Decentralized Coordinated Precoding Design in Cell-Free Massive MIMO Systems for URLLC

Enyu Shi*, Graduate Student Member, IEEE*, Jing Zhang,
Jiayi Zhang*, Senior Member, IEEE*,
Derrick Wing Kwan Ng*, Fellow, IEEE*, and Bo Ai*, Fellow, IEEE*

***Abstract*—Cell-free massive multiple-input multiple-output (MIMO) is a promising network to offer huge improvement of the achievable rate compared with conventional cellular massive MIMO systems. However, the commonly adopted Shannon-type achievable rate is only valid in the long block length regime that is not applicable to the emerging short-packet communication. To realize ultra-reliable and low-latency communication (URLLC) in cell-free massive MIMO systems, we optimize the precoding vector at the access points (APs) to maximize the minimum user rate in both the centralized and decentralized fashion. The design takes into account the impact of URLLC and we propose path-following algorithms (PFA) to address the considered problem which generates a sequence of advanced feasible points and converges to at least a locally optimal solution of the design problem. Moreover, we investigate the requirement of the precoding schemes, the length of the transmission duration, the number of antennas equipped at each AP, and the size of each AP cluster on the URLLC rate. Numerical results show that the decentralized PFA precoding can achieve 80% of the 95%-likely URLLC rate of the centralized precoding and 89% of the average URLLC rate with only 12% computational complexity of the centralized precoding.**

***Index Terms*—Cell-free massive MIMO, nonconvex optimization, precoding, URLLC.**

I. Introduction

Ultra-reliable and low-latency communication (URLLC) is one of the generic applications required to be covered in the fifth-generation (5 G) [1], [2], [3]. As a result, it has been attracted significant interests since it enables several innovative usages, especially in industrial production, such as remote heavy industrial machines operation and factory automation [4], [5]. However, compared with conventional communication systems, the achievable rate under URLLC is quite different since short blocklength is adopted to shorten the latency such that the classical Shannon-sense capacity no longer holds. Specifically, the URLLC rate is a complicated function of the transmission power, the precoding vector, the bandwidth, the transmission time, and the decoding error probability [6]. Indeed, guaranteeing URLLC represents unique challenges to resource allocation design due to the non-convexity introduced by the finite blocklength. In the literature, much attention has been devoted to designing effective resource allocation algorithms that support URLLC [7], [8], [9]. However, the systems considered in these works are all cellular networks and their performance is known to be limited by severe inter-cell interference.

Cell-free massive multiple-input multiple-output (MIMO) architecture is a new promising solution to overcome the issue discussed above [10], [11], [12], [13]. It reaps the advantages of massive MIMO and network MIMO, since massive distributed access points (APs) facilitate coherent signal transmission to serve all the users without any cell boundaries [14], [15], [16]. However, current literature focuses on the resource allocation in cell-free massive MIMO systems for URLLC is still limited. For example, in [17], the authors applied the path-following algorithm (PFA) for optimizing the power allocation with a special class of conjugate beamforming to maximize the users' minimum URLLC rate and the energy efficiency. However, an adaptive and optimized precoding design at the APs is generally more effective that the fixed one. Besides, in [18], the upper bounds of the uplink and downlink decoding error probabilities (DEPs) were derived by using the saddlepoint method to support URLLC. While the closed-form expression of DEP can characterize the performance, it is generally intractable for the the design of cooperatively efficient resource allocation. As such, there is an emerging need for designing the precoding with the performance metric of the URLLC rate.

Motivated by the above discussion, the PFA-based precoding design for maximizing the users' minimum URLLC rate is studied in this correspondence. First, a PFA-based centralized precoding design is proposed which generates a sequence of feasible points and converges to a locally optimal solution of the design optimization problem. Second, we propose a decentralized PFA-based precoding design by dividing the APs into several non-overlapping cooperative clusters in which the APs only share the data and instantaneous channel state information (CSI) in each cluster to design the precoding vectors to reduce the computational complexity. Simulation results show that compared with the centralized precoding, the decentralized PFA precoding can achieve 80% of the 95%-likely URLLC rate and 89% of the average URLLC rate with only 12% of the computational complexity of the counterpart. We also investigate the impact of the precoding schemes, the length of transmission duration, and the size of the AP cluster on the URLLC rate via extensive simulations.

II. System Model

We consider a cell-free massive MIMO system, which consists of L APs and K single-antenna users that are distributed arbitrarily over a large area. We assume that each AP is equipped with N antennas. Moreover, all the APs are connected with each other and a central processing unit (CPU) via dedicated fronthaul links with sufficient capacity. All APs serve all users on the same time-frequency resource through time division duplex (TDD) operation [19].

The channel coefficient between AP l and user k, $\mathbf{h}_{kl} \in \mathbb{C}^{N\times 1}$, is assumed to follow a correlated Rayleigh fading distribution. We adopt a classic block fading model for modeling the channels such that $\mathbf{h}_{kl}$ remains constant in t channel uses of the time-frequency blocks and experience an independent realization in every block. Note that the channel coefficients can be acquired at the APs by existing channel

Manuscript received 5 May 2022; revised 25 August 2022; accepted 23 September 2022. Date of publication 28 September 2022; date of current version 13 February 2023. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grants 2022JBZX030 and 2022JBQY004, in part by the Frontiers Science Center for Smart High-speed Railway System, in part by the Natural Science Foundation of Jiangsu Province, Major Project under Grant BK20212002, in part by ZTE Corporation, and in part by State Key Laboratory of Mobile Network and Mobile Multimedia Technology. The work of Derrick Wing Kwan Ng was supported by Australian Research Council Discovery Project under Grant DP210102169. The review of this article was coordinated by Dr. Chau Yuen. *(Corresponding author: Jiayi Zhang.)*

Enyu Shi, Jing Zhang, and Jiayi Zhang are with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China, and also with the Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University, Beijing 100044, China (e-mail: 21111047@bjtu.edu.cn; 19120167@bjtu.edu.cn; jiayizhang@bjtu.edu.cn).

Derrick Wing Kwan Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2033, Australia (e-mail: w.k.ng@unsw.edu.au).

Bo Ai is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: boai@bjtu.edu.cn).

Digital Object Identifier 10.1109/TVT.2022.3210253

estimation algorithms [20] and this is beyond the scope of this work as we aim to optimize the precoding for URLLC. Therefore, we assume that perfect CSI is available at the APs.

In the downlink payload data transmission phase, the received signal at user k can be expressed as $y_k = \sum_{l=1}^{L} \mathbf{h}_{kl}^H \mathbf{w}_{kl} s_k + \sum_{l=1}^{L} \mathbf{h}_{kl}^H \sum_{i \neq k}^{K} \mathbf{w}_{il} s_i + n_k$, where $s_i \sim \mathcal{N}_{\mathbb{C}}(0,1)$ at AP l, $\mathbf{w}_{il} \in \mathbb{C}^{N \times 1}$ is the precoding vector for user i at AP l, and $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ represents the thermal noise at user k. Then, the corresponding effective signal-to-interference-plus-noise ratio (SINR) is given as

$$\varphi_k = \frac{\left|\mathbf{h}_k^H \mathbf{w}_k\right|^2}{\sum_{i \neq k}^{K} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2}, \tag{1}$$

where $\mathbf{h}_k = [\mathbf{h}_{k1}^H, \ldots, \mathbf{h}_{kL}^H]^H \in \mathbb{C}^{LN \times 1}$ and $\mathbf{w}_i = [\mathbf{w}_{i1}^H, \ldots, \mathbf{w}_{iL}^H]^H \in \mathbb{C}^{LN \times 1}$. By treating the inter-user interference $\mathbf{h}_{kl}^H \sum_{i \neq k}^{K} \mathbf{w}_{il} s_i$ as Gaussian noise, where $p_{il}^{\mathrm{dl}} \triangleq \|\mathbf{w}_{il}\|^2$ is the power allocated to user i at AP l, the achievable rate in nats/sec/Hz for user k for the case of sufficiently long blocklength is given by the Shannon rate function $\tilde{R}_k = \ln(1 + \varphi_k)$, and the achievable URLLC rate in nats/sec/Hz for user k can be approximated as [8, eq. (30)]

$$R_k = \ln\left(1 + \varphi_k\right) - \sqrt{\frac{1}{tB} \times V_k} \times Q^{-1}\left(\epsilon\right), \tag{2}$$

where t is the transmission duration, B is the communication bandwidth, V_k is the channel dispersion [8] which can be expressed as $V_k = 1 - \frac{1}{(1+\varphi_k)^2}$, $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function, i.e., $Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt$, and ϵ is the decoding error probability. Note that (2) is the normal approximation when the channel $\mathbf{h}_k$ is assumed to be quasi-static and deterministic over the transmission duration t. The subtrahend in (2) captures the rate penalty due to the finite block length, tB.

III. Max-Min Rate Based Precoding Design

A. Centralized Precoding Design

In the centralized precoding design, the optimization of the precoding vectors takes place at the CPU, where the estimate of the global instantaneous CSI $\mathbf{h}_{kl}, \forall k \in \{1, \ldots, K\}, \forall l \in \{1, \ldots, L\}$, available.

The centralized max-min URLLC rate optimization problem can be expressed as

$$\max_{\mathbf{w}} \min_{k=1,\ldots,K} \left\{R_k\left(\mathbf{w}\right)\right\} \tag{3a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \left\|\mathbf{w}_{kl}\right\|^2 \leq p_{\max}, \forall l, \tag{3b}$$

where $\mathbf{w} = \{\mathbf{w}_{kl} : k = 1, \ldots, K, l = 1, \ldots, L\}$ and $p_{\max}$ is the maximum power at each AP. The problem (3a) is non-convex due to the URLLC rate function $R_k(\mathbf{w})$. With the help of [8], we apply the PFA to develop a concave lower bound for $R_k(\mathbf{w})$.

Without loss of generality, the URLLC rate expression for user k can be rewritten as $R_k(\mathbf{w}) = f_k(\mathbf{w}) - a g_k(\mathbf{w})$, where $a = Q^{-1}(\epsilon)/\sqrt{tB}$, $f_k(\mathbf{w}) = \ln(1 + \varphi_k(\mathbf{w}))$, and $g_k(\mathbf{w}) = \sqrt{1 - 1/(1 + \varphi_k(\mathbf{w}))^2}$. Now, we aim to establish a convex lower bound for $f_k(\mathbf{w})$ and a concave upper bound for $g_k(\mathbf{w})$.

Let $\mathbf{w}^{(n)}$ be a feasible point for (3a) that is computed from the $(n-1)$th iteration of the iterative PFA.

1) Lower Bounding for $f_k(\mathbf{W})$: According to [8], the following inequality holds for all $\mathbf{x} \in \mathbb{C}^{M_1}, \mathbf{y} \in \mathbb{C}^{M_2}$ and $\bar{\mathbf{x}} \in \mathbb{C}^{M_1}, \bar{\mathbf{y}} \in \mathbb{C}^{M_2}$

$$\ln\left(1 + \frac{\|\mathbf{x}\|^2}{\|\mathbf{y}\|^2 + \sigma^2}\right) \geq a - \frac{\|\bar{\mathbf{x}}\|^2}{2\mathcal{R}\{\bar{\mathbf{x}}^H \mathbf{x}\} - \|\bar{\mathbf{x}}\|^2} - b\|\mathbf{x}\|^2 - c\|\mathbf{y}\|^2. \tag{4}$$

Applying the inequality in (4) for $x = \mathbf{h}_k^H \mathbf{w}_k$, $y = \mathcal{L}_k(\mathbf{w})$, $\bar{x} = \mathbf{h}_k^H \mathbf{w}_k^{(n)}$, $\bar{y} = \mathcal{L}_k(\mathbf{w}^{(n)})$, where $\mathcal{L}_k(\mathbf{w})$ arranges $\mathbf{h}_k^H \mathbf{w}_i, i \neq k$ into a vector of dimension $K - 1$, we obtain

$$\begin{aligned} f_k\left(\mathbf{w}\right) \geq \bar{a}_k^{(n)} &- \frac{\left|\mathbf{h}_k^H \mathbf{w}_k^{(n)}\right|^2}{2\mathcal{R}\left\{\left(\mathbf{w}_k^{(n)}\right)^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{w}_k\right\} - \left|\mathbf{h}_k^H \mathbf{w}_k^{(n)}\right|^2} \\ &- \bar{b}_k^{(n)} \left|\mathbf{h}_k^H \mathbf{w}_k\right|^2 - \bar{c}_k^{(n)} \sum_{i \neq k} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 \triangleq f_k^{(n)}\left(\mathbf{w}\right), \end{aligned} \tag{5}$$

with the constraint of

$$2\mathcal{R}\left\{\left(\mathbf{w}_k^{(n)}\right)^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{w}_k\right\} - \left|\mathbf{h}_k^H \mathbf{w}_k^{(n)}\right|^2 > 0, \tag{6}$$

where $\bar{a}_k^{(n)} = f_k(\mathbf{w}^{(n)}) + 2 - \frac{|\mathbf{h}_k^H \mathbf{w}_k^{(n)}|^2}{\beta_k^{(n)}} \frac{\sigma^2}{\alpha_k^{(n)}}, 0 < \bar{b}_k^{(n)} = \frac{\bar{a}_k^{(n)}}{\beta_k^{(n)} |\mathbf{h}_k^H \mathbf{w}_k^{(n)}|^2}$, $0 < \bar{c}_k^{(n)} = \frac{|\mathbf{h}_k^H \mathbf{w}_k^{(n)}|^2}{\beta_k^{(n)} \alpha_k^{(n)}}$, $\alpha_k^{(n)} \triangleq \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{w}_i^{(n)}|^2 + \sigma^2$, and $\beta_k^{(n)} \triangleq \sum_{i=1}^{K} |\mathbf{h}_k^H \mathbf{w}_i^{(n)}|^2 + \sigma^2$. According to [8], the function $f_k^{(n)}(\mathbf{w})$ is concave over the trust region (6) and achieves the same value as $f_k(\mathbf{w})$ at $\mathbf{w}^{(n)}$, $f_k^{(n)}(\mathbf{w}^{(n)}) = f_k(\mathbf{w}^{(n)})$.

2) Upper Bounding for $g_k(\mathbf{W})$: Since the function $f(x) = \sqrt{x}$ is concave on $x > 0$, the following inequality for all $x > 0$ and $\bar{x} > 0$ holds true

$$\sqrt{x} = f(x) \leq f\left(\bar{x}\right) + \left.\frac{\partial f(x)}{\partial x}\right|_{x=\bar{x}} (x - \bar{x}) = \frac{\sqrt{\bar{x}}}{2} + \frac{x}{2\sqrt{\bar{x}}}, \tag{7}$$

where $\frac{\partial f(x)}{\partial x}$ refers to the partial derivative of the function $f(x) \leq f(\bar{x})$ with respect to x. Applying the inequality in (7) for $x = 1 - 1/(1 + \varphi_k(\mathbf{w}))^2$ and $\bar{x} = 1 - 1/(1 + \varphi_k(\mathbf{w}^{(n)}))^2$ and using

$$\begin{aligned} &\left(\sum_{i \neq k} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right)^2 \Big/ \left(\sum_{i=1}^{K} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right)^2 \\ &\geq \frac{4\alpha_k^{(n)}}{\left(\beta_k^{(n)}\right)^2} \left(\sum_{i \neq k} \left(2\mathcal{R}\left\{\left(\mathbf{h}_k^H \mathbf{w}_i^{(n)}\right)^* \mathbf{h}_k^H \mathbf{w}_i\right\} - \left|\mathbf{h}_k^H \mathbf{w}_i^{(n)}\right|^2\right) + \sigma^2\right) \\ &\quad - \frac{2\left(\alpha_k^{(n)}\right)^2}{\left(\beta_k^{(n)}\right)^3} \left(\sum_{i=1}^{K} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right) - \frac{\left(\sum_{i \neq k} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right)^2}{\left(\beta_k^{(n)}\right)^2}, \end{aligned} \tag{8}$$

with the constraints of

$$\sum_{i=1}^{K} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2 \leq 2\beta_k^{(n)}, \tag{9}$$

$$\begin{aligned} &\frac{1}{\left(\beta_k^{(n)}\right)^2} \left(\sum_{i=1}^{K} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right) \leq \frac{2}{\alpha_k^{(n)}} \\ &\times \left(\sum_{i \neq k} \left(2\mathcal{R}\left\{\left(\mathbf{h}_k^H \mathbf{w}_i^{(n)}\right)^* \mathbf{h}_k^H \mathbf{w}_i\right\} - \left|\mathbf{h}_k^H \mathbf{w}_i^{(n)}\right|^2\right) + \sigma^2\right), \end{aligned} \tag{10}$$

Algorithm 1: Path-Following Algorithm for Solving Problem (3a).

```
Initialization: Iterate the convex problem (14) until the convergence to obtain an initial point $\mathbf{w}^{(0)}$. Set $n = 0$.
Using (5) to obtain a concave lower bound for $f_k(\mathbf{w})$ with constraint (6).
Using (11) to obtain a convex upper bound for $g_k(\mathbf{w})$ with constraints (9) and (10).
Using (12) to obtain a concave lower bound for $R_k(\mathbf{w})$ under the trust region constrained by (6), (9), and (10).
Repeat until (3a) converges : Solve the convex problem (13) to generate $\mathbf{w}^{(n+1)}$.
```

we have

$$g_k\left(\mathbf{w}\right) \leq d_k^{(n)} - \frac{4\alpha_k^{(n)} e_k^{(n)}}{\left(\beta_k^{(n)}\right)^2} \left(\sum_{i\neq k} \left(2\mathcal{R}\left\{\left(\mathbf{h}_k^H \mathbf{w}_i^{(n)}\right)^* \mathbf{h}_k^H \mathbf{w}_i\right\}\right.\right.$$

$$\left.\left. - \left|\mathbf{h}_k^H \mathbf{w}_i^{(n)}\right|^2\right) + \sigma^2\right) + \frac{2\left(\alpha_k^{(n)}\right)^2 e_k^{(n)}}{\left(\beta_k^{(n)}\right)^3}\left(\sum_{i=1}^{K}\left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right)$$

$$+ \frac{\left(\sum_{i\neq k}\left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma^2\right)^2 e_k^{(n)}}{\left(\beta_k^{(n)}\right)^2} \triangleq g_k^{(n)}\left(\mathbf{w}\right), \tag{11}$$

where $0 < d_k^{(n)} = \frac{\sqrt{1-1/(1+\varphi_k(\mathbf{w}^{(n)}))^2}}{2} + \frac{1}{2\sqrt{1-1/(1+\varphi_k(\mathbf{w}^{(n)}))^2}}$, and $0 < e_k^{(n)} = \frac{1}{2\sqrt{1-1/(1+\varphi_k(\mathbf{w}^{(n)}))^2}}$. The function $g_k^{(n)}(\mathbf{w})$ is convex and achieves the same value as $g_k(\mathbf{w})$ at $\mathbf{w}^{(n)}$, $g_k^{(n)}(\mathbf{w}^{(n)}) = g_k(\mathbf{w}^{(n)})$.

3) Concave Lower Bound for $R_k(\mathbf{W})$: By applying (5) and (11), we have $R_k(\mathbf{w}) \geq f_k^{(n)}(\mathbf{w}) - a g_k^{(n)}(\mathbf{w}) \triangleq R_k^{(n)}(\mathbf{w})$, under the trust region constrained by (6), (9), and (10). The function $R_k^{(n)}(\mathbf{w})$ is concave and matches with the function $R_k(\mathbf{w})$ at $\mathbf{w}^{(n)}$:

$$R_k\left(\mathbf{w}^{(n)}\right) = R_k^{(n)}\left(\mathbf{w}^{(n)}\right). \tag{12}$$

At the nth iteration, we solve the following convex problem with the computational complexity $\mathcal{O}((LNK)^3(2K+1))$ to generate the next feasible point $\mathbf{w}^{(n+1)}$:

$$\max_{\mathbf{w}} \min_{k=1,\ldots,K} \left\{R_k^{(n)}\left(\mathbf{w}\right)\right\} \quad \text{s.t.} \quad \text{(3b), (6), (9), (10).} \tag{13}$$

According to (5) and (11), we can conclude that $\min_{k=1,\ldots,K} R_k(\mathbf{w}^{(n+1)}) \geq \min_{k=1,\ldots,K} R_k(\mathbf{w}^{(n)}), \ \forall n,$ which guarantees the monotonicity in convergence.

According to [8], [17], [21], [22], it is important to have a proper initial point $\mathbf{w}^{(0)}$ with the positive URLLC rate. Thus, we start from any random point $\mathbf{w}^{(0)}$ satisfying the convex power constraint $\sum_{k=1}^{K}\left|\mathbf{w}_{kl}\right|^2 \leq K, \forall l$ and (6), and then iterate

$$\max_{\mathbf{w}} \min_{k=1,\ldots,K} f_k^{(n)}\left(\mathbf{w}\right) \quad \text{s.t.} \quad \text{(3b),} \tag{14}$$

The solution obtained by these iterations can be adopted as the feasible initial point $\mathbf{w}^{(0)}$. Finally, Algorithm 1 provides the pseudo-code for the applied path-following procedure.

B. Decentralized Precoding Design

The previously proposed centralized precoding design requires all the APs to upload the instantaneous CSI to the CPU, which put a significant burden on the fronthaul signaling. Besides, the computational complexity of the centralized precoding design can be exceedingly high for a huge number of antennas. As such, there is a desire for designing the precoding in a decentralized manner which only requires local instantaneous CSI at the APs. In practice, the APs can be divided into several non-overlapping cooperation clusters in which the APs in the same cluster shares both the data and the instantaneous CSI to design the precoding vectors. The APs in different clusters only have the knowledge of the statistical CSI, such as the mean and the variance. Note that although APs are divided into clusters, each user is served by all the APs instead of the APs in the cluster which the user resides in.

Assume each cluster contains M APs, therefore, there are L/M clusters in the network. As stated before, each AP can obtain the instantaneous CSI of the APs in the same cluster and the statistical CSI of the APs in different clusters. Therefore, the virtual SINR of user k in cluster $\mathcal{L}$ for designing the precoding vector can be expressed as

$$\varphi_{k\mathcal{L}}^{\mathrm{V}}\left(\mathbf{w}_{k\mathcal{L}}\right) = \frac{\left|\sum_{l\in\mathcal{L}} \mathbf{h}_{kl}^H \mathbf{w}_{kl} + \sum_{\bar{l}\notin\mathcal{L}} \mathbb{E}\left\{\mathbf{h}_{k\bar{l}}^H\right\} \mathbf{w}_{k\bar{l}}\right|^2}{\sum_{i\neq k}^{K}\left|\sum_{l\in\mathcal{L}} \mathbf{h}_{kl}^H \mathbf{w}_{il} + \sum_{\bar{l}\notin\mathcal{L}} \mathbb{E}\left\{\mathbf{h}_{k\bar{l}}^H\right\} \mathbf{w}_{i\bar{l}}\right|^2 + \sigma^2}. \tag{15}$$

Since we consider Rayleigh fading channels, we have $\mathbb{E}\{\mathbf{h}_{k\bar{l}}^H\} = \mathbf{0}$. Therefore, (15) can be written as

$$\varphi_{k\mathcal{L}}^{\mathrm{V}}\left(\mathbf{w}_{k\mathcal{L}}\right) = \frac{\left|\sum_{l\in\mathcal{L}} \mathbf{h}_{kl}^H \mathbf{w}_{kl}\right|^2}{\sum_{i\neq k}^{K}\left|\sum_{l\in\mathcal{L}} \mathbf{h}_{kl}^H \mathbf{w}_{il}\right|^2 + \sigma^2}. \tag{16}$$

The decentralized max-min URLLC rate optimization problem can be expressed as

$$\max_{\mathbf{w}_{\mathcal{L}}^{\mathrm{V}}} \min_{k=1,\ldots,K} R_{k\mathcal{L}}^{\mathrm{V}}\left(\mathbf{w}_{\mathcal{L}}^{\mathrm{V}}\right)$$

$$\text{s.t.} \quad \sum_{k=1}^{K}\left|\mathbf{w}_{k\mathcal{L}}\right|^2 \leq p_{\max}, \forall l \in \mathcal{L}, \tag{17}$$

where $\mathbf{w}_{\mathcal{L}}^{\mathrm{V}}$ represents the precoding vectors designed for all the users by APs in cluster $\mathcal{L}$ according to (16), and $R_{k\mathcal{L}}^{\mathrm{V}}(\mathbf{w}_{k\mathcal{L}}^{\mathrm{V}}) = \ln\left(1 + \varphi_{k\mathcal{L}}^{\mathrm{V}}(\mathbf{w}_{k\mathcal{L}}^{\mathrm{V}})\right) - \sqrt{\frac{1}{tB} \times V_{k\mathcal{L}}^{\mathrm{V}}} \times Q^{-1}(\epsilon), \quad V_{k\mathcal{L}}^{\mathrm{V}} = 1 - \frac{1}{(1+\varphi_{k\mathcal{L}}^{\mathrm{V}}(\mathbf{w}_{k\mathcal{L}}^{\mathrm{V}}))^2}$.

The problem (17) can be solved in a similar approach as the one for (3a). When the problem (17) has been solved for all the clusters, we can obtain the precoding vector for user k by

$$\mathbf{w}_k = \left[\left(\mathbf{w}_{k1}^{\mathrm{V}}\right)^H, \ldots, \left(\mathbf{w}_{k(L/M)}^{\mathrm{V}}\right)^H\right]^H. \tag{18}$$

Then, the URLLC rate of user k can be obtained by computing (2) using the precoding vector obtained from (18). The computational complexity for each iteration in decentralized precoding design is $\mathcal{O}(((\frac{L}{M})NK)^3(2K+1))$. Compared with the centralized precoding, the computational complexity decreased by M^3.

IV. Numerical Results

In this section, we evaluate the performance of the proposed PFA precoding design for the centralized and the decentralized fashion and investigate the impact of the precoding schemes, the length of transmission duration t, the number of antennas equipped at each AP N, and the size of the AP cluster M on the URLLC rate. We first describe

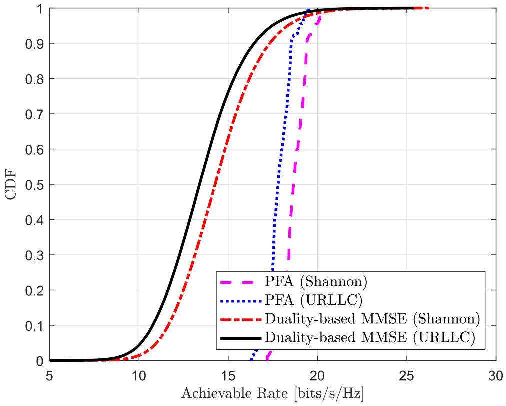

Fig. 1. CDF of the achievable rate achieved by the centralized PFA precoding and the duality-based MMSE precoding with $t = 0.05$ ms, $B = 1$ MHz, $K = 6$, and $N = 4$.

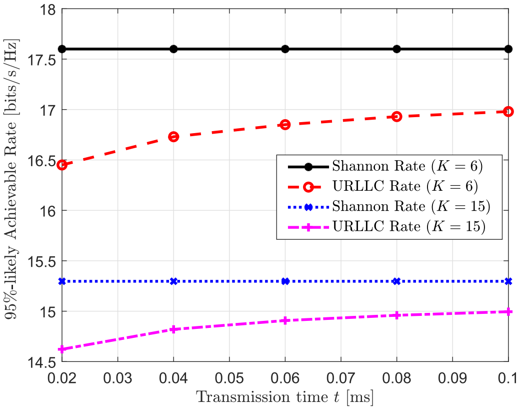

Fig. 2. Optimized 95%-likely achievable rate versus the transmission time t with $N = 4$ and $B = 1$ MHz.

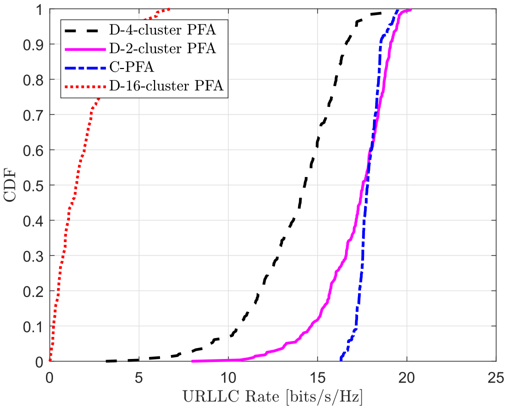

Fig. 3. CDF of the URLLC rate achieved by the PFA precoding in the centralized and decentralized way with $t = 0.05$ ms and $N = 4$.

our adopted simulation parameters. We adopt the similar parameters setting as in [15] as the basis to establish our simulation system model. L APs and K users are deployed in a rectangular area of 96×48 m^2. In particular, the APs are deployed on a rectangle grid. The area is wrapped around at the edges to avoid the boundary effects [15]. The horizontal spacing between APs are 24 m, and the vertical spacing is 12 m. The K users are deployed randomly. We adopt a similar propagation model as in [14]. Besides, we set $L = 16$, $\tau_p = 3$, and $\epsilon = 10^{-5}$. Note that in all the figures, the achievable rates are calculated in bits/s/Hz.

Fig. 1 shows the cumulative distribution functions (CDFs) of the achievable rate per user achieved by the proposed PFA centralized precoding and the duality-based MMSE precoding with $t = 0.05$ ms, $B = 1$ MHz, $K = 6$, and $N = 4$ which is given by

$$\mathbf{w}_k = \frac{\mathbf{v}_k}{\left\|\mathbf{v}_{kl}\right\|}, \quad \mathbf{v}_k = p\left(\sum_{i=1}^{K} p\mathbf{h}_i\mathbf{h}_i^H + \sigma^2\mathbf{I}_{LN}\right)^{-1}\mathbf{h}_k, \qquad (19)$$

where p is the transmit power intend for each user at each AP. It can be observed that the proposed PFA centralized precoding scheme performs very well. The achievable rate per user distribution with the proposed PFA centralized precoding almost uniformly outperforms the duality-based MMSE precoding, and the former is more steeper. Specifically, applying the PFA centralized precoding leads to 32% improvement in terms of average URLLC rate and 65% improvement in terms of 95%-likely URLLC rate. Note that the duality-based MMSE precoding in (19) is only a heuristic solution utilizing the uplink-downlink duality and cannot effectively minimize the MSE $\mathbb{E}\{|y_k - s_k|^2|\mathbf{h}_{kl}\}$. Moreover, compared with the PFA centralized precoding, the duality-based MMSE precoding has a lower computational complexity since it only requires $\frac{N^2L^2K+NLK}{2} + \frac{N^3L^3-NL}{3} + N^2L^2$ complex-valued multiplications. Besides, as expected, the performance of Shannon rate serves as a performance upper bound of the URLLC rate at the expense of infinitely long code length.

Fig. 2 plots the optimized 95%-likely achievable rate by Algorithm 1 versus the transmission time t with $N = 4$ and $B = 1$ MHz. As expected, the URLLC rate increases along with the transmission time t according to the expression of the URLLC rate. Note that the Shannon rate is fixed since it is computed assuming a sufficient long blocklength, e.g., $t \to \infty$. Besides, when the number of user increases from 6 to 15, we can observe that the achievable rate decreases since there are more users competing for limited resources that reduces the flexibility of the resource allocation for effective beamforming. The performance gap between the Shannon rate and URLLC rate is also reduced with the increasing number of users as the performance of these two scheme is limited by the user with the weakest channel gain.

Fig. 3 shows the performance of the PFA precoding in the centralized and decentralized fashion in terms of the URLLC rate. The curve “C-PFA” represents the URLLC rate computed using the centralized PFA precoding design. Also, the curve “D-4-cluster,” “D-2-cluster,” and “D-16-cluster” stand for the performance of the decentralized PFA precoding design with 4 APs, 8 APs, and 1 AP in each cluster, respectively. The first observation from Fig. 3 is that compared with the centralized PFA precoding, the 95%-likely URLLC rate with the decentralized PFA precoding is generally lower. This is because when the decentralized PFA precoding is adopted, only the instantaneous CSI within the cluster and the statistical CSI outside the cluster are used for optimization in each cluster. As there is a mismatch between the statistical CSI and the instantaneous CSI, the optimization for the decentralized setting is less effective for the utilization of the system resources. Besides, the performance of the 2-cluster decentralized PFA precoding outperforms the centralized PFA precoding for the strong users. The reason is that the performance of the centralized PFA precoding is always limited by the worst-case users, since substantial resources are allocated to equalize all the SINRs, while the decentralized PFA precoding benefits from being more scalable. Compared with the 2-cluster decentralized PFA precoding, when adopting the 4-cluster or 16-cluster decentralized PFA precoding, the mismatch between the statistical CSI and the instantaneous CSI is pronounced, so the performance is the worse. Specifically, compared with the centralized precoding, the 95%-likely URLLC rate is reduced from 16.73 bits/s/Hz to 13.25 bits/s/Hz with the 2-cluster decentralized PFA precoding and to 8.95 bits/s/Hz with the 4-cluster decentralized PFA precoding. Moreover, when the fully distributed 16-cluster decentralized PFA precoding is adopted, the 95%-likely URLLC rate is only 0.17 bits/s/Hz. However, since the computational complexity is also reduced, the performance loss of adopting the 2-cluster decentralized PFA precoding instead of the centralized precoding is tolerable. In particular, the 2-cluster decentralized PFA precoding achieves 80% of the 95%-likely URLLC rate, 89% of the average URLLC rate, and 12% of the computational complexity of the centralized precoding. The second observation is that the CDF of users' URLLC rate is not as steep as the counterpart when the decentralized PFA precoding design is adopted. The reason is that the optimization target of each cluster contains virtual SINR rather than the actual SINR, leading to under utilisation of system resources.

V. Conclusion

In this correspondence, we considered the precoding design in the cell-free massive MIMO system for URLLC in the centralized and decentralized fashion. PFA was designed for maximizing the users' minimum URLLC rate and its performance was evaluated with different settings of the transmission time, the number of antennas per AP, and the size of the AP cluster. Simulation results showed that the centralized PFA precoding design can effectively improve the performance of 95%-likely achievable rate and the decentralized PFA precoding with a reasonable setting can approach the performance of the former but with low computational complexity. In the future, we will jointly optimize the precoding vector, the cluster formation, and the number of APs in each cluster in a distributed fashion for URLLC.

References

[1] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

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

[3] J. Zhang et al., "RIS-aided next-generation high-speed train communications: Challenges, solutions, and future directions," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 145–151, Dec. 2021.

[4] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[5] L. Liu, Y. Zhou, W. Zhuang, J. Yuan, and L. Tian, "Tractable coverage analysis for hexagonal macrocell-based heterogeneous UDNs with adaptive interference-aware CoMP," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 503–517, Jan. 2018.

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

[7] C. She, C. Yang, and T. Q. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Tran. Commun.*, vol. 66, no. 5, pp. 2266–2280, May 2018.

[8] 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. 2020.

[9] S. He, Z. An, J. Zhu, J. Zhang, Y. Huang, and Y. Zhang, "Beamforming design for multiuser URLLC with finite blocklength transmission," *IEEE Transa. Wireless Commun.*, vol. 20, no. 12, pp. 8096–8109, Dec. 2021.

[10] J. Zhang, J. Zhang, E. Björnson, and B. Ai, "Local partial zero-forcing combining for cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8459–8473, Dec. 2021.

[11] J. Zhang, J. Zhang, D. W. K. Ng, S. Jin, and B. Ai, "Improving sum-rate of cell-free massive MIMO with expanded compute-and-forward," *IEEE Trans. Signal Process.*, vol. 70, no. 11, pp. 202–215, Nov. 2021.

[12] J. Zheng, J. Zhang, E. Björnson, Z. Li, and B. Ai, "Cell-free massive MIMO-OFDM for high-speed train communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 2823–2839, Oct. 2022.

[13] Z. Wang, J. Zhang, B. Ai, C. Yuen, and M. Debbah, "Uplink performance of cell-free massive MIMO with multi-antenna users over jointly-correlated rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7391–7406, Sep. 2022.

[14] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2019.

[15] 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.

[16] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[17] 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.

[18] A. Lancho, G. Durisi, and L. Sanguinetti, "Cell-free massive MIMO for URLLC: A finite-blocklength analysis," 2022, *arXiv:2207.00856*.

[19] E. Shi et al., "Wireless energy transfer in RIS-aided cell-free massive MIMO systems: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 60, no. 3, pp. 26–32, Mar. 2022.

[20] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, no. 3–4, pp. 154–655, 2017.

[21] C. Xing, S. Wang, S. Chen, S. Ma, H. V. Poor, and L. Hanzo, "Matrix-monotonic optimization—Part I: Single-variable optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 738–754, 2021.

[22] C. Xing, S. Wang, S. Chen, S. Ma, H. V. Poor, and L. Hanzo, "Matrix-monotonic optimization—Part II: Multi-variable optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 179–194, 2021.