Statistical Delay and Error-Rate Bounded QoS Provisioning Over mmWave Cell-Free M-MIMO and FBC-HARQ-IR Based 6G Wireless Networks

Xi Zhang, *Fellow, IEEE*, Jingqing Wang, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—As a new and dominating 6G mobile-networks' service class for time-sensitive traffics, massive ultra-reliable and low latency communications (mURLLC) has received tremendous attention. One of key 6G enabling-techniques for achieving mURLLC lies in how to efficiently support statistical delay and error-rate bounded quality-of-services (QoS) provisioning for real-time data-transmissions over time-varying wireless networks. Towards this end, several emerging wireless techniques, including finite blocklength coding (FBC), hybrid automatic repeat request with incremental redundancy (HARQ-IR) protocol, millimeter wave (mmWave), cell-free (CF) massive multiple-input multiple-output (m-MIMO), etc., have been shown to be 6G promising enablers to significantly improve various QoS performances. However, integrating these techniques with the statistical delay and error-rate bounded QoS provisioning theory for mURLLC has imposed many new difficulties not encountered before. To overcome these challenges, in this paper we propose the statistical delay-and-error-rate-bounded QoS provisioning system architecture over mmWave user-centric CF m-MIMO and FBC-HARQ-IR based 6G wireless networks. First, we establish the comprehensive system models by accurately characterizing the integrations of above-described 6G promising techniques with statistical QoS provisioning theory. Then, we integrate FBC with HARQ-IR protocol to derive the channel capacity as a function of error probability. Finally, we obtain the closed-form expressions for effective capacities under our proposed schemes. We also conduct a set of simulations to validate and evaluate our proposed FBC-HARQ-IR based mmWave user-centric CF m-MIMO schemes.**

***Index Terms*—Statistical delay and error-rate bounded QoS, 6G mobile wireless networks, mURLLC, mmWave user-centric CF m-MIMO, FBC-HARQ-IR, effective capacity.**

I. Introduction

THE DELAY-BOUNDED statistical quality-of-service (QoS) theory [1], [2] has been proposed as a promising technique to support the explosively growing demands of time-sensitive wireless multimedia applications over the upcoming sixth generation (6G) multimedia mobile wireless networks. Given a specific delay bound, the design issues of QoS provisioning for multimedia wireless services have received considerable research attention. Because of the highly time-varying nature of wireless fading channels, deterministic delay-bounded QoS constraints are no longer feasible to characterize queuing behaviors of multimedia wireless services. Towards this end, the concept of *statistical QoS provisioning* [3]–[5], in terms of effective capacity and delay-bounded violation probabilities, has been proposed to support time-sensitive wireless communications over 6G multimedia mobile wireless networks.

In addition, as one of the 6G standard traffic services, massive ultra-reliable and low latency communications (mURLLC) [6] requires short-packet data communications to support time-sensitive wireless multimedia services. Along this direction, finite blocklength coding (FBC) is proposed in [7] and the results have shown that the codeword blocklength can be as short as 100 channel symbols for reliable communications. The authors of [8] have exploited recent results of non-asymptotic coding rate to derive the goodput over additive white Gaussian noise (AWGN) channels and the energy-efficiency spectral-efficiency tradeoff. The maximum achievable coding rate using FBC over AWGN channels has been derived in [9]. The authors of [10] have investigated different properties of channel codes that approach the fundamental limits of a given memoryless wireless channel in the finite blocklength regime.

Moreover, due to the ultra-low latency and high reliability guarantees of mURLLC, the traditional automatic repeat and request (ARQ) scheme, which requires acknowledgement (ACK) or non-ACK (NACK) feedback, is no longer efficient for supporting delay-sensitive wireless multimedia applications. Towards this end, researchers have developed hybrid automatic repeat request (HARQ) [11] protocols, including HARQ with incremental redundancy (HARQ-IR) [12] and HARQ chase combining (HARQ-CC) [13], to adaptively control the transmission rate based on decoding feedback. There has been a great deal of research focusing on the performance analyses of HARQ protocols while being integrated with FBC. The authors of [14] have compared the link-level system performance with HARQ-IR and HARQ-CC, and shown that HARQ-IR can significantly improve channel

Manuscript received October 1, 2019; revised January 15, 2020; accepted February 17, 2020. Date of publication July 1, 2020; date of current version August 20, 2020. The work of Xi Zhang and Jingqing Wang was supported in part by the U.S. National Science Foundation under Grant CCF-2008975, Grant ECCS-1408601, and Grant CNS-1205726, and in part by the U.S. Air Force under Grant FA9453-15-C-0423. The work of H. Vincent Poor was supported in part by the U.S. National Science Foundation under Grant CCF-0939370 and Grant CCF-1513915. *(Corresponding author: Xi Zhang.)*

Xi Zhang and Jingqing Wang are with Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xizhang@ece.tamu.edu; wang12078@tamu.edu).

H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2020.3000804

coding rate as compared with HARQ-CC. The authors of [12] have studied power allocation policies using the HARQ-IR protocol when analyzing reliable downlink data transmissions under QoS constraints. The impact of fixed transmission rate, queuing constraints, and hard-deadline limitations on the throughput has been studied in [15] while applying the HARQ-IR protocol.

On the other hand, massive multiple-input multiple-output (m-MIMO) systems [16] with large antenna arrays implemented at the base stations (BSs) to simultaneously serve large numbers of mobile users have been developed to support the explosively increasing number of mobile devices while addressing the wireless spectrum scarcity problem. When designing m-MIMO systems, inter-cell interference is becoming a major bottleneck and cannot be removed due to a cell-centric implementation. To solve the design issues for the traditional m-MIMO systems, cell-free m-MIMO [17] has been proposed as a promising network architecture for 6G multimedia mobile wireless networks. Such a distributed cell-free m-MIMO system architecture model significantly increases throughput as well as user coverage, and the co-processing at multiple access points (APs) can suppress the inter-cell interference. On the other hand, cell-free m-MIMO system requires higher capacity of backhaul connections and the co-processing at the APs increases backhaul overhead. To reduce this backhaul overhead, a "user-centric" approach has been proposed for cell-free m-MIMO systems, where each mobile user is served by a selected subset of APs that are within the user-centric cluster. Two AP selection strategies for user-centric cell-free m-MIMO systems are proposed in [18], including received-power-based selection and largest-large-scale-fading-based selection. The authors of [19] have shown that such a user-centric cell-free m-MIMO approach outperforms the pure cell-free m-MIMO approach in terms of achievable data-rate per-user for the vast majority of mobile users in the network. The authors of [20] have studied the downlink performance of cell-free m-MIMO systems in terms of the minimum data-rate among all users.

When integrating m-MIMO techniques with mmWave techniques, the major design issues, such as accurate channel estimation, have been investigated to reduce the hardware complexity as well as power consumption over the mmWave m-MIMO based wireless fading channels. To resolve such problems, the authors of [21] have developed low-complexity multiuser hybrid analog/digital precoding algorithms with limited feedbacks. Although the traditional suboptimal approaches for selecting analog precoders and combiners can avoid exhaustive search, they still require some high-complexity operations. Due to the sparsity characteristics of mmWave wireless fading channels, researchers have developed a dictionary learning method [22] to solve the hybrid beamforming optimization problem in a low-complexity way. The authors of [23] have proposed an algorithm for adapting dictionaries in order to achieve sparse signal representations. The authors of [22] have developed a dictionary learning-based channel estimation model such that a dictionary is learned from comprehensively collected channel measurements. Furthermore, when being integrated with mmWave techniques, the assumptions and analytical results in [17], [20], [24] for cell-free m-MIMO systems cannot be directly applied in mmWave frequency bands. The authors of [25] have introduced and analyzed the user-centric and cell-free system architectures at millimeter wave frequencies. The authors of [26] have proposed downlink power control algorithms to maximize the global energy efficiency in mmWave user-centric and cell-free m-MIMO architectures. However, how to efficiently integrate mmWave with cell-free m-MIMO architecture models in the finite blocklength regime under statistical delay and error-rate bounded constraints is still an open problem.

