled URLLC traffic, and presented a distributed risk-aware ML solution for the corresponding radio resource management problem (RRM). In [21], the authors introduced ML and fountain codes into millimeter wave hybrid access, and proposed an adaptive channel assignment method for URLLC. The work in [22] characterized the wireless connectivity over dynamic channels via statistical learning methods, and measured the reliability of wireless connectivity in terms of the probability of channel blocking events. However, notice that the main drawback of ML-based mechanisms in the context of URLLC lies in the big data requirements, e.g., for model training, specially in dynamic environments, and/or the exploitation of latency-unfriendly feedback/signaling channels.

Furthermore, I-CSI might not be available for URLLC under tight latency constraints. Thus, herein we focus on exploiting URLLC channel history for precoding design in heterogeneous scenarios. Notice that the system may provide service to a more significant number of devices by enabling a harmonious coexistence of URLLC and eMBB services in the same resource blocks (time-frequency) [12].

B. Contributions

In our work, we focus on transmit power minimization through precoding design in heterogeneous scenarios with coexisting URLLC and eMBB DL users, and no I-CSI availability for URLLC services. Specifically, our contributions are four-fold:

- We formulate a precoding optimization problem concerning transmit power minimization while ensuring URRLC and eMMB coexistence with different CSI availability. Furthermore, we exploit the Chernoff bound to stochastically model, impose, and guarantee the reliability requirements of the URLLC user based on its channel history.

- We propose an algorithm that leverages existing I-CSI-based precoding methods to solve the optimization problem. This allows taking advantage of efficient state-of-art precoders with relatively low implementation difficulty. We show that the algorithm complexity grows with the number of iterations ζ and the third power of the total number of users $(K+1)^3$.

- We evaluate the performance of the proposed algorithm with ZF precoding and the transmit power minimization (TPM) precoding with per-user SINR constraints. We show that ZF outperforms TPM in poor channel conditions, e.g., Rayleigh fading, while TPM exhibits superior performance in more deterministic channels, e.g., Rician fading with a significant κ factor.

- We analyze the impact of ζ, the Chernoff bound auxiliary variable r, and the number of past channel measurements L on the system performance. We show that large values of r and ζ may reduce the transmit power significantly but at the cost of affecting the reliability performance in practice. We also determine the confidence levels required to reach the target outage probabilities. For instance, outage probabilities below 10^{-3} are achievable with more than 99% confidence for both precoding methods in favorable line-of-sight (LOS) conditions with 16 transmit antennas at the BS, and given $L = 500$ samples of URLLC channel history.

The work is structured as follows. In Section II, we describe the system model, main assumptions, and formulate the optimization problem. In Section III, we present the history-based beamforming design, the proposed algorithm, and discuss ZF and TPM-based implementations. In Section IV, we illustrate numerical results and validate the performance of the proposed algorithm. Finally, Section V concludes the paper.

Notation: Uppercase and lowercase boldface letters denote matrices and vectors, respectively. Superscript $(\cdot)^*$ denotes complex conjugate, $(\cdot)^H$ depicts the Hermitian operator, and $(\cdot)^{-1}$ represents the matrix inverse operation. $|| \cdot ||$ represents the norm of a vector, and $||\cdot||_F$ the Frobenius norm. Moreover, $\mathcal{CN}(\mathbf{v}, \mathbf{R})$ denotes a complex Gaussian distribution with mean vector $\mathbf{v}$ and covariance matrix $\mathbf{R}$. Finally, $\mathcal{U}(\upsilon_1, \upsilon_2)$ represents a uniform distribution in the range $[\upsilon_1, \upsilon_2]$, and $Q(\cdot)$ depicts the Q-function.

II. SYSTEM MODEL

We consider a scenario where a BS equipped with M antennas spatially multiplexes one URLLC user and K eMBB users, with $K + 1 \leq M$, within a resource block, *i.e.*, time-frequency resource, as depicted in Fig. 1 (a). Nevertheless, notice that the BS may serve multiple URLLC users within different resource blocks as shown in Fig. 1 (b), while herein,

TABLE I
MAIN SYMBOLS USED THROUGHOUT THE PAPER

Symbol	Definition
M	number of transmit antennas at the BS
K	total number of eMBB users
$\mathbf{h}_k$	channel vector between the BS and user k
$\tilde{\mathbf{h}}_{0,l}$	$l-$th past channel measurement of the URLLC user
$\mathbf{w}_k$	precoder intended to user k
$\mathbf{u}_k$	normalized precoder intended to user k
γ_k	SINR at user k
γ_k^{tar}	SINR target at user k
σ^2	noise power
s_k	complex baseband signal corresponding to user k
p_k	power allocated to user k
p_{max}	maximum transmit power at the BS
ξ	outage probability target
L	number of past measurements of the URRLC channel
$\hat{\mu}$	sample mean in the Chernoff bound framework
$\hat{s}$	sample standard deviation in the Chernoff bound framework
μ_{UB}	upper bound of the population mean
α	confidence of the upper bound μ_{UB}
ζ	number of generated channel coefficients
d_r	radius of the network deployment area
δ	path loss exponent
ψ	path gain
$\mathcal{O}_u$	outage probability of the URLLC user
CV	confidence value of $\log_{10} \mathcal{O}_u$
MV	estimated mean value of $\log_{10} \mathcal{O}_u$
SD	estimated standard deviation of $\log_{10} \mathcal{O}_u$

we focus on a single resource block operation without loss of generality. Scheduled eMBB users are continuously receiving data in the DL, and their QoS is guaranteed provided that their target SINR, γ_k^{tar} for user $k = 1, 2, \ldots K$, is surpassed. The I-CSI of the eMBB users is obtained via training/feedback methods prior to the DL transmissions, and is assumed to be always known at the BS. On the other hand, URLLC users with strict latency and reliability requirements do not receive data all the time in the DL, which agrees with the typically sporadic data transmissions in many URLLC use cases [18]. Moreover, cooperation for I-CSI is unaffordable under the considered latency constraints due to the delays that the required procedures would introduce [5]. Therefore, I-CSI of the URLLC link is not available at the BS before a DL transmission takes place. Instead, we assume that if the URLLC user does not receive critical information, it participates in frequent CSI acquisition also via training/feedback, while the BS performs I-CSI-based transmit beamforming to serve the K eMBB users. On the other hand, when a transmission to a URLLC user is required, the BS uses both the I-CSI of the eMMB users and the past channel measurements of the corresponding URLLC channel for the beamforming design.

A. Signal Model

The received signal at the $k-$th user at a certain time-frequency resource is given by

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \sum_{i \neq k} \mathbf{h}_k^H \mathbf{w}_i s_i + n_k, \tag{1}$$

where $\mathbf{h}_k \in \mathbb{C}^{M \times 1}$ represents the complex vector containing the channel coefficients between the M antennas of the BS and user k, $\mathbf{w}_k \in \mathbb{C}^{M \times 1}$ is the precoding vector intended to user
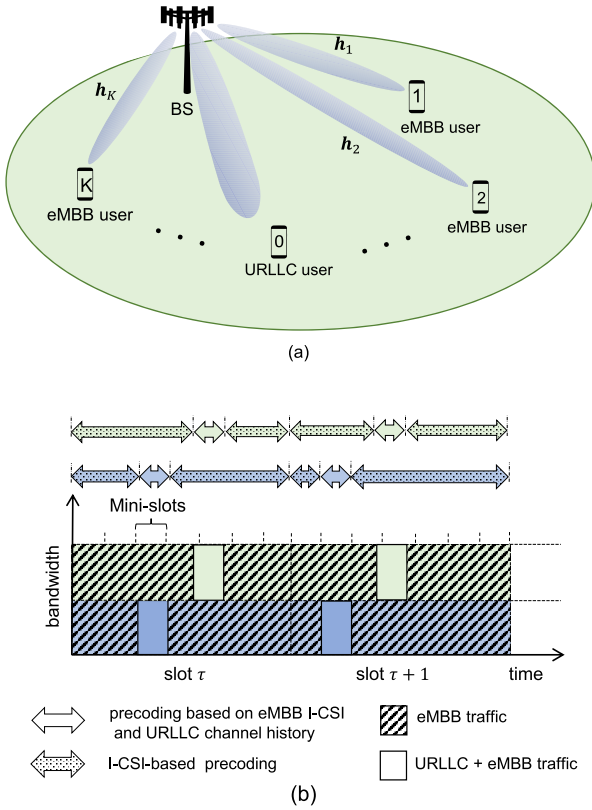
Fig. 1. (a) System model, and (b) multiplexing scheme. A BS serves one URLLC node and a set of K eMBB users within a time-frequency block in the DL channel. The transmit beams are perfectly focused on the direction of the eMBB users since they exploit the available I-CSI, but the beam towards the URRLC user is not perfectly oriented, neither sharp, since its design depends on imperfect channel statistics. Pure eMBB transmissions are served via I-CSI-based precoding, while the BS implements a hybrid precoding when a URLLC service is triggered and will coexist with the eMBB services, for which it leverages eMBB I-CSI and URLLC CSI history.

k, and s_k denotes the complex baseband signal transmitted to user k such that $\mathbb{E}\{s_k^* s_k\} = 1$ and $\mathbb{E}\{s_k^* s_i\} = 0 \ \forall k \neq i$. Finally, $n_k \sim \mathcal{CN}(0, \sigma^2)$ represents the additive white Gaussian noise. We use indexing from 0 to K with index 0 referring to the URLLC user. The SINR at user k is

$$\gamma_k(\{\mathbf{w}_k\}) = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{i \neq k} |\mathbf{h}_k^H \mathbf{w}_i|^2 + \sigma^2}. \tag{2}$$

B. Problem Formulation

Herein, we aim at configuring the transmit precoding that satisfies the URLLC and eMBB related QoS constraints of instantaneous DL transmissions with minimum power. For this, we set the following optimization problem

$$\mathbf{P1} : \underset{\{\mathbf{w}_i\}_{\forall i}}{\text{minimize}} \ \sum_{i=0}^{K} ||\mathbf{w}_i||_2^2 \tag{3a}$$

$$\text{subject to } \mathcal{O}_u = \text{P}\{\gamma_0(\{\mathbf{w}_k\}) < \gamma_0^{tar}\} \leq \xi, \tag{3b}$$

$$\gamma_k(\{\mathbf{w}_k\}) > \gamma_k^{tar}, \quad k = 1, 2, \ldots, K, \tag{3c}$$

where $\gamma_0(\{\mathbf{w}_k\})$ represents the receive SINR at the URLLC user. Moreover, γ_0^{tar} is the SINR threshold required to achieve

a successful URLLC transmission, ξ is maximum allowable outage probability, and γ_k^{tar} depicts the required SINR for proper operation of eMBB user k. In general, (3b) guarantees that the outage probability of the URLLC user, $\mathcal{O}_u$, is kept below the outage target.

Note that the objective function (3a) in $\mathbf{P1}$ is convex. On the other hand, the constraint (3c) is usually rearranged to solve the problem (without (3b)) via second order cone programming or semi-definite relaxation, among others. However, constraint (3b) is not convex and difficult to handle in general since the channel distribution is assumed unknown, and therefore, an expression for the outage probability is not available, which is typical in practical systems. Even if we consider an empirical approximation to the channel fading probability distribution, a large number of measurement samples, i.e., at least $10/\xi$, would be required, and the mathematical complexity for solving the optimization problem would be high anyway. Since the I-CSI of the URLLC link is not available to solve $\mathbf{P1}$, we rely on past URLLC channel measurements $\{\tilde{\mathbf{h}}_{0,1} \ \tilde{\mathbf{h}}_{0,2} \ldots \tilde{\mathbf{h}}_{0,L}\}$, and the I-CSI of eMBB users to design all precoders.

III. HISTORY-BASED BEMFORMING DESIGN

The main difficulty in solving $\mathbf{P1}$ lies in efficiently addressing the constraint (3b) given a limited number L of past channel samples of the URLLC link. To address this, herein, we apply the Chernoff bound to reformulate (3b) as

$$\mu = e^{r\gamma_0^{tar}} \mathbb{E}\{e^{-r\gamma_0(\{\mathbf{w}_k\})}\} \leq \xi, \quad \forall r > 0, \tag{4}$$

where the expectation $\mathbb{E}\{e^{-r\gamma_0(\{\mathbf{w}_k\})}\}$ is taken over $\mathbf{h}_0$, and r is an auxiliary variable. Since a set of L channel realizations is available, the expectation could be approximated to the sample mean as

$$\hat{\mu} = \frac{1}{L} \sum_{j=1}^{L} e^{r(\gamma_0^{tar} - \gamma_{0,j}(\{\mathbf{w}_k\}))}, \tag{5}$$

where $\gamma_{0,j}$ is given by

$$\gamma_{0,j}(\{\mathbf{w}_k\}) = \frac{|\tilde{\mathbf{h}}_{0,j}^H \mathbf{w}_k|^2}{\sum_{i \neq k} |\tilde{\mathbf{h}}_{0,j}^H \mathbf{w}_i|^2 + \sigma^2}. \tag{6}$$

Note that $\hat{\mu} \to \mu$ holds only when $L \to \infty$ due to the law of large numbers. Therefore, since the value of $\hat{\mu}$ in (5) might significantly deviate from the population mean for a limited number L of channel measurements, we proceed as follows.

Observe that the population mean can be bounded with $100 \times \alpha\%$ confidence using percentiles over the distribution of the sample mean as follows

$$\text{P}\{\mu_{UB} \geq \mu\} = \alpha$$

$$\text{P}\left\{(\hat{\mu} - \mu_{UB})\frac{\sqrt{L}}{\hat{s}} \leq (\hat{\mu} - \mu)\frac{\sqrt{L}}{\hat{s}}\right\} = \alpha$$

$$\text{P}\left\{(\hat{\mu} - \mu_{UB})\frac{\sqrt{L}}{\hat{s}} \leq \varphi\right\} = \alpha$$

$$F_\varphi\left((\hat{\mu} - \mu_{UB})\frac{\sqrt{L}}{\hat{s}}\right) = 1 - \alpha$$

$$\mu_{UB} \triangleq \hat{\mu} - \frac{\hat{s}}{\sqrt{L}} F_\varphi^{-1}(1 - \alpha), \tag{7}$$

where μ_{UB} is an upper bound of the population mean, $\hat{s}$ is the sample standard deviation, $F_{\varphi}(\cdot)$ depicts the cumulative distribution function of $\varphi = (\hat{\mu} - \mu)\frac{\sqrt{L}}{\hat{s}}$, and $F_{\varphi}^{-1}(\cdot)$ its inverse. The sample standard deviation is given by

$$\hat{s} = \sqrt{\frac{1}{L-1}\sum_{j=1}^{L}\left(e^{r(\gamma_0^{tar} - \gamma_{0,j}(\{\mathbf{w}_k\}))} - \hat{\mu}\right)^2}. \tag{8}$$

The sample mean $\hat{\mu}$ tends to follow a normal distribution around μ as L increases. However, since the population standard deviation is unknown we must rely on $\hat{s}$. In such case, the Student's t distribution, which converges to a normal distribution when the number of samples goes to infinity, must be used. Notice that this distribution turns out to be more useful than the normal distribution for smaller number of samples due to its heavier tail. Then, the problem **P1** is rewritten as

$$\textbf{P2}: \minimize_{\{\mathbf{w}_i\}\forall i, r}\ \sum_{i=0}^{K}||\mathbf{w}_i||_2^2 \tag{9a}$$

$$\text{subject to}\ \mu_{UB} \leq \xi, \tag{9b}$$

$$\gamma_k(\{\mathbf{w}_k\}) > \gamma_k^{tar}, \quad k = 1, 2, \ldots, K, \tag{9c}$$

$$r > 0. \tag{9d}$$

Note that **P2** is still difficult to solve analytically, while common solvers (*e.g.*, genetic algorithm (GA), particle swarm optimization (PSO)) do not often provide feasible solutions. This comes from the high non-linearity of (9b), and the difficulty to configure appropriately the solvers, which often requires lots of computational resources. Alternatively, we propose taking advantage of existing solutions in the literature related to problem **P1** that do not consider the reliability constraint. Next, we provide specific details, which includes a heuristic for solving **P2**.

A. Proposed Algorithm

The proposed algorithm leverages existing I-CSI -based precoding schemes (such as those in [23] and [24]), but here, they are fed with random channel vectors. The random channel vectors are generated using statistics obtained from the channel history of the URLLC link, specifically, the sample mean $\bar{\mathbf{m}}$ and sample covariance matrix $\mathbf{C}$, which are given by

$$\bar{\mathbf{m}} = \frac{1}{L}\sum_{j=1}^{L}\tilde{\mathbf{h}}_{0,j}, \tag{10}$$

$$\mathbf{C} = \frac{1}{L-1}\sum_{j=1}^{L}(\tilde{\mathbf{h}}_{0,j} - \bar{\mathbf{m}})(\tilde{\mathbf{h}}_{0,j} - \bar{\mathbf{m}})^H. \tag{11}$$

With these statistics, we generate ζ random channel vectors such that

$$\mathbf{h}_{0,t} \sim \mathcal{CN}(\bar{\mathbf{m}}, \mathbf{C}), \quad \text{for } t = 1, 2, \ldots, \zeta. \tag{12}$$

Then, for each generated vector, we define a channel matrix

$$\mathbf{H}_t = [\mathbf{h}_{0,t}\ \mathbf{h}_1 \ldots \mathbf{h}_K], \tag{13}$$

where the last K columns correspond to the I-CSI of the eMBB users. To compute the individual precoders $\mathbf{w}_{k,t}$,

we first determine separately the normalized precoders and their associated power $p_{k,t}$ such that $\mathbf{w}_{k,t} = \sqrt{p_{k,t}}\mathbf{u}_{k,t}$.

The $t-$th matrix of normalized precoders has the structure

$$\mathbf{U}_t = [\mathbf{u}_{0,t}\ \mathbf{u}_{1,t}\ldots\mathbf{u}_{K,t}], \tag{14}$$

and can be obtained using standard I-CSI-based precoders as illustrated in Section III-B. Meanwhile, the power for each user is computed after randomly assigning a target SINR to the URLLC user as described next. In this specific case, the URLLC target SINR is drawn from a uniform distribution, *i.e.*,

$$\gamma_{0,t}^{tar} \sim \mathcal{U}(\gamma_{0,min}, \gamma_{0,max}), \tag{15}$$

where $\gamma_{0,min} = 2^{r_0} - 1$ with r_0 as the spectral efficiency in bps/Hz, and $\gamma_{0,max}$ depicts the maximum theoretically achievable SINR, which corresponds to that obtained in an interference-free setup where all the power is allocated to the URLLC link and the BS uses MRT precoding. Then, the system of equations

$$\begin{bmatrix} p_{0,t}|\mathbf{h}_{0,t}^H\mathbf{u}_{0,t}|^2 - \gamma_{0,t}^{tar}\sum_{i\neq 0}p_{i,t}|\mathbf{h}_{0,t}^H\mathbf{u}_{i,t}|^2 \\ p_{1,t}|\mathbf{h}_1^H\mathbf{u}_{1,t}|^2 - \gamma_1^{tar}\sum_{i\neq 1}p_{i,t}|\mathbf{h}_1^H\mathbf{u}_{i,t}|^2 \\ \vdots \\ p_{K,t}|\mathbf{h}_K^H\mathbf{u}_{K,t}|^2 - \gamma_K^{tar}\sum_{i\neq K}p_{i,t}|\mathbf{h}_K^H\mathbf{u}_{i,t}|^2 \end{bmatrix} = \begin{bmatrix} \gamma_{0,t}^{tar}\sigma^2 \\ \gamma_1^{tar}\sigma^2 \\ \vdots \\ \gamma_K^{tar}\sigma^2 \end{bmatrix} \tag{16}$$

must hold in order to satisfy the SINR requirements of all the users. Hence, the power of each user can be obtained from solving (16). After this step, we must check that the power allocation does not exceed the maximum available power p_{max}. If the power allocation is infeasible, *i.e.*, $\sum_{k=0}^{K}p_{k,t} > p_{max}$, we draw a new URLLC SINR target sample according to (15), compute the users' power allocation that satisfies (16), and repeat this process until the power constraint is fulfilled

After the power allocation, the $t-$th precoding matrix is formed as

$$\mathbf{W}_t = [\sqrt{p_{0,t}}\mathbf{u}_{0,t}\ \sqrt{p_{1,t}}\mathbf{u}_{1,t}\cdots\sqrt{p_{K,t}}\mathbf{u}_{K,t}]. \tag{17}$$

Then, the column vectors of matrix $\mathbf{W}_t$ are substituted into (2) and the SINR is computed for each of the L past channel measurements of the URLLC user. Next, we compute μ_{UB} according to (7). Finally, the precoding matrix that minimizes the transmit power while satisfying constraint (9b) constitutes the solution. Mathematically, the optimum index t is obtained by solving

$$\textbf{P3}: \ t^{opt} = \arg\min_{t\in[1,\zeta]}||\mathbf{W}_t||_F^2, \tag{18a}$$

$$\text{subject to}\quad \mu_{UB}(\mathbf{W}_t) \leq \xi. \tag{18b}$$

Observe that **P3** is always feasible given sufficiently large values of ζ and r. To prove it, we must consider the limit case when $r \to \infty$ and $\zeta \to \infty$. Under these conditions, there will always be at least a vector $\mathbf{h}_{0,t}$ from the set ζ whose associated precoders $\mathbf{w}_k$ ensure that $\gamma_{0,j}(\mathbf{w}_k) \geq \gamma_0^{tar}, \ \forall j$ since the image of $\gamma_{0,j}$ is the entire non-negative real domain. Now, from (7) and (18b), we have

$$\mu_{UB} = \hat{\mu} - \frac{\hat{s}}{\sqrt{L}}F_{\varphi}^{-1}(1 - \alpha) \leq \xi, \tag{19}$$

Algorithm 1 Multi-Antenna Precoding for URLLC and eMBB Coexistence

---

    Inputs: r, $\{\tilde{\mathbf{h}}_{0,j}\}$, $\{\mathbf{h}_k\}$
    Outputs: $\{\mathbf{w}_k\}$
1: $p_T \leftarrow p_{max}$
2: Compute $\bar{\mathbf{m}}$ and $\mathbf{C}$ according to (10), (11)
3: **for** $t = 1$ to ζ **do**
4:     Generate $\mathbf{h}_{0,t}$ with (12)
5:     Compute $\mathbf{H}_t$ and $\mathbf{U}_t$ according to (13), (14)
6:     Draw $\gamma_{0,t}^{tar}$ according to (15)
7:     Compute $p_{k,t}$ according to (16)
8:     **if** $\sum_{k=0}^{K} p_{k,t} \leq p_{max}$ **then**
9:         compute $\mathbf{W}_t$ according to (17)
10:       compute μ_{UB} according to (7)
11:       **if** $\mu_{UB} \leq \xi$ and $||\mathbf{W}_t||_F^2 < p_T$ **then**
12:         $\mathbf{W}^{opt} \leftarrow \mathbf{W}_t$
13:         $p_T \leftarrow ||\mathbf{W}_t||_F^2$
14:       **end if**
15:     **else**
16:       go to step 6
17:     **end if**
18: **end for**

---

with $0 \leq \xi \leq 1$. Now, computing the limits

$$\lim_{r \to \infty} \hat{\mu} = 0, \quad \lim_{r \to \infty} \hat{s} = 0, \quad \lim_{r \to \infty} \varphi = -\mu. \quad (20)$$

Therefore, $F_\varphi^{-1}(1 - \alpha)$ converges to $1 - F_\mu^{-1}(1 - \alpha)$, and μ_{UB} to zero. Then, (19) becomes true independently of the reliability requirement ξ. This implies that for large values of ζ and r, there will be at least one solution $\mathbf{W}_t$ for **P3**. It is worth noting that there is a trade-off on the selection of ζ. On the one hand, very large values of ζ imply large processing times, which is not suitable for practical systems. On the other hand, very small values of ζ may not guarantee to find a feasible solution for the problem.

Notice that the variable r controls the tightness of the Chernoff bound. As previously stated, $\gamma_{0,j}(\{\mathbf{w}_k\}) > \gamma_0^{tar}$ must be usually satisfied, thus, if very large values of r are used, the empirical distribution of $e^{r(\gamma_0^{tar} - \gamma_{0,j}(\{\mathbf{w}_k\}))}$ gets farther from a Gaussian distribution, therefore, slowing the convergence of $\hat{\mu}$ to μ. Consequently, this may also lead to the selection of precoders that do not ensure the reliability target in practice. Therefore, the value of r must be controlled to guarantee the expected results. Relatively small values, *e.g.*, $r < 10$, are suitable as discussed in Section IV.

Noteworthy, the precoding solutions obtained with our proposal ensure the reliability level as long as the number of channel measurements is large enough to compute the sample mean of the constraint in (9b) with confidence of $100 \times \alpha\%$.

Algorithm 1 encloses the required steps for the solution of the initial problem given a certain r. The complexity of the algorithm is mainly dominated by the number of iterations and the solution of the system of equations (steps 3 and 7). The former increases the complexity by ζ, while the latter by $(K + 1)^3$, leading to a total complexity of $O(\zeta(K + 1)^3)$.

B. Precoding Methods

Herein, we consider two precoding schemes, ZF, and TPM,[1] and illustrate how to obtain the normalized precoders (14) in such cases. Moreover, in case of ZF, we show how the procedure can be significantly simplified.

1) ZF: Under ZF, the links become noise-limited since the interference term is removed. The normalized ZF precoding vector is given by $\mathbf{u}_k = \mathbf{z}_k/||\mathbf{z}_k||$ with

$$[\mathbf{z}_1 \ldots \mathbf{z}_K] = \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}, \quad (21)$$

where $\mathbf{H} = [\mathbf{h}_1 \ldots \mathbf{h}_K]$ depicts a matrix containing all instantaneous channel column vectors from the BS to all users. Notice that ZF requires the computation of the pseudo-inverse of a $K \times K$ matrix which might be computationally costly. Fortunately, the computation complexity decreases as M grows as in massive MIMO systems since the term $(\mathbf{H}^H\mathbf{H})^{-1}/M$ converges to the identity matrix. In such asymptotic regime, the expression in (21) becomes $[\mathbf{z}_1 \ldots \mathbf{z}_K] = \mathbf{H}$, which matches the MRT precoding [24].

Back to the proposed algorithm, for each $\mathbf{H}_t$ (13), one evaluates expression (14) as

$$\mathbf{U}_t = \left[\frac{\mathbf{z}_{0,t}}{||\mathbf{z}_{0,t}||} \frac{\mathbf{z}_{1,t}}{||\mathbf{z}_{1,t}||} \cdots \frac{\mathbf{z}_{K,t}}{||\mathbf{z}_{K,t}||} \right], \quad (22)$$

where

$$[\mathbf{z}_{0,t} \ \mathbf{z}_{1,t} \ldots \mathbf{z}_{K,t}] = \mathbf{H}_t(\mathbf{H}_t^H\mathbf{H}_t)^{-1}. \quad (23)$$

With ZF, the SINR expression in (2) reduces to

$$\gamma_k(\{\mathbf{w}_{k,t}\}) = \frac{|\mathbf{h}_k^H\mathbf{w}_{k,t}|^2}{\sigma^2} = \frac{p_{k,t}}{||\mathbf{z}_{k,t}||^2\sigma^2}. \quad (24)$$

Therefore, to determine the power allocation required to find (17), we perform (15) and isolate the powers $p_{k,t}$ from (24) after accordingly replacing $\gamma_k(\{\mathbf{w}_{k,t}\})$ by $\gamma_{0,t}^{tar}$ and γ_k^{tar}. Notice that the above SINR expression only contains one variable $p_{k,t}$, thus, solving the system of equations in (16) can be avoided. Finally, one must proceed to determine $\mathbf{W}_t$ in (17) and μ_{UB} according to (7) to then choose the precoder with the minimum allocated power, *i.e.,* the solution of **P3**.

2) TPM: The normalized TPM precoding vector is given by [23]

$$\mathbf{u}_k = \frac{\left(\mathbf{I}_M + \sum_{i=0}^{K} \frac{\lambda_i}{\sigma^2}\mathbf{h}_i\mathbf{h}_i^H\right)^{-1}\mathbf{h}_k}{\left\|\left(\mathbf{I}_M + \sum_{i=0}^{K} \frac{\lambda_i}{\sigma^2}\mathbf{h}_i\mathbf{h}_i^H\right)^{-1}\mathbf{h}_k\right\|}, \quad (25)$$

where $\mathbf{I}_M$ depicts the identity matrix and $\{\lambda_i\}$ represent the Lagrange multipliers used to solve the original problem in **P1** without the URLLC constraint. The latter can be computed from fixed-point equations as

$$\lambda_k = \frac{\sigma^2}{\left(1 + \frac{1}{\gamma_k}\right)\mathbf{h}_k^H\left(\mathbf{I}_M + \sum_{i=0}^{K} \frac{\lambda_i}{\sigma^2}\mathbf{h}_i\mathbf{h}_i^H\right)^{-1}\mathbf{h}_k}. \quad (26)$$

---

[1]This precoder matches the structure of the optimal receive beamforming in the UL channel, *i.e.,* MMSE, if we equate the parameters $\{\lambda_k\}$ to the UL transmit powers [23].

<div style="text-align:center">

TABLE II

SIMULATION PARAMETERS

</div>

Parameter	Value	Parameter	Value
M	$\{8, 16\}$	K	4
δ	3.5	p_{max}	47 dBm
d_r	500 m	γ_0^{tar}	−11.44 dB
κ_0	$\{0, 2, 5, 10\}$	$\gamma_k^{tar}(k \neq 0)$	$\{0, 10\}$ dB
r	10	ξ	$\{10^{-3}, 10^{-4}\}$
L	250, 500, 3500	α	0.99
$\kappa_k(k \neq 0)$	$\{0, 2\}$	ζ	3000

Given the above precoding structure, we can perform (10)-(13), and then use (26)-(25) to compute $\mathbf{U}_t$. The next step is the random SINR assignment (15). Notice that with TPM, the interference term is not perfectly removed as in ZF. Therefore, to compute the power allocation $p_{k,t}$, one must unavoidably find the solution of the system of equations in (16). Again, to conclude, one must compute the precoding matrix $\mathbf{W}_t$ given that $\mathbf{w}_t = \sqrt{p_{k,t}}\mathbf{u}_k$, determine μ_{UB} according to (7), and find the solution to **P3**.

C. Practicalities

To reduce the processing delay, the computations related to the proposed algorithm may be executed in instances where only the eMBB users are receiving data. Notice that during this time, the BS will not use the precoders obtained from the proposed algorithm for current DL transmissions, but the ones computed in parallel using predefined precoding schemes. In general, the algorithm should run every time the I-CSI of eMBB users is updated. Meanwhile, to reduce the hardware resource utilization, it might not be necessary to run the algorithm every time a new URLLC channel measurement is obtained, since the statistics will not be considerably modified.

The proposed algorithm also considers a single URLLC transmission per resource block. This is because concurrent URLLC transmissions would cause a strong mutual interference due to the use of imperfect statistics on the precoder design. Notice that we refer here to URLLC users with very tight latency requirements that do not allow I-CSI acquisition. Other URLLC services with not that stringent latency demands might be treated in the same way as eMBB users.

Finally, the storage space required to save the channel measurement will be upper bounded by $\sum_{\forall u} 10/\xi_u \times B \times M$ with ξ_u and B as the user-specific outage target and the number of bits required for quantization, respectively. For instance, for 10 URLLC users, $\xi_u = 10^{-3}$, $B = 8$, and $M = 8$, the maximum storage space that would be required is 6400000 bits (0.8 MB). Notice that a finite quantization level may affect the performance in practice, which could be considered in future works.

IV. NUMERICAL RESULTS

We consider that the BS serves one URLLC and four eMBB users within a resource block. The users are uniformly deployed in an area of radius d_r around the BS. The distance between the BS and user k is denoted as d_k, while the path gain experienced by the latter is given by $\psi_k = d_k^{-\delta}$, where δ

depicts the path loss exponent. Moreover, we use the Rician fading model due to its potential to cover different scenarios by properly tuning the parameter κ_k, from non LOS (NLOS) setups as in Rayleigh fading ($\kappa_k = 0$) to fully deterministic LOS scenarios ($\kappa_k \rightarrow \infty$). Specifically, the Rician channel model is given by [25]

$$\mathbf{h}_k = \sqrt{\psi_k}\left(\sqrt{\frac{\kappa_k}{\kappa_k + 1}}\mathbf{h}_{k,LOS} + \sqrt{\frac{1}{\kappa_k + 1}}\mathbf{h}_{k,NLOS}\right), \quad (27)$$

where $\psi_k\sqrt{\kappa_k/(\kappa_k + 1)}\mathbf{h}_{k,LOS}$ represents the deterministic LOS propagation component, and $\psi_k\mathbf{h}_{k,NLOS}/\sqrt{\kappa_k + 1}$ represents the scattering component with $\mathbf{h}_{k,NLOS} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. For simplicity, we set $\mathbf{h}_{k,LOS}$ as a vector of ones and assume the same γ_k^{tar} and LOS factor κ_k for all the eMMB users. The remaining simulation parameters are shown in Table II. Notice that the noise power corresponds to a bandwidth of 10 MHz, while γ_0^{tar} comes from assuming a packet of 32 bytes transmitted over 0.256 ms and 10 MHz, i.e., $\gamma_0^{tar} = 2^{32\times 8 \text{bits}/(0.256\times 10^{-3}\text{s}\times 10^7 \text{Hz})} - 1 = 0.0718 = -11.44$ dB.

In Sections IV-A, IV-B and IV-C, we evaluate the performance of the proposed algorithm for an instantaneous network realization, including a given URLLC channel history, network deployment, and I-CSI of the eMBB users.[2] Meanwhile, we present statistics obtained over 5×10^3 network realizations in Section IV-D.

A. On the Configuration of r and ζ

Fig. 2 shows the impact of the configuration of the parameter r on the attainable outage probability (Fig. 2 (a)) and transmit power (Fig. 2(b)) for the proposed algorithm with both ZF- and TPM-based precoding. Herein, we set the outage target to 10^{-3}, and assume $L = 500$ channel measurements. Notice that the outage probability decreases as r decreases, but at the cost of attaining higher-power precoders. Meanwhile, adopting relatively large values of r leads to reduced transmit power precoders that may not satisfy the outage constraint. Interestingly, ZF outperforms TPM in terms of transmit power regardless of the value of r. Noteworthy, the TPM precoding is the optimal for transmit power minimization only with I-CSI availability. Nevertheless, the increment of κ_0 brings a reduction in the performance gap between both precoding methods since the channel becomes more deterministic and the randomly generated coefficients are closer to the actual channel realizations. Notice that the increment of κ_0 also eases the selection of r. In the following, we set $r = 10$, which allows reliably meeting the outage constraint for both ZF and TPM-based precoding mechanisms as illustrated in Fig. 2.

Fig. 3 shows the impact of the number of random generated vectors on the total transmit power for both precoding schemes. Observe that as ζ increases, the obtained precoder gets probabilistically closer to the optimum one. Interestingly, the transmit power reduction is not significant for $\zeta > 3000$ in most configurations. Therefore, adopting higher values is not convenient as it increases the processing times. Observe for

---

[2]Given a certain seed for the generation of random numbers, we obtained a single network realization. We repeatedly tested many seeds and verified that the performance trends remain similar.
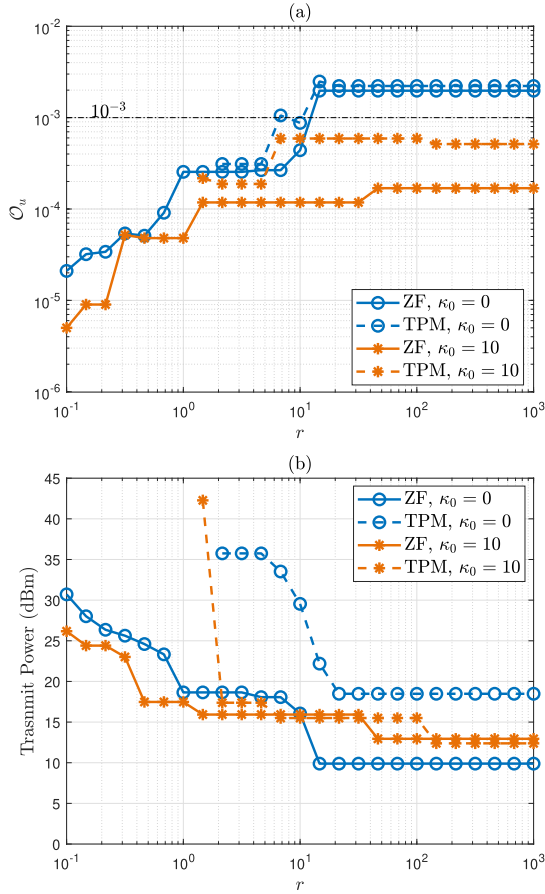
Fig. 2. (a) Outage probability, and (b) total transmit, power as a function of r. The URLLC channel is subject to Rayleigh fading ($\kappa_0 = 0$) and Rician fading ($\kappa_0 = 10$), while $\xi = 10^{-3}$, $\gamma_k^{tar} = 10$ dB $\forall k \neq 0$, and $L = 500$ channel measurements.
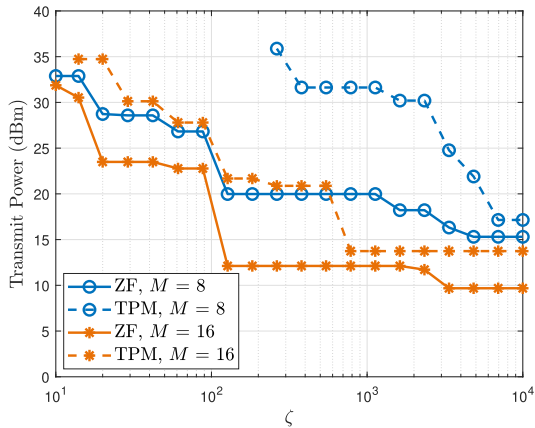

Fig. 3. Total transmit power as a function of ζ. The URLLC channel is subject to Rayleigh fading ($\kappa_0 = 0$), while $\xi = 10^{-3}$, $\gamma_k^{tar} = 10$ dB $\forall k \neq 0$, and $L = 500$ channel measurements.
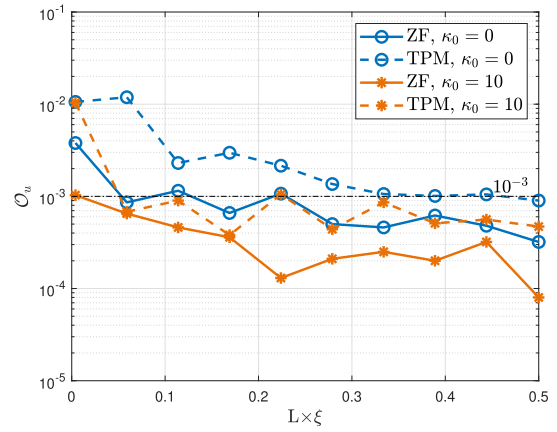

Fig. 4. Outage probability vs L$\times\xi$ for Rayleigh ($\kappa_0 = 0$) and Rician ($\kappa_0 = 10$) fading scenarios for ZF and TPM with $\gamma_k^{tar} = 10$ dB $\forall k \neq 0$.
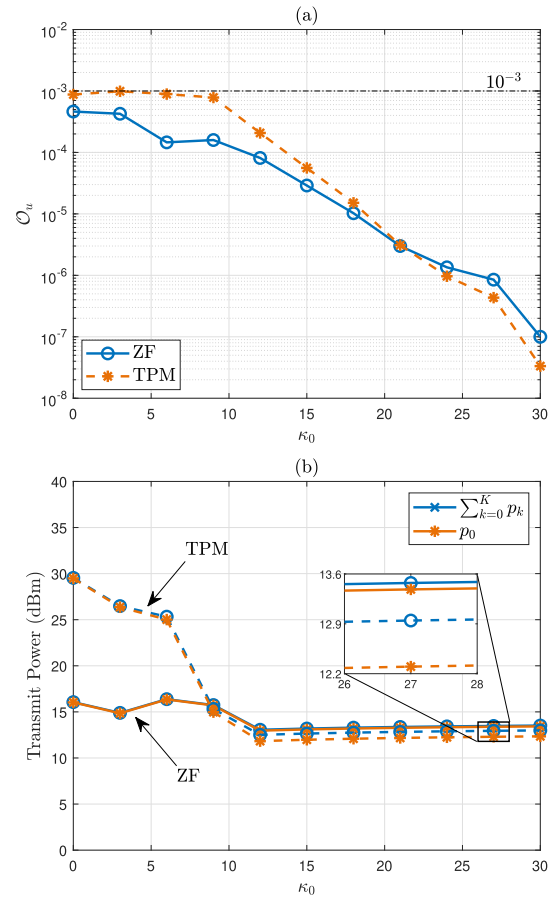

Fig. 5. (a) Outage probability, and (b) total transmit power and transmit power for URLLC as a function of κ_0. We set $\gamma_k^{tar} = 10$ dB $\forall k \neq 0$, $\xi = 10^{-3}$, and consider $L = 500$ channel measurements.

B. On the Performance Impact of the Number of Measurements and URLLC LOS Channel Factor

Fig. 4 shows the minimum number of channel measurements required to keep the outage target below the threshold ξ. For ZF, around 250 past channel measurements are enough to achieve the target $\xi = 10^{-3}$ when $\kappa_0 = 0$, i.e., Rayleigh fading, while the number drops below 50 when

a relatively small ζ, the problem may not be feasible, as it occurs for $\zeta < 280$ for TPM and $\zeta < 12$ for ZF, with $M = 8$. However, the problem becomes feasible as ζ gets relatively larger, which confirms the statement in Section III-A about the feasibility of the problem as $\zeta \to \infty$.

$\kappa_0 = 10$. However, these figures considerably increase for TPM, being around 470 and 120, respectively. The reduction of the required number of samples as κ_0 increases is because the generated channel vectors get more concentrated around the actual (more deterministic) channel realizations.

Fig. 5 (a) displays the achievable outage probabilities for different values of κ_0. It is worth noting that the outage probabilities even go below 10^{-5} as κ_0 increases. This guarantees high reliability levels, even if the outage target is more stringent, in scenarios where the BS and users have a relatively strong LOS component. Meanwhile, Fig. 5 (b) shows the behavior of the transmit power for different values of the LOS component. The increment of κ_0 leads to lower transmit powers which converge due to the fixed number of random generations ζ. The figure also depicts the fact that most of the power is required for the URRLC service, evincing the significant costs of achieving high reliability. Also notice that ZF outperforms TPM when $\kappa_0 \leq 8$, while TPM performs better as κ_0 increases since the generated coefficients are closer to the actual channel realizations of the URLLC user.

In the following, we focus only on the performance of ZF. This is for simplicity and given ZF outperforms TPM precoding in poor channel conditions. Moreover, we consider $L = 250$ channel measurements for which, given $r = 10$, the URLLC constraint is already met as shown in Fig. 4.

C. On the Performance Impact of the Number of eMBB Users and Their SINR Target

Fig. 6 shows the behavior of the outage probability (Fig. 6 (a)) and transmit power (Fig. 6 (b)) for different SINR targets $\gamma_k^{tar} \; \forall k \neq 0$. Notice that the outage probabilities tend to increase with the SINR since larger eMBB transmission powers cause larger interference levels to the URRLC link. Meanwhile, higher SINR requirements of the eMMB users lead to more power allocated to them, therefore, increasing the total transmit power. Moreover, there is an increment on the transmit power intended to the URRLC user as tighter reliability targets are set. For instance, the required transmit power at $\gamma_k^{tar} = 10$ dB ($k \neq 0$) is approximately 17 dBm and 23 dBm, for $\xi = 10^{-3}$ and $\xi = 10^{-4}$, respectively. Notice that we are considering the worst possible case (Rayleigh fading), where the powers requirements are higher due to the lack of a LOS component.

Fig. 7 (a) shows the achieved outage probability as a function of the number of eMBB devices that are simultaneously served within a resource block for $\xi = 10^{-3}$ and $\xi = 10^{-4}$. Note that the outage probability tends to increase with the number of devices, since the precoding needs to cope with larger interference levels. Therefore, higher transmit powers are required to achieve the targeted outage probabilities in the URLLC link, which is depicted in Fig. 7 (b).

D. On the Statistics of the Achievable Outage Probability and Allocated Transmit Power

Different from the previous results, herein we obtain statistics for 5×10^3 randomly generated network realizations, i.e.,
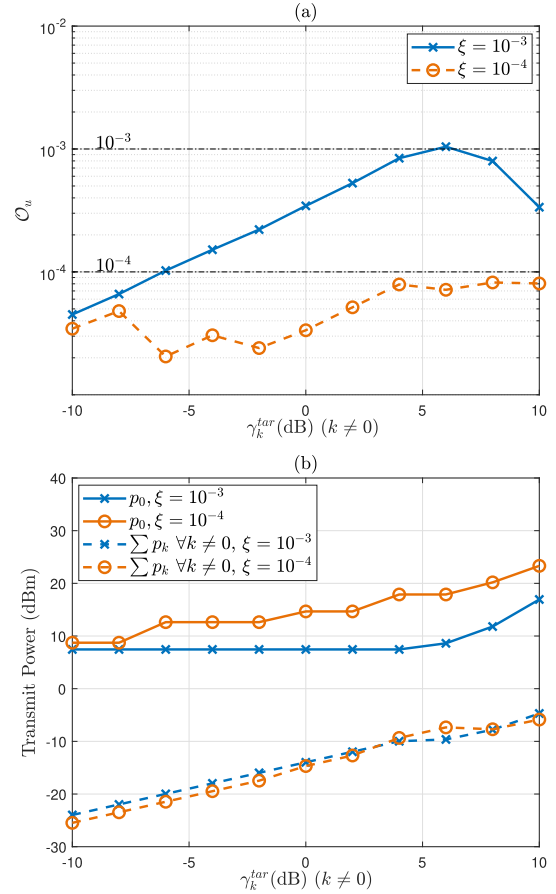

Fig. 6. (a) Outage probability, and (b) transmit power of both, eMBB and URLLC, DL transmissions, as a function of γ_k^{tar} ($k \neq 0$). URLLC channel is subject to Rayleigh fading ($\kappa_0 = 0$). We set $L = 250, 3500$ channel measurements for $\xi = 10^{-3}, 10^{-4}$, respectively.

different network deployments, channel history, and I-CSI of URLLC and eMBB users, respectively.

Fig. 8 shows the empirical probability density function (PDF) of the outage probability exponent, i.e., $\log_{10} \mathcal{O}_u$, for ZF with a target $\xi = 10^{-3}$. Notice that the histograms for $L = 250$ (Fig. 8 (a)) and $L = 500$ (Fig. 8 (b)) approximately match a Gaussian PDF, whose parameters are obtained by standard curve fitting and are displayed in Table III. Here, the confidence value for $\xi \leq 10^{-3}$ is obtained as

$$CV = \left(1 - Q\left[\frac{\xi - MV}{SD}\right]\right) \times 100, \qquad (28)$$

where MV depicts the estimated mean, and SD the estimated standard deviation. In ZF precoding, we obtain $CV \approx 89.01\%$ and $CV \approx 94.96\%$ with $M = 8$ and $M = 16$, respectively, and exploiting $L = 250$. Note that the use of more antennas moves the mean of the distribution to the left due to the diversity gain. The improvement is, however, small because the algorithm reduces the transmit power while pushing $\mathcal{O}_u$ close to the target. It is worth highlighting that the distributions can also move to the left in scenarios with larger κ_0, and/or by exploiting more URLLC past channel measurements, e.g., $L = 500$. Indeed, the chances of exceeding ξ decrease considerably with the increment of L. Specifically, for
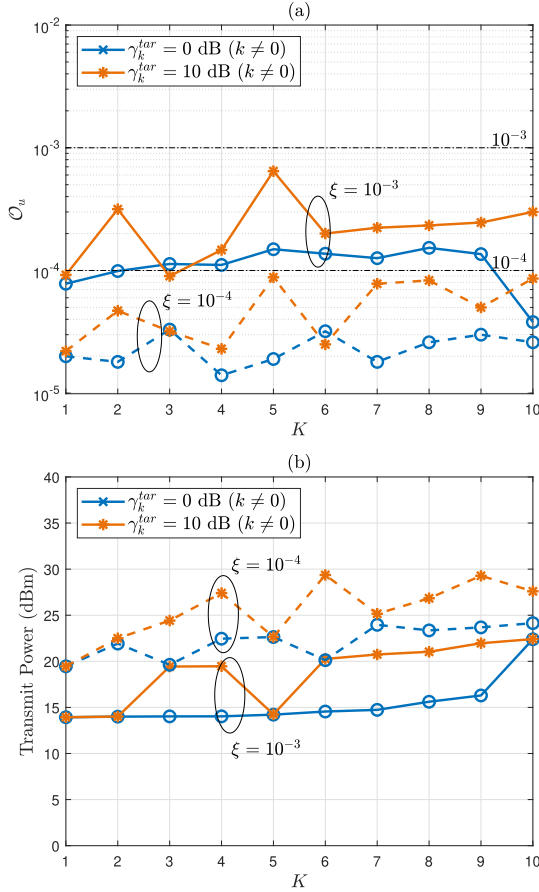
Fig. 7. (a) Outage probability, and (b) total transmit power, as a function of the number of simultaneously served eMBB devices. The URLLC channel is subject to Rayleigh fading ($\kappa_0 = 0$). We set $M = 16$, and $L = 250, 3500$ channel measurements for $\xi = 10^{-3}, 10^{-4}$, respectively.
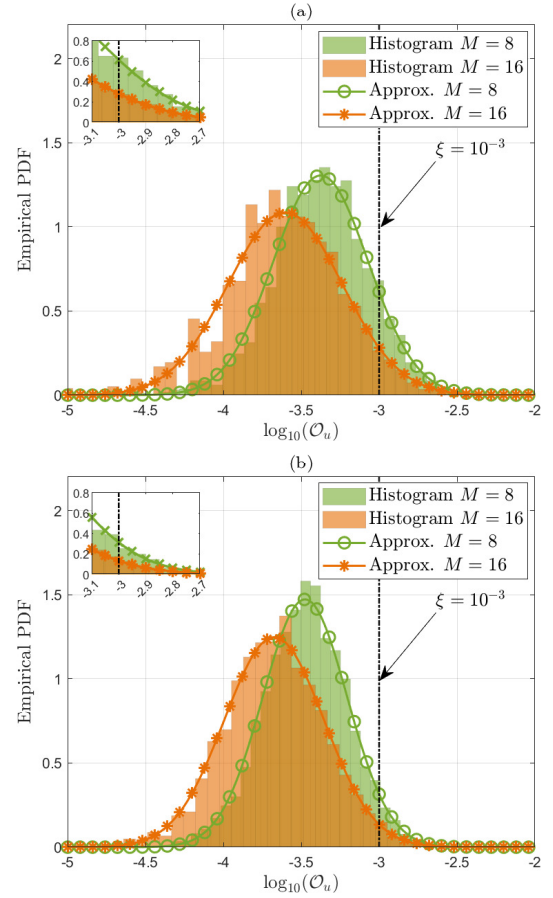

Fig. 8. Empirical PDF of the outage probability and approximation to a Gaussian distribution for 5000 network realizations with (a) $L = 250$, and (b) $L = 500$ channel measurements. All users are subject to Rayleigh fading ($\kappa_k = 0 \ \forall k$). We set $\gamma_k^{tar} = 10$ dB $\forall k \neq 0$, and $\xi = 10^{-3}$.

$L = 500$, the confidence levels increase up to $CV = 96.05\%$ and $CV = 98.25\%$ for $M = 8$ and $M = 16$, respectively. Notice that the estimated values are close to the ones obtained with Monte Carlo (MC) simulations which means that a Gaussian distribution is a good approximation. The confidence levels can also be increased at the cost of incurring in higher transmit powers, *i.e.*, smaller ζ or r. Meanwhile, TPM again exhibits a poor performance in Rayleigh fading, but it considerably improves as the value of κ_0 gets larger. The use of more more past channel measurements would strongly improve the CV for this precoding method. Also note that the confidence levels for both precoding methods increase and even reach values above 99% as the LOS components get stronger.

Fig. 9 shows the (approximately Gaussian) empirical PDF of the total transmit power in dBm when using $M = 8$ and $M = 16$ for $L = 500$ with ZF under Rayleigh fading. The mean transmit power is approximately 16.41 dBm and 14.52 dBm for $M = 8$ and $M = 16$, respectively. This power reduction of about 1.9 dB is the cause that the mean outage probabilities illustrated in Fig. 8 did not experience a larger reduction as previously discussed. Finally, we would like to highlight that under a similar setup with $M = 8$, TPM exhibits an even poorer performance since it requires on

TABLE III
FITTING PARAMETERS AND CONFIDENCE FOR $\log_{10} \mathcal{O}_u$

Precod.	κ_0	L	M	MV	SD	$CV(\%)$	$MC(\%)$
ZF	0	250	8	-3.375	0.305	89.01	89.50
	0	250	16	-3.603	0.367	94.96	95.46
	0	500	8	-3.476	0.271	96.05	96.26
	0	500	16	-3.674	0.319	98.25	98.46
	2	250	8	-3.929	0.604	93.81	95.60
	2	250	16	-5.358	0.840	99.75	>99.99
	2	500	8	-4.729	0.815	98.30	99.40
	2	500	16	-6.643	0.847	99.91	> 99.99
	5	250	8	-5.088	1.078	97.36	97.40
	5	250	16	-6.158	0.539	>99.99	>99.99
	5	500	8	-5.245	1.167	97.28	98.40
	5	500	16	-6.570	0.646	>99.99	>99.99
TPM	0	250	8	-2.858	0.247	28.30	26.00
	0	250	16	-2.975	0.221	45.51	43.20
	0	500	8	-2.973	0.228	45.29	42.10
	0	500	16	-3.061	0.284	58.36	57.20
	2	250	8	-3.590	0.610	83.31	82.60
	2	250	16	-5.180	0.819	99.61	99.46
	2	500	8	-3.600	0.393	93.65	95.92
	2	500	16	-5.390	0.869	99.73	99.70
	5	250	8	-5.087	1.075	97.41	97.40
	5	250	16	-6.150	0.539	>99.99	>99.99
	5	500	8	-5.257	1.138	97.63	98.40
	5	500	16	-6.569	0.646	>99.99	>99.99

average 22.14 dBm of transmit power to satisfy the reliability requirements. Nevertheless, this behavior changes in scenarios
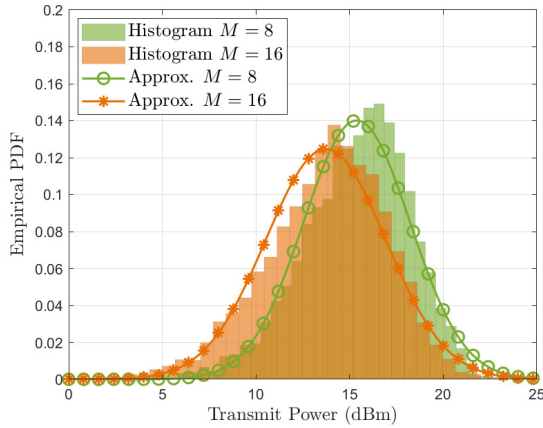
Fig. 9. Empirical PDF of the total transmit power and approximation to a Gaussian distribution for 5000 network realizations. All users are subject to Rayleigh fading ($\kappa_k = 0 \; \forall k$). We set $\gamma_k^{tar} = 10$ dB $\forall k \neq 0$, and $\xi = 10^{-3}$.

with enhanced channel conditions, *e.g.,* Rician fading with $\kappa_0 = 10$, since 12.31 dBm and 11.37 dBm of transmit power are required by ZF and TPM, respectively.

V. CONCLUSION

In this paper, we considered the I-CSI of multiple eMBB links and the channel measurement's history of one URLLC user for DL multi-antenna beamforming design. In our proposal, we leveraged the Chernoff bound to stochastically model, impose, and guarantee the reliability requirements of the URLLC user based on its channel history. Moreover, our proposed precoding design relies on properly modified I-CSI-based precoding methods. We illustrated our approach by adopting ZF and TPM precodings with per-user SINR constraints, whose performance was assessed through simulations. We showed that ZF outperforms TPM in scenarios with poor channel conditions, while TPM exhibits a better performance as the channel becomes more deterministic, *i.e.,* with greater LOS. For instance, in Rayleigh fading with 500 past URLLC measurements, eight antennas at the BS, and for an outage probability target of 10^{-3}, the mean transmit power of ZF and TPM are 16.41 dBm and 22.14 dBm, respectively. However, in Rician fading with a LOS of 10 dB, the figures drop to 12.31 dBm and 11.37 dBm, respectively. Finally, we determined the confidence levels required to achieve the target outage probabilities, which can be larger than 99% when operating in favorable LOS conditions.

REFERENCES

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

[2] J. Lorca et al., "Deliverable D2.1: Scenarios, KPIs, use cases and baseline system evaluation," E2E-Aware Optim. Adv. Netw. Edge 5G New Radio (ONE5G), Berlin, Germany, Tech. Rep. D2.1, 2017, vol. 2.

[3] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–6.

[4] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, "Radio resource allocation and retransmission schemes for URLLC over 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 896–904, Apr. 2019.

[5] O. L. A. Lopez, N. H. Mahmood, H. Alves, C. M. Lima, and M. Latva-Aho, "Ultra-low latency, low energy, and massiveness in the 6G era via efficient CSIT-limited scheme," *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 56–61, Nov. 2020.

[6] N. Bornhorst and M. Pesavento, "Beamforming for multi-group multicasting with statistical channel state information using second-order cone programming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 3237–3240.

[7] W. Tabikh, Y. Yuan-Wu, and D. Slock, "Beamforming design with combined channel estimate and covariance CSIT via random matrix theory," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–5.

[8] S. Qiu, D. Chen, D. Qu, K. Luo, and T. Jiang, "Downlink precoding with mixed statistical and imperfect instantaneous CSI for massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3028–3041, Apr. 2018.

[9] A.-S. Bana, G. Xu, E. De Carvalho, and P. Popovski, "Ultra reliable low latency communications in massive multi-antenna systems," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 188–192.

[10] A. A. Nasir, H. D. Tuan, H. H. Nguyen, M. Debbah, and H. V. Poor, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[11] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–6.

[12] Z. Wang and V. W. S. Wong, "Joint resource block allocation and beamforming with mixed-numerology for eMBB and URLLC use cases," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

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

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

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

[16] A. K. Bairagi et al., "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.

[17] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

[18] 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.

[19] N. H. Mahmood, O. A. López, H. Alves, and M. Latva-Aho, "A predictive interference management algorithm for URLLC in beyond 5G networks," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 995–999, Mar. 2021.

[20] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42–48, Mar. 2019.

[21] Q. Huang et al., "Machine-learning-based cognitive spectrum assignment for 5G URLLC applications," *IEEE Netw.*, vol. 33, no. 4, pp. 30–35, July/Aug. 2019.

[22] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Predictive ultra-reliable communication: A survival analysis perspective," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1221–1225, Apr. 2021.

[23] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[24] M. R. Khandaker and K.-K. Wong, "Signal processing for massive MIMO communications," in *Academic Press Library in Signal Processing*, vol. 7. Amsterdam, The Netherlands: Elsevier, 2018, pp. 367–401.

[25] J. R. Hampton, *Introduction to MIMO Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

Dian Echevarría Pérez (Graduate Student Member, IEEE) was born in Cienfuegos, Cuba, in 1989. He received the B.Sc. degree in telecommunications and electronic engineering from the Central University of Las Villas, Cuba, in 2013, and the M.Sc. degree (Hons.) in wireless communications engineering from the University of Oulu, Finland, in 2021. He is currently pursuing the Ph.D. degree with the Centre for Wireless Communications (CWC), Oulu, Finland. From 2013 to 2018, he worked as a Specialist in telematics at Cuban Telecommunications Company (ETECSA). In 2019, he joined CWC. His research interests are full-duplex communications, wireless RF energy transfer, and radio-resource allocation for ultra-reliable low-latency communications.

Hirley Alves (Member, IEEE) is currently an Associate Professor and the Head of the Machine-Type Wireless Communications Group, 6G Flagship, Centre for Wireless Communications, University of Oulu. He actively works on massive connectivity and ultra-reliable low latency communications for future wireless networks, 5GB and 6G, full-duplex communications, and physical-layer security. In addition, he leads the URLLC activities for the 6G Flagship Program. He has received several awards and has been the organizer, the chair, the TPC, and a tutorial lecturer for several renowned international conferences. He is the General Chair of the ISWCS'2019 and the General Co-Chair of the First 6G Summit, Levi 2019, and ISWCS 2021.

Onel L. Alcaraz López (Member, IEEE) was born in Sancti-Spíritus, Cuba, in 1989. He received the B.Sc. degree (Hons.) in electrical engineering from the Central University of Las Villas, Cuba, the M.Sc. degree in electrical engineering from the Federal University of Paraná, Brazil, in 2017, and the D.Sc. degree (Hons.) in electrical engineering from the University of Oulu, Finland, in 2020. From 2013 to 2015, he worked as a Specialist in telematics at Cuban Telecommunications Company (ETECSA). He is a Collaborator to the 2016 Research Award given by the Cuban Academy of Sciences, a co-recipient of the 2019 IEEE European Conference on Networks and Communications (EuCNC) Best Student Paper Award, a recipient of the 2020 Best Doctoral Thesis Award granted by Academic Engineers and Architects in Finland TEK and Tekniska Föreningen i Finland TFiF in 2021, and a recipient of the 2022 Young Researcher Award in the field of technology in Finland. He has authored the book entitled *Wireless RF Energy Transfer in the Massive IoT Era: Towards Sustainable Zero-Energy Networks* (Wiley, December 2021). He currently holds an Assistant Professorship (a tenure track) in sustainable wireless communications engineering with the Centre for Wireless Communications (CWC), Oulu, Finland. His research interests include wireless connectivity, sustainable and dependable IoT, energy harvesting, wireless RF energy transfer, machine-type communications, and cellular-enabled positioning systems.