To effectively overcome the above-mentioned challenges, in this paper we integrate an mmWave user-centric cell-free m-MIMO system with the FBC-HARQ-IR technique over 6G wireless networks. In particular, we establish mmWave user-centric cell-free m-MIMO based system models. Then, we apply the dictionary learning method to design a low-complexity beam-training algorithm for solving the beam-training optimization problem. We also apply the FBC-HARQ-IR protocol to determine the channel capacity as well as error probability using FBC. Based on the information theoretic results in QoS theory, we characterize QoS metrics in terms of error probability and derive the corresponding effective capacity function for our proposed FBC-HARQ-IR based mmWave cell-free m-MIMO schemes. We also conduct a set of simulations to validate and evaluate our proposed mmWave user-centric cell-free m-MIMO schemes by implementing statistical delay and error-rate bounded QoS provisioning in the finite blocklength regime.

The rest of this paper is organized as follows: Section II establishes mmWave user-centric cell-free m-MIMO based system models. Section III designs the dictionary learning based beam-training algorithm. Section IV derives the channel capacity and error probability using the FBC-HARQ-IR protocol. Section V derives and analyzes statistical delay and error-rate bounded QoS metrics and the effective capacity function in the finite blocklength regime. Section VI evaluates and analyzes the system performance for our proposed FBC-HARQ-IR based mmWave user-centric cell-free m-MIMO schemes. The paper concludes with Section VII.

II. The Network Architecture and System Models

A. The mmWave User-Centric Cell-Free m-MIMO and FBC-HARQ-IR System Architecture

Fig. 1 shows our proposed mmWave user-centric cell-free m-MIMO and FBC-HARQ-IR system architecture, assuming that there are K_{a} randomly located APs over a large area and K_{u} mobile users ($K_{\mathrm{u}} \ll K_{\mathrm{a}}$). As shown in Fig. 1, we consider a mmWave user-centric cell-free m-MIMO network model, where each mobile user is served by coherent joint transmissions from a selected subset of APs which are within the user-centric cluster. Fig. 2 shows the system transmitting-group and receiving-group model at PHY-layer for our proposed mmWave cell-free m-MIMO and FBC-HARQ-IR based 6G multimedia mobile wireless networks. We assume that each AP is equipped with N_{T} antennas and L_{T} RF chains, while each mobile user is equipped with a single antenna ($N_{\mathrm{T}} > K_{\mathrm{u}}, N_{\mathrm{T}} \geq L_{\mathrm{T}}$, and $L_{\mathrm{T}} < K_{\mathrm{u}}$). Define $\mathcal{G}(k)$ as the

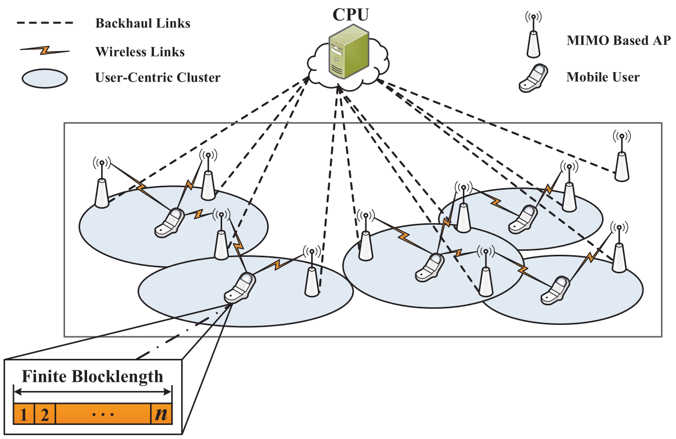


Fig. 1. The system architecture model for user-centric cell-free m-MIMO based 6G multimedia mobile wireless networks.

group of mobile users served by the kth AP, where $|\mathcal{G}(k)| = L_{\mathrm{T}} < K_u$. Define $\mathcal{K}(m)$ as the set of clustered APs that serve mobile user m. All APs are connected to a central processing unit (CPU) through backhaul links.

B. The mmWave User-Centric Cell-Free m-MIMO Based System Models in the Finite Blocklength Regime

As shown in Fig. 3, each time interval is divided into three phases: 1) Large-scale beam-training phase for hybrid precoder design and user selection, where each AP chooses an optimal RF precoder and selects a group of mobile users $\mathcal{G}(k)$ with the best channel quality; 2) Small-scale uplink training phase, where mobile users send uplink pilot symbols to the APs, then each AP estimates the wireless channels to all mobile users based on the received pilot symbols; 3) Downlink finite-blocklength data transmission phase, where the APs use the knowledge of channel estimation obtained in the previous small-scale uplink training phase, precode, and transmit the finite-blocklength data to mobile users under HARQ-IR protocol, as shown in Fig. 3. HARQ-IR protocol will be discussed in detail in Section IV. There, we define n_{p} as the number of channel uses for uplink pilot training symbols and n_{d} as the number of channel uses reserved for transmitting L equal-length downlink-data blocks with $\widehat{n}$ symbols each for implementing HARQ-IR protocol, i.e., $n_{\mathrm{d}} = L\widehat{n}$. Define n as the total number of channel uses for both uplink pilot training and downlink data transmission phases. Thus, we have $n = n_{\mathrm{p}} + n_{\mathrm{d}} = n_{\mathrm{p}} + L\widehat{n}$ (see Fig. 3).

1) Large-Scale Beam-Training Phase: The goal of large-scale beam-training phase is to develop an efficient algorithm to design an optimal RF precoder $\mathbf{F}_{k,m}^{\mathrm{R,opt}} \in \mathbb{C}^{N_{\mathrm{T}} \times L_{\mathrm{T}}}$ and user selection group $\mathcal{G}(k)$. Define an equivalent channel's impulse response vector, denoted by $\widetilde{\boldsymbol{h}}_{k,m}$, between the kth AP and mobile user m as in the following equation:

$$\widetilde{\boldsymbol{h}}_{k,m} = \boldsymbol{h}_{k,m}\mathbf{F}_{k,m}^{\mathrm{R}} \tag{1}$$

where $\boldsymbol{h}_{k,m} \in \mathbb{C}^{1 \times N_{\mathrm{T}}}$ represents the channel's impulse response vector from the kth AP to mobile user m and $\mathbf{F}_{k,m}^{\mathrm{R}} \in \mathbb{C}^{N_{\mathrm{T}} \times L_{\mathrm{T}}}$ is the analog precoder from AP k to mobile user m. Correspondingly, we can formulate the maximization problem $\mathbf{P_1}$ which selects the best mobile user and the optimal precoding matrix $\left\{m^{\mathrm{opt}}, \mathbf{F}_{k,m}^{\mathrm{R,opt}}\right\}$ as follows:

$$\mathbf{P_1}: \left\{m^{\mathrm{opt}}, \mathbf{F}_{k,m}^{\mathrm{R,opt}}\right\} = \arg \max_{\left\{m,\, \mathbf{F}_{k,m}^{\mathrm{R}}\right\}} \left\{\left\|\boldsymbol{h}_{k,m}\mathbf{F}_{k,m}^{\mathrm{R}}\right\|_F^2\right\} \tag{2}$$

$$\text{s.t. C1: } \mathbf{F}_{k,m}^{\mathrm{R}} \in \boldsymbol{\mathcal{F}}_c, \quad \forall m;$$

$$\text{C2: } \left\|\mathbf{F}_{k,m}^{\mathrm{R}}\mathbf{F}_{k,m}^{\mathrm{B}}\right\|_F^2 = 1, \quad \forall m, \tag{3}$$

where $\|\mathbf{M}\|_F$ is the Frobenius norm of matrix $\mathbf{M}$, which is defined as $\sqrt{\mathrm{Tr}\left((\mathbf{M}^\dagger)\mathbf{M}\right)}$ where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix and $(\cdot)^\dagger$ denotes the Hermitian transpose of a matrix, $\mathbf{F}_{k,m}^{\mathrm{B}}$ is the digital precoder from AP k to mobile user m, and $\boldsymbol{\mathcal{F}}_c = \{\boldsymbol{f}_c(1), \ldots, \boldsymbol{f}_c(N_{\mathrm{T}})\}$ is the beam steering codebook stored at the APs, where $\boldsymbol{f}_c(i)$ is given by

$$\boldsymbol{f}_c(i) = \frac{1}{\sqrt{N_{\mathrm{T}}}}\left[1, e^{\jmath\pi\left(-1+\frac{(2i-1)}{N_{\mathrm{T}}}\right)}, \ldots, e^{\jmath\pi(N_{\mathrm{T}}-1)\left(-1+\frac{2i-1}{N_{\mathrm{T}}}\right)}\right]^T \tag{4}$$

for $i = 1, \ldots, N_{\mathrm{T}}$ where $\jmath = \sqrt{-1}$ and $[\cdot]^T$ is the transpose of a vector. Note that the vector $\boldsymbol{f}_c(i)$ has the same structure as the antenna array's response vector. To solve the above optimization problem $\mathbf{P_1}$, we can apply the exhaustive search method. However, the complexity of such exhaustive search method is too high, especially for m-MIMO scenario. Accordingly, we design low-complexity suboptimal beam-training algorithm to solve $\mathbf{P_1}$ and its corresponding user selection algorithm in Section III.

2) Small-Scale Uplink Training Phase: Define the pilot training sequence from all K_{u} mobile users as $\boldsymbol{s}_m \triangleq \left[s_m^{(1)}, \ldots, s_m^{(n_{\mathrm{p}})}\right]$ and $\|\boldsymbol{s}_m\|_F^2 = 1$. During the small-scale uplink training phase, we can derive the received signal, denoted by $\boldsymbol{y}_k^{(l)}$, at the kth AP for transmitting the lth training data block as in the following equation:

$$\boldsymbol{y}_k^{(l)} = \sum_{m=1}^{K_{\mathrm{u}}} \sqrt{\mathcal{P}_{\mathrm{p}}}\widetilde{\boldsymbol{h}}_{k,m}s_m^{(l)} + \boldsymbol{n}_k^{(l)}, \quad l = 1, \ldots, n_{\mathrm{p}} \tag{5}$$

where $\mathcal{P}_{\mathrm{p}}$ is the uplink pilot transmit power from each mobile user to the AP; $s_m^{(l)}$ denotes the pilot training signal sent from mobile user m to the kth AP; $\widetilde{\boldsymbol{h}}_{k,m}$ represents the equivalent channel's impulse response vector between mobile user m and the kth AP, given in Eq. (1); and $\boldsymbol{n}_k^{(l)}$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ^2. Then, we can derive the minimum mean-squared error (MMSE) channel estimation as follows:

$$\begin{aligned}\widehat{\boldsymbol{h}}_{k,m} &= \frac{\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^\dagger\right]}{\sum_{m=1}^{K_{\mathrm{u}}}\left(\mathbf{F}_{k,m}^{\mathrm{R}}\right)^\dagger\mathbb{E}_{\boldsymbol{h}_{k,m}}\left[\boldsymbol{h}_{k,m}\left(\boldsymbol{h}_{k,m}\right)^\dagger\right]\mathbf{F}_{k,m}^{\mathrm{R}} + \sigma^2\mathbf{I}_{K_u}} \\ &= \frac{\left\|\left(\mathbf{F}_{k,m}^{\mathrm{R}}\right)^\dagger\mathbb{E}_{\boldsymbol{h}_{k,m}}\left[\boldsymbol{h}_{k,m}\left(\boldsymbol{h}_{k,m}\right)^\dagger\right]\mathbf{F}_{k,m}^{\mathrm{R}}\right\|^2}{\sum_{m=1}^{K_{\mathrm{u}}}\left(\mathbf{F}_{k,m}^{\mathrm{R}}\right)^\dagger\mathbb{E}_{\boldsymbol{h}_{k,m}}\left[\boldsymbol{h}_{k,m}(\boldsymbol{h}_{k,m})^\dagger\right]\mathbf{F}_{k,m}^{\mathrm{R}} + \sigma^2\mathbf{I}_{K_u}}\end{aligned} \tag{6}$$

where $\|\cdot\|$ is the Euclidean norm of a matrix, $\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{h}_{k,m}}[\cdot]$ are the expectations with respect to $\widetilde{\boldsymbol{h}}_{k,m}$ and $\boldsymbol{h}_{k,m}$, respectively, and $\mathbf{I}_{K_{\mathrm{u}}}$ is the identity matrix of size K_{u}.

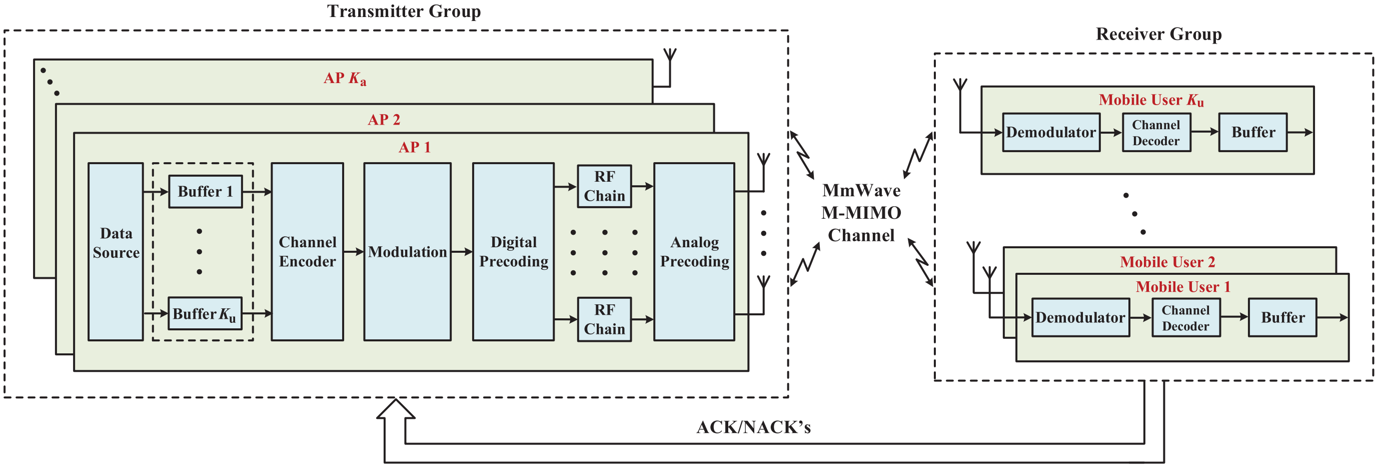


Fig. 2. The system transmitting-group and receiving-group model at PHY-layer for our proposed mmWave cell-free m-MIMO and FBC-HARQ-IR based 6G multimedia mobile wireless networks.

3) User-Centric Downlink Finite-Blocklength Data Transmission Phase: We define the transmit signal matrix as $\mathbf{X}_m^{n_d} \triangleq \left[\boldsymbol{x}_m^{(1)}, \ldots, \boldsymbol{x}_m^{(n_d)}\right]$ and receive signal vector as $\boldsymbol{y}_m^{n_d} \triangleq \left[y_m^{(1)}, \ldots, y_m^{(n_d)}\right]$, respectively. Considering Rayleigh block-fading channel, we can derive the received signal, denoted by $\boldsymbol{y}_m^{n_d} \in \mathbb{C}^{1\times n_d}$ from the kth AP to the mth mobile user for transmitting n_d finite-blocklength data blocks as follows:

$$\boldsymbol{y}_m^{n_d} = \sum_{k\in K(m)} \boldsymbol{h}_{k,m}\mathbf{F}_{k,m}^{\mathrm{R}}\mathbf{F}_{k,m}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m}\right)^{\frac{1}{2}}\mathbf{X}_m^{n_d} + \sum_{\substack{m'=1\\ m'\neq m}}^{K_u}\sum_{k\in\mathcal{K}(m')}\boldsymbol{h}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{R}}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\mathbf{X}_{m'}^{n_d} + \boldsymbol{n}_m \tag{7}$$

where $\mathbf{X}_m^{n_d}$ and $\mathbf{X}_{m'}^{n_d}$ are the signals sent to mobile user m and mobile user m', respectively; $\boldsymbol{h}_{k,m} \in \mathbb{C}^{1\times N_T}$ represents the channel's impulse response vector from the kth AP to mobile user m; $\mathbf{F}_{k,m}^{\mathrm{R}} \in \mathbb{C}^{N_T\times L_T}$ and $\mathbf{F}_{k,m'}^{\mathrm{R}} \in \mathbb{C}^{N_T\times L_T}$ are the analog precoders for mobile user m and mobile user m', respectively; $\mathbf{F}_{k,m}^{\mathrm{B}} \in \mathbb{C}^{L_T\times N_T}$ and $\mathbf{F}_{k,m'}^{\mathrm{B}} \in \mathbb{C}^{L_T\times N_T}$ represent the digital precoders for mobile user m and mobile user m', respectively; $\boldsymbol{\Omega}_{k,m} \in \mathbb{C}^{N_T\times N_T}$ and $\boldsymbol{\Omega}_{k,m'} \in \mathbb{C}^{N_T\times N_T}$ denote the power allocation matrices which allocate total transmit power among N_T streams at the kth AP to mobile user m and mobile user m', respectively; and $\boldsymbol{n}_m$ is the AWGN with zero mean and covariance $\mathbf{I}_{n_d}\sigma^2$. Define $\boldsymbol{\Omega}_{k,m} \triangleq \mathrm{diag}\left\{\omega_{k,m}^{(1)}, \ldots, \omega_{k,m}^{(N_T)}\right\}$ as the power allocation matrix, where diag$\{\cdot\}$ represents the diagonal matrix. In addition, we normalize the precoding matrices as $\left\|\mathbf{F}_{k,m}^{\mathrm{R}}\mathbf{F}_{k,m}^{\mathrm{B}}\right\|_F^2 = 1$. As a result, the received signal at mobile user m can be rewritten as in the following equation:

$$\boldsymbol{y}_m^{n_d} = \sum_{k\in\mathcal{K}(m)} \widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m}\right)^{\frac{1}{2}}\mathbf{X}_m^{n_d} + \sum_{\substack{m'=1\\ m'\neq m}}^{K_u}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\mathbf{X}_{m'}^{n_d} + \boldsymbol{n}_m. \tag{8}$$

In addition, we can determine the signal-to-interference-plus-noise ratio (SINR), denoted by γ_m, for mobile user m as follows:

$$\gamma_m = \frac{\sum\limits_{k\in\mathcal{K}(m)}\left\|\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m}\right)^{\frac{1}{2}}\right\|^2}{\left\|\sum\limits_{\substack{m'=1\\ m'\neq m}}^{K_u}\sum\limits_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right\|^2 + \sigma^2}. \tag{9}$$

III. Dictionary Learning Based Low-Complexity Hybrid Precoder Design

Although the traditional suboptimal methods for selecting analog precoder and combiner can avoid exhaustive search, it still involves some high-complexity matrix operations. Due to the sparsity characteristics of mmWave wireless fading channels, we apply the dictionary learning method [22] to solve the optimization problem in a low-complexity way. During the dictionary learning based beam-training phase, using Eq. (4), we can define the following over-complete beam steering codebook for the RF precoder to achieve the sparse representation for solving the optimization problem:

$$\widetilde{\boldsymbol{f}}_c(i) = \frac{1}{\sqrt{N_T}}\left[1, e^{\jmath\pi\left(-1+\frac{(2i-1)}{M_T}\right)}, \ldots, e^{\jmath\pi(N_T-1)\left(-1+\frac{2i-1}{M_T}\right)}\right]^T \tag{10}$$

for $i = 1, \ldots, M_T$ where $M_T > N_T$. Thus, we can rewrite the beam-training codebook at all APs as $\widetilde{\boldsymbol{\mathcal{F}}}_c = \left\{\widetilde{\boldsymbol{f}}_c(1), \ldots, \widetilde{\boldsymbol{f}}_c(M_T)\right\}$. Such an over-complete matrix introduces redundancy to the original beamforming codebook matrix, improving both flexibility and capability of sparse representation. Using the singular value decomposition (SVD), the channel's impulse response vector $\boldsymbol{h}_{k,m}$ can be derived as follows:

$$\boldsymbol{h}_{k,m} = \mathbf{U}_{k,m}\boldsymbol{\Sigma}_{k,m}\left(\mathbf{V}_{k,m}\right)^{\dagger} \tag{11}$$

where the columns of $\mathbf{U}_{k,m}$ is the left singular vector of $\boldsymbol{h}_{k,m}$; $\boldsymbol{\Sigma}_{k,m}$ denotes the diagonal matrix containing the singular values of $\boldsymbol{h}_{k,m}$; and the rows of $\left(\mathbf{V}_{k,m}\right)^{\dagger}$ represent the right

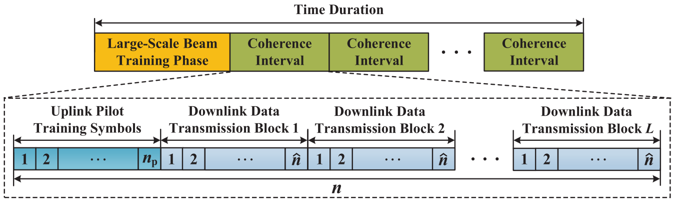

Time Duration
Large-Scale Beam Training Phase
Coherence Interval
Coherence Interval
...
Coherence Interval
Uplink Pilot Training Symbols
Downlink Data Transmission Block 1
Downlink Data Transmission Block 2
Downlink Data Transmission Block L
1
2
...
n_p
$\hat{n}$
n

Fig. 3. Frame structure with large-scale beam-training, small-scale uplink training, and finite-blocklength downlink data transmission phases using HARQ-IR protocol in mmWave user-centric cell-free m-MIMO scheme, where n_{p} is the number of channel uses for uplink pilot training symbols, n is the total number of channel uses for both uplink pilot training and downlink data transmission phases, L is the number of finite-blocklength data blocks for the downlink data transmission using HARQ-IR protocol and $\widehat{n}$ is the blocklength of each data block using HARQ-IR protocol. Thus, $n_{\mathrm{d}} = L\widehat{n}$ and $n = n_{\mathrm{p}} + n_{\mathrm{d}} = n_{\mathrm{p}} + L\widehat{n}$.

singular vectors of $\boldsymbol{h}_{k,m}$. Define an optimal precoder matrix as $\mathbf{F}_{k,m}^{\mathrm{opt}} \triangleq \mathbf{V}_{k,m}$. Then, we define $\mathbf{D}_{k,m} \triangleq [\mathbf{D}_{k,1}, \ldots, \mathbf{D}_{k,K_{\mathrm{u}}}]$ as the dictionary of beamforming codewords learned from large-scale beam-training phase at AP k for mobile user m. Motivated by $\mathbf{P_1}$, we need to minimize the "distance" between the layered precoder $\left\{\mathbf{F}_{k,m}^{\mathrm{R}}, \mathbf{F}_{k,m}^{\mathrm{B}}\right\}$ and an optimal precoder $\mathbf{F}_{k,m}^{\mathrm{opt}}$. As a result, converting $\mathbf{P_1}$, we can formulate the following maximization problem $\mathbf{P_2}$ by using the concept of the Fubini-Study distance [27]:

$$\mathbf{P_2}: \arg \max_{\left\{m, \mathbf{F}_{k,m}^{\mathrm{R}}, \mathbf{F}_{k,m}^{\mathrm{B}}\right\}} \left\{ \left\| \left(\mathbf{F}_{k,m}^{\mathrm{opt}}\right)^{\dagger} \mathbf{F}_{k,m}^{\mathrm{R}} \mathbf{F}_{k,m}^{\mathrm{B}} \right\| \right\} \tag{12}$$

subject to constraints C1 and C2 given in Eq. (3), where $\|\cdot\|$ denotes the Euclidean norm. To solve the maximization problem $\mathbf{P_2}$, we can formulate an equivalent optimization problem to minimize the Frobenius norm of the error between the two precoders. Accordingly, the overall error minimization problem for designing an optimal beam-training codebook for our proposed cell-free mmWave m-MIMO schemes can be reconstructed as follows:

$$\mathbf{P_3}: \arg \min_{\left\{m, \mathbf{F}_{k,m}^{\mathrm{R}}, \mathbf{F}_{k,m}^{\mathrm{B}}\right\}} \left\{ \left\| \mathbf{F}_{k,m}^{\mathrm{opt}} - \mathbf{F}_{k,m}^{\mathrm{R}} \mathbf{F}_{k,m}^{\mathrm{B}} \right\|_F \right\} \tag{13}$$

subject to constraints C1 and C2 given in Eq. (3). Correspondingly, using the dictionary learning approach, we can reformulate problem $\mathbf{P_3}$ into the following minimization problem $\mathbf{P_4}$:

$$\begin{aligned} \mathbf{P_4}: \arg &\min_{\left\{m, \mathbf{F}_{k,m}^{\mathrm{R}}, \mathbf{F}_{k,m}^{\mathrm{B}}\right\}} \left\{ \left\| \mathbf{F}_{k,m}^{\mathrm{opt}} - \mathbf{D}_{k,m} \mathbf{F}_{k,m}^{\mathrm{B}} \right\|_F \right\} && (14) \\ \text{s.t. C3: } &\left\| \mathbf{F}_{k,m}^{\mathrm{B}} \left(\mathbf{F}_{k,m}^{\mathrm{B}}\right)^{\dagger} \right\|_0 = L_{\mathrm{T}}, \quad \forall m; \\ \text{C4: } &\left\| \mathbf{D}_{k,m} \mathbf{F}_{k,m}^{\mathrm{B}} \right\|_F^2 = 1, \quad \forall m, && (15) \end{aligned}$$

where $\|\mathbf{M}\|_0$ represents the ℓ_0-pseudo-norm that counts the number of non-zero entries in matrix $\mathbf{M}$. For our dictionary learning based beam-training algorithm, our goal is to minimize the object function given in Eq. (14) iteratively. To solve the minimization problem $\mathbf{P_4}$, we need to proceed with the following stages. 1) Sparse coding stage: We fix the dictionary $\mathbf{D}_{k,m}$ and find the best matrix $\mathbf{F}_{k,m}^{\mathrm{B}}$ by applying any suitable approximation pursuit method. In this paper, we apply the orthogonal matching pursuit (OMP) algorithm [28]. 2) Dictionary update stage: In the dictionary update stage, the algorithm searches for a better dictionary by updating one column at a time. During each iteration, all columns in $\mathbf{D}_{k,m}$ is fixed except $[\mathbf{D}_{k,m}]_{:,i}$, where $[\mathbf{D}_{k,m}]_{:,i}$ represents the ith column of matrix $\mathbf{D}_{k,m}$. Also, $[\mathbf{D}_{k,m}]_{:,i}$ and the corresponding matrix $\mathbf{F}_{k,m}^{\mathrm{B}}$ are updated for achieving the minimum overall representation error in the optimization problem $\mathbf{P_4}$ at the end of each iteration. We define $\mathbf{D}_{k,m}^{(\ell)}$ and $\mathbf{F}_{k,m}^{(\ell,\mathrm{B})}$ as the updated dictionary and the sparse representation matrix after ℓth iteration, respectively. **Algorithm 1** is the pseudo-code outlining our proposed dictionary learning based beam-training algorithm.

Algorithm 1 Dictionary Learning Based Beam-Training Algorithm

```
Input: $K_{\mathrm{u}}, L_{\mathrm{T}}$, and $N_{\mathrm{T}}$
Initialization: Set the initial dictionary $\mathbf{D}_{k,m}^{(0)} = \widetilde{\boldsymbol{\mathcal{F}}}_c \in \mathbb{R}^{N_{\mathrm{T}} \times M_{\mathrm{T}}}$, $\mathbf{F}_{k,m}^{\mathrm{R}} = \emptyset, \mathcal{G}_{\mathrm{k}} = \emptyset$, and $\ell = 1$
Dictionary Learning:
repeat
  Sparse coding stage:
  for $m = 1 : K_{\mathrm{u}}$ do
    Use OMP method to determine $\mathbf{F}_{k,m}^{(\ell,\mathrm{B})}$ for each example $\left[\mathbf{D}_{k,m}^{(\ell)}\right]_{:,i}$
    Select the best matrix $\mathbf{F}_{k,m}^{(\ell,\mathrm{B})}$ and the strongest user by solving $\arg \min_{\{m, \mathbf{F}_{k,m}^{(\ell,\mathrm{B})}\}} \left\{ \left\| \mathbf{F}_{k,m}^{\mathrm{opt}} - \mathbf{D}_{k,m}^{(\ell)} \mathbf{F}_{k,m}^{(\ell,\mathrm{B})} \right\|_F \right\}$
    $\mathcal{G}_{\mathrm{k}} = \mathcal{G}_{\mathrm{k}} \bigcup \{m^{\mathrm{opt}}\}$
  end for
  Dictionary update stage:
  for $i = 1 : M_{\mathrm{T}}$ do
    Update $\mathbf{F}_{k,m}^{(\ell,\mathrm{B})}$ and $\left[\mathbf{D}_{k,m}^{(\ell)}\right]_{:,i}$ using Eq. (14)
  end for
  Set $\ell \leftarrow \ell + 1$
until convergence
```

Assume that the sparse coding stage is perfectly conducted, we can retrieve the best approximations to $\mathbf{F}_{k,m}^{\mathrm{R}}$ that contains no more than e_0 non-zero entries. In this case, when fixing the dictionary $\mathbf{D}_{k,m}$, the overall representation error given in Eq. (14) will be decreased after each iteration. In addition, during the dictionary update stage, an additional reduction or no change in the overall representation error is guaranteed, while not violating the constraints. As a result, such series of iterations ensures a monotonic overall representation error reduction, which indicates that the convergence to a local minimum is guaranteed.

IV. The HARQ-IR Protocol in the Finite Blocklength Regime

Unlike the traditional ARQ protocol, HARQ protocol enables the receiver to exploit the received information from previous HARQ transmission rounds to increase the successful decoding probability of a data packet. Define M_m as the total bits of data packet that is intended to be transmitted to mobile user m. Each finite-blocklength codeword with length n_{d} is divided into L equal-length blocks with $\widehat{n}$ symbols each for implementing HARQ-IR protocol, i.e., $n_{\mathrm{d}} = L\widehat{n}$ (see Fig. 3), and will be transmitted consecutively in the following time slots. Accordingly, we define the codeword with L finite-blocklength data blocks as $\mathbf{X}_m^{n_{\mathrm{d}}} \triangleq \left[\boldsymbol{x}_m^{(1)}, \ldots, \boldsymbol{x}_m^{(L)}\right]$. Denote by N_l the number of HARQ-IR retransmissions, where $1 \leq N_l \leq L$. Under HARQ-IR protocol, if the received data packet can be successfully decoded at the receiver, an ACK will be sent back to the transmitter, and the corresponding data packet will be removed from buffer. Otherwise, an NACK is sent back to the transmitter and another data block will be transmitted until the codeword is successfully decoded at the receiver or the maximum number of transmissions for the packet is reached. If a data packet cannot be correctly decoded at the end of the Lth HARQ retransmission round, it will be discarded from buffer at the transmitter due to transmission delay bound violation. In this section, by using the suboptimal beam-training precoder design given by **Algorithm 1** in Section III, we can characterize the channel capacity as well as the error probability using our proposed HARQ-IR based mmWave user-centric cell-free m-MIMO system models.

Definition 1 ($(\widehat{n}, N_l, M_m, \epsilon_m)$-code): We define a message set $\mathcal{M}_m = \{1, \ldots, M_m\}$ and a message $W \in \mathcal{M}_m$ which is uniformly distributed on $\mathcal{M}_m$. Under HARQ-IR protocol, we define an $(\widehat{n}, N_l, M_m, \epsilon_m)$-code ($\epsilon_m \in [0, 1)$) as follows:

- An encoder Υ: $\{1, \ldots, M_m\} \mapsto \mathbb{C}^{N_{\mathrm{T}} \times \widehat{n}N_l}$ that maps the message $W \in \{1, \ldots, M_m\}$ to a codeword $\mathbf{X}_m^{\widehat{n}N_l}$ with length $\widehat{n}N_l$ which satisfies the following maximum power constraint:

$$\left\|\mathbf{X}_m^{\widehat{n}N_l}\right\|^2 \leq \sqrt{\widehat{n}N_l\overline{\mathcal{P}}_m} \tag{16}$$

  where $\overline{\mathcal{P}}_m$ denotes the average transmit power for mobile user m.
- A decoder $\mathcal{D}$: A decoder $\left\{\mathcal{D}_{\boldsymbol{h}_{k,m}}\right\}_{\boldsymbol{h}_{k,m} \in \mathbb{C}^{1 \times N_{\mathrm{T}}}}$: $\mathbb{C}^{1 \times N_{\mathrm{T}}} \times \mathbb{C}^{N_{\mathrm{T}} \times \widehat{n}N_l} \mapsto \{1, \ldots, M_m\} \bigcup \{e\}$, where e represents the error event.

We apply the *threshold decoding rule* [29], i.e., $\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right) > \beta_m$, where $\beta_m \triangleq \log_2 \frac{M_m - 1}{2}$ denotes the decoding threshold and $\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right)$ is defined as the *information density* for the codeword of finite blocklength $\widehat{n}N_l$, which can be expressed as in the following equation:

$$\begin{aligned}\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right) &= \frac{1}{\widehat{n}N_l N_{\mathrm{T}}} \sum_{j=1}^{\widehat{n}N_l} \boldsymbol{i}_{m,j} \\ &\triangleq \frac{1}{\widehat{n}N_l N_{\mathrm{T}}} \sum_{j=1}^{\widehat{n}N_l} \log_2 \frac{P_{y_m^{(j)}|\widetilde{\boldsymbol{h}}_{k,m}, \boldsymbol{x}_m^{(j)}}\left(y_m^{(j)}|\widetilde{\boldsymbol{h}}_{k,m}, \boldsymbol{x}_m^{(j)}\right)}{P_{y_m^{(j)}|\widetilde{\boldsymbol{h}}_{k,m}}\left(y_m^{(j)}|\widetilde{\boldsymbol{h}}_{k,m}\right)}\end{aligned} \tag{17}$$

where $\boldsymbol{y}_m^{\widehat{n}N_l}$ is the received signal with length $\widehat{n}N_l$, $P_{y_m^{(j)}|\widetilde{\boldsymbol{h}}_{k,m}, \boldsymbol{x}_m^{(j)}}$ and $P_{y_m^{(j)}|\widetilde{\boldsymbol{h}}_{k,m}}$ denote the conditional probabilities and $\boldsymbol{i}_{m,j}$ denotes the random variable with the same distribution of the information density $\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right)$. In addition, using HARQ-IR protocol, we can define the initial transmission rate, denoted by $R_{m,\mathrm{in}}$, for mobile user m as follows:

$$R_{m,\mathrm{in}} \triangleq \frac{\log_2 M_m}{\widehat{n}} \text{ bits/channel use.} \tag{18}$$

Accordingly, the data transmission rate, denoted by R_{m,N_l}, at the end of N_lth HARQ-IR retransmission to mobile user m can be defined as follows:

$$R_{m,N_l} \triangleq \frac{\log_2 M_m}{\widehat{n}N_l} = \frac{R_{m,\mathrm{in}}}{N_l} \text{ bits/channel use.} \tag{19}$$

Under the dependence testing (DT) bound, previous results [30] have shown that there exists an (n, M_m, ϵ_m)-code and average error probability, denoted by ϵ_m, not exceeding the following constraint:

$$\begin{aligned}\epsilon_m \leq \mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\Bigg[&P_{Y_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}, X_m^{\widehat{n}N_l}}\Bigg(\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right) \\ &< \log_2\left(\frac{M_m - 1}{2}\right)\Bigg) \\ &+ \frac{M_m - 1}{2} P_{\overline{Y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\Bigg(\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \overline{\boldsymbol{y}}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right) \\ &> \log_2\left(\frac{M_m - 1}{2}\right)\Bigg)\Bigg]\end{aligned} \tag{20}$$

where $P_{Y_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}, X_m^{\widehat{n}l}}$ and $P_{\overline{Y}_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}}$ denote the conditional probabilities and $\overline{\boldsymbol{y}}_m^{\widehat{n}N_l}$ follows the same distribution as the output signal $\boldsymbol{y}_m^{\widehat{n}N_l}$ and is independent of the input signal $\mathbf{X}_m^{\widehat{n}N_l}$. When calculating the average decoding error probability, we consider two different error events, i.e., miss-detection error and confusion error. Using the Berry-Esseen Theorem [31], we can derive the miss-detection error probability which is given in the first term on the right-hand side in Eq. (20) as follows:

$$\begin{aligned}&P_{Y_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}, X_m^{\widehat{n}l}}\left(\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right) < \log_2\left(\frac{M_m - 1}{2}\right)\right) \\ &= Q\left(\frac{\widehat{n}N_l C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right) - \log_2\left(\frac{M_m - 1}{2}\right)}{\sqrt{\widehat{n}N_l V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)}}\right. \\ &\qquad \left. -2\left(\frac{\log_2 2}{\sqrt{2\pi}} + B_1\right)\frac{1}{\sqrt{\widehat{n}N_l}}\right)\end{aligned} \tag{21}$$

where $C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ denotes the channel capacity, $V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ is the channel dispersion, and

$$B_1 = \frac{6\mathcal{S}\left[\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l}; \boldsymbol{y}_m^{\widehat{n}N_l}, \widetilde{\boldsymbol{h}}_{k,m}\right)\right]}{\left[V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)\right]^{\frac{3}{2}}} \tag{22}$$

where $\mathcal{S}[\cdot]$ is the third moment operator. Furthermore, according to [30], we can obtain the following confusion error probability which is given in the second term on the right-hand side in Eq. (20):

$$
\begin{aligned}
&\frac{M_m-1}{2}P_{\overline{Y}_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right)\geq\log_2\left(\frac{M_m-1}{2}\right)\right)\\
&=\mathbb{E}_{P_{\overline{Y}_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}}}\left[\exp\left\{-\sum_{j=1}^{\widehat{n}N_l}\boldsymbol{i}_{m,j}\right\}\mathbb{1}_{\left\{\sum_{j=1}^{\widehat{n}N_l}\boldsymbol{i}_{m,j}>\log_2\left(\frac{M_m-1}{2}\right)\right\}}\right]\\
&\leq 2\left(\frac{\log_2 2}{\sqrt{2\pi}}+B_2\right)\frac{1}{\sqrt{\widehat{n}N_l}}
\end{aligned}
\tag{23}
$$

where $\mathbb{E}_{P_{\overline{Y}_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}}}[\cdot]$ is the expectation with respect to $P_{\overline{Y}_m^{\widehat{n}l}|\widetilde{\boldsymbol{h}}_{k,m}}$, $\mathbb{1}_{\{\mathcal{A}\}}$ is the indicator function of the hypothesis test $\mathcal{A}$'s result, and

$$
\begin{aligned}
B_2=\ &\frac{1}{V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)}\Big\{12\mathcal{S}\Big[\Big|\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right)\\
&\qquad-\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right)\right]\Big|\Big]\Big\}.
\end{aligned}
\tag{24}
$$

Correspondingly, we can obtain the average decoding error probability, denoted by ϵ_{m,N_l}, for data packets with length $\widehat{n}N_l$ under HARQ-IR protocol over mmWave cell-free m-MIMO based 6G multimedia mobile wireless networks as follows:

$$
\begin{aligned}
\epsilon_{m,N_l}&\leq Q\left(\frac{\widehat{n}N_lC_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)-\log_2\left(\frac{M_m-1}{2}\right)}{\sqrt{\widehat{n}N_lV_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)}}\right)+\frac{B_1+B_2}{\sqrt{\widehat{n}N_l}}\\
&\approx Q\left(\frac{\widehat{n}N_lC_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)-\log_2\left(\frac{M_m-1}{2}\right)}{\sqrt{nK/N_{\mathrm{T}}}}\right)+\frac{B_1+B_2}{\sqrt{\widehat{n}N_l}}.
\end{aligned}
\tag{25}
$$

Using Taylor's expansion for inverse Q function, we can derive an upper bound on the finite blocklength coding rate for our proposed mmWave cell-free m-MIMO schemes as in the following equations:

$$
\begin{aligned}
\frac{\log_2 M_m}{\widehat{n}N_l}&\leq C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)-\sqrt{\frac{V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)}{\widehat{n}N_l}}Q^{-1}\left(\epsilon_{m,N_l}-\frac{B_1+B_2}{\widehat{n}N_l}\right)\\
&\approx C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)-\sqrt{\frac{V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)}{\widehat{n}N_l}}Q^{-1}(\epsilon_{m,N_l})+O\left(\frac{1}{\sqrt{\widehat{n}N_l}}\right)
\end{aligned}
\tag{26}
$$

where $f(x)=O(g(x))$ if and only if there exists a positive real number M and a real number x_0 such that $|f(x)|\leq Mg(x)$ for all $x\geq x_0$. As a result, for our proposed $(\widehat{n},N_l,M_m,\epsilon_m)$-code, we can derive the approximate decoding error probability for transmitting data packet of length $\widehat{n}N_l$ to mobile user m under perfect CSI as follows:

$$
\epsilon_{m,N_l}\approx Q\left(\sqrt{\frac{\widehat{n}}{N_lV_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)}}\left(N_lC_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)-R_{m,\text{in}}\right)\right)
\tag{27}
$$

where $R_{m,\text{in}}$ is the initial data transmission rate for mobile user m, specified by Eq. (18).

Considering the non-vanishing error probability, it is challenging to derive the closed-form expression of the channel capacity for our proposed mmWave user-centric cell-free m-MIMO schemes compared with the traditional m-MIMO schemes. In the following theorem, we give the concrete expression to derive an lower bound on the channel capacity $C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime.

Theorem 1: The lower bound on the channel capacity $C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ for our proposed mmWave user-centric cell-free m-MIMO and FBC-HARQ-IR based 6G mobile wireless networks is given as follows:

$$
\begin{aligned}
&C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)\\
&\quad\geq\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\Bigg[\log_2\Bigg\{\det\Bigg(\left(\left(\boldsymbol{\Xi}_{k,m}\right)^{-1}-\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\right)\\
&\qquad+\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\widetilde{\boldsymbol{h}}_{k,m}\Bigg)\Bigg\}\Bigg]\\
&\qquad-\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\log_2\left\{\det\left(\left(\boldsymbol{\Xi}_{k,m}\right)^{-1}-\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\right)\right\}\right]
\end{aligned}
\tag{28}
$$

where $\det(\cdot)$ is the determinant of a matrix and

$$
\begin{aligned}
\boldsymbol{\Xi}_{k,m}\triangleq\Bigg\{&\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\Bigg[\boldsymbol{h}_{k,m}\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\mathbf{F}_{k,m'}^{\mathrm{R}}\mathbf{F}_{k,m'}^{\mathrm{B}}\boldsymbol{\Omega}_{k,m'}\\
&\times\left(\mathbf{F}_{k,m'}^{\mathrm{R}}\mathbf{F}_{k,m'}^{\mathrm{B}}\right)^{\dagger}\left(\boldsymbol{h}_{k,m}\right)^{\dagger}\Bigg]+\sigma^2\Bigg\}^{-1}.
\end{aligned}
\tag{29}
$$

Proof: To derive the lower bound on the channel capacity, we first need to analyze the mutual information $I\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)$ as follows [32]:

$$
I\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)=H\left(\mathbf{X}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)-H\left(\mathbf{X}_m^{\widehat{n}N_l}|\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right)
\tag{30}
$$

where $H(\cdot)$ represents the function of information entropy. We can derive $H\left(\mathbf{X}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)$ in Eq. (30) as follows:

$$
H\left(\mathbf{X}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)=\log_2\left(\pi e\mathbf{I}_{N_T}\right).
\tag{31}
$$

Then, define $\widehat{\mathbf{X}}_m^{\widehat{n}N_l}$ as the linear MMSE estimate of $\mathbf{X}_m^{\widehat{n}N_l}$ given $\boldsymbol{y}_m^{\widehat{n}N_l}$ and $\widetilde{\boldsymbol{h}}_{k,m}$. Correspondingly, using the suboptimal beam-training precoders $\mathbf{F}_{k,m}^{\mathrm{R}}$ and $\mathbf{F}_{k,m}^{\mathrm{B}}$ derived in **Algorithm 1** in Section III, we can obtain the following equation:

$$
\widehat{\mathbf{X}}_m^{\widehat{n}N_l}=\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\boldsymbol{\Xi}_{k,m}\boldsymbol{y}_m^{\widehat{n}N_l}
\tag{32}
$$

where

$$
\begin{aligned}
\boldsymbol{\Xi}_{k,m}\triangleq\Bigg\{&\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\Bigg[\boldsymbol{h}_{k,m}\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\mathbf{F}_{k,m'}^{\mathrm{R}}\mathbf{F}_{k,m'}^{\mathrm{B}}\boldsymbol{\Omega}_{k,m'}\\
&\times\left(\mathbf{F}_{k,m'}^{\mathrm{R}}\mathbf{F}_{k,m'}^{\mathrm{B}}\right)^{\dagger}\left(\boldsymbol{h}_{k,m}\right)^{\dagger}\Bigg]+\sigma^2\Bigg\}^{-1}.
\end{aligned}
\tag{33}
$$

Correspondingly, we can derive an upper bound on the conditional entropy $H\left(\mathbf{X}_m^{\widehat{n}N_l}|\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right)$ as in the following equation:

$$\begin{aligned} &H\left(\mathbf{X}_m^{\widehat{n}N_l}|\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right) \\ &\quad\leq \mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left(\mathbf{X}_m^{\widehat{n}N_l}-\widehat{\mathbf{X}}_m^{\widehat{n}N_l}\right)\left(\mathbf{X}_m^{\widehat{n}N_l}-\widehat{\mathbf{X}}_m^{\widehat{n}N_l}\right)^{\dagger}\right] \\ &\quad= \log_2\left\{\pi e \det\left(\mathbf{I}_{N_T}-\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\boldsymbol{\Xi}_{k,m}\widetilde{\boldsymbol{h}}_{k,m}\right)\right\} \end{aligned} \tag{34}$$

where $\mathbf{I}_{N_T}$ is the identity matrix of size N_T. Then, plugging Eqs. (31) and (34) back into Eq. (30), we can derive a lower bound on the mutual information $I\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)$ as follows [33]:

$$\begin{aligned} &I\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right) \\ &\quad\geq \log_2\left(\pi e\mathbf{I}_{N_T}\right)-\log_2\left\{\pi e \det\left[\mathbf{I}_{N_T}-\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\boldsymbol{\Xi}_{k,m}\widetilde{\boldsymbol{h}}_{k,m}\right]\right\} \\ &\quad\geq \log_2\left\{\det\left[\mathbf{I}_{N_T}+\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\left(\left(\boldsymbol{\Xi}_{k,m}\right)^{-1}-\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\right)^{-1}\right.\right. \\ &\qquad\qquad\left.\left.\times\widetilde{\boldsymbol{h}}_{k,m}\right]\right\}. \end{aligned} \tag{35}$$

Accordingly, we can derive a lower bound on the channel capacity as follows:

$$\begin{aligned} &C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)=\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\max_{p_{\mathbf{X}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}}\left\{I\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)\right\}\right] \\ &\geq \mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\log_2\left\{\det\left[\mathbf{I}_{N_T}+\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\left(\left(\boldsymbol{\Xi}_{k,m}\right)^{-1}\right.\right.\right.\right. \\ &\quad\left.\left.\left.\left.-\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\right)^{-1}\widetilde{\boldsymbol{h}}_{k,m}\right]\right\}\right] \\ &=\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\log_2\left\{\det\left[\left(\boldsymbol{\Xi}_{k,m}\right)^{-1}-\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}+\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\right.\right.\right. \\ &\quad\left.\left.\left.\times\widetilde{\boldsymbol{h}}_{k,m}\right]\right\}\right]-\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\log_2\left\{\det\left(\left(\boldsymbol{\Xi}_{k,m}\right)^{-1}-\widetilde{\boldsymbol{h}}_{k,m}\left(\widetilde{\boldsymbol{h}}_{k,m}\right)^{\dagger}\right)\right\}\right] \end{aligned} \tag{36}$$

which is a lower bound on the channel capacity $C_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ as shown in Eq. (28). Therefore, we complete the proof of Theorem 1. ■

Traditionally, it is challenging to derive the closed-form expression of the channel dispersion for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime. Using the suboptimal beam-training precoder design given by **Algorithm 1** in Section III, we can derive an upper bound on the channel dispersion $V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ for our proposed FBC-HARQ based mmWave user-centric cell-free m-MIMO scheme as summarized in the following theorem.

Theorem 2: The upper bound on the channel dispersion $V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ for our proposed mmWave user-centric cell-free m-MIMO and FBC-HARQ-IR based 6G mobile wireless networks is given as follows:

$$\begin{aligned} V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)\leq 8\widehat{n}N_l&\left\{\frac{3\overline{\mathcal{P}}_m}{\sigma^4}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\left(\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\right.\right.\right.\right. \\ &\left.\left.\left.\left.\times\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right)\right\|^2\right]+2\right\}. \end{aligned} \tag{37}$$

Proof: To derive the upper bound on the channel dispersion $V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$, we need to proceed with the following steps. First, we start with variance as in the following equation:

$$\begin{aligned} &\mathrm{Var}\left[\boldsymbol{i}\left(\mathbf{X}_m^{\widehat{n}N_l};\boldsymbol{y}_m^{\widehat{n}N_l},\widetilde{\boldsymbol{h}}_{k,m}\right)\right] \\ &\quad=\mathrm{Var}\left[\log_2\left(\frac{P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}\right)}{P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)}\right)\right] \\ &\quad\leq 2\left(\mathrm{Var}\left[\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}\right)\right)\right]\right. \\ &\qquad\left.+\mathrm{Var}\left[\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)\right)\right]\right) \end{aligned} \tag{38}$$

where $\mathrm{Var}[\cdot]$ represents the variance and $P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}}$ and $P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}$ are the conditional probabilities. Second, using the suboptimal beam-training precoder $\mathbf{F}_{k,m}^{\mathrm{B}}$ derived in **Algorithm 1** in Section III, we can apply the Poincará inequality to derive the following equation:

$$\begin{aligned} &\mathrm{Var}\left[\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)\right)\right] \\ &\quad\leq\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\nabla\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)\right)\right\|^2\middle|\mathbf{X}_m^{\widehat{n}N_l}\right] \\ &\quad=\frac{1}{\sigma^2}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right.\right.\right. \\ &\qquad\left.\left.\left.\times\mathbf{X}_{m'}^{\widehat{n}N_l}\middle|\boldsymbol{y}_m^{\widehat{n}N_l}\right]-\boldsymbol{y}_m^{\widehat{n}N_l}\right\|^2\middle|\mathbf{X}_m^{\widehat{n}N_l}\right]. \end{aligned} \tag{39}$$

where ∇ is the Nabla operator. Third, we define

$$\begin{aligned} &\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\widehat{\mathbf{X}}_{m'}^{\widehat{n}N_l}\right] \\ &\quad\triangleq\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\mathbf{X}_{m'}^{\widehat{n}N_l}\middle|\boldsymbol{y}_m^{\widehat{n}N_l}\right]. \end{aligned} \tag{40}$$

Accordingly, we can have:

$$\begin{aligned} &\mathrm{Var}\left[\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)\right)\right] \\ &\quad\leq\frac{1}{\sigma^2}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\boldsymbol{y}_m^{\widehat{n}N_l}-\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right.\right. \\ &\qquad\left.\left.\times\widehat{\mathbf{X}}_{m'}^{\widehat{n}N_l}\right\|^2\middle|\mathbf{X}_m^{\widehat{n}N_l}\right] \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{\sigma^2}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\boldsymbol{y}_m^{\widehat{n}N_l}\middle|\mathbf{X}_m^{\widehat{n}N_l}\right\|^2 \middle| \mathbf{X}_m^{\widehat{n}N_l}\right] + \frac{2}{\sigma^2} \\
&\times \mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\widehat{\mathbf{X}}_{m'}^{\widehat{n}N_l}\right\|^2\middle|\mathbf{X}_m^{\widehat{n}N_l}\right] \\
&\leq \frac{6\widehat{n}N_l\overline{\mathcal{P}}_m}{\sigma^4}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right\|^2\right] \\
&\quad +4\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\frac{\boldsymbol{n}_m}{\sigma^2}\right\|^2\middle|\mathbf{X}_m^{\widehat{n}N_l}\right] \\
&= \frac{6\widehat{n}N_l\overline{\mathcal{P}}_m}{\sigma^4}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right\|^2\right] \\
&\quad +4\widehat{n}N_l. \qquad (41)
\end{aligned}$$

Following the similar procedures for obtaining $\mathrm{Var}\left[\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m}\right)\right)\right]$ in Eq. (41), we can derive an upper bound on $\mathrm{Var}\Big[\log_2\Big(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}}$ $\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}\right)\Big)\Big]$ for all $\mathbf{X}_m^{\widehat{n}N_l}$ as follows:

$$\begin{aligned}
&\mathrm{Var}\left[\log_2\left(P_{\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}}\left(\boldsymbol{y}_m^{\widehat{n}N_l}|\widetilde{\boldsymbol{h}}_{k,m},\mathbf{X}_m^{\widehat{n}N_l}\right)\right)\right] \\
&\leq \frac{6\widehat{n}N_l\overline{\mathcal{P}}_m}{\sigma^4}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\left[\left\|\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}}\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\right\|^2\right] \\
&\quad +4\widehat{n}N_l \qquad (42)
\end{aligned}$$

Finally, substituting Eqs. (41) and (42) back into Eq. (38), we can derive an upper bound on the channel dispersion $V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right)$ as follows:

$$\begin{aligned}
V_m\left(\widetilde{\boldsymbol{h}}_{k,m}\right) \leq 8\widehat{n}N_l\Bigg(&\frac{3\overline{\mathcal{P}}_m}{\sigma^4}\mathbb{E}_{\widetilde{\boldsymbol{h}}_{k,m}}\Bigg[\Bigg\|\sum_{m'=1}^{K_{\mathrm{u}}}\sum_{k\in\mathcal{K}(m')}\widetilde{\boldsymbol{h}}_{k,m}\mathbf{F}_{k,m'}^{\mathrm{B}} \\
&\times\left(\boldsymbol{\Omega}_{k,m'}\right)^{\frac{1}{2}}\Bigg\|^2\Bigg] + 2\Bigg) \qquad (43)
\end{aligned}$$

which is Eq. (37), completing the proof of Theorem 2. ■

V. Effective Capacity For Statistical Delay and Error-Rate Bounded QoS Provisioning in the Finite Blocklength Regime

In this section, by using the decode error probability function given in Eq. (27) in the previous Section IV, we can then characterize the analytical relationship between the statistical delay and error-rate bounded QoS metrics/schemes and decode error probability function. In addition, we derive the corresponding effective capacity function under HARQ-IR protocol for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime.

A. Statistical Delay and Error-Rate Bounded QoS Metrics Under Constant Arrival Rate in the Finite Blocklength Regime

Statistical delay-bounded QoS guarantees [34] have been extensively studied for analyzing queuing behavior for time-varying arrival and service processes. Traditionally, the effective capacity measures queue-length process which is independent of the decoding error at the receiver. We measure the delay-bounded QoS requirements for mURLLC traffics under HARQ-IR protocol in the finite blocklength regime as follows: a data packet of size $\widehat{n}N_l$ bits can be successfully delivered to the receiver within a queueing delay of no more than $D_{m,\mathrm{th}}$ symbols with a probability of at least $(1-\epsilon_{\mathrm{q},m})$, where $\epsilon_{\mathrm{q},m}$ is the delay violation probability for mobile user m.

Based on the Large Deviation Principle (LDP) [35], under sufficient conditions, the queue-length process Q_m converges in distribution to a random variable $Q_m(\infty)$ such that

$$-\lim_{Q_{m,\mathrm{th}}\to\infty}\frac{\log\left(\Pr\left\{Q_m(\infty)>Q_{m,\mathrm{th}}\right\}\right)}{Q_{m,\mathrm{th}}}=\theta_m \qquad (44)$$

where $Q_{m,\mathrm{th}}$ represents the buffer-size overflow threshold at mobile user m and θ_m is defined as the *QoS exponent* for mobile user m. To be more specific, Eq. (44) states that the probability of the queue-length process exceeding a certain threshold $Q_{m,\mathrm{th}}$ decays exponentially fast as the threshold $Q_{m,\mathrm{th}}$ increases. As shown in [2], the larger θ_m corresponds to the more stringent QoS requirement, while the smaller θ_m leads to the looser delay constraint, which implies the system can only provide a looser QoS guarantee.

Due to the non-vanishing error probability cased by the finite blocklength data transmissions, the traditional queuing behavior and *effective capacity* measurement approaches are no longer appropriate for our proposed FBC-HARQ-IR based mmWave user-centric cell-free m-MIMO schemes. As a result, we need to derive new analytical model for characterizing QoS metrics/schemes and the corresponding *effective capacity function* under statistical delay and error-rate bounded QoS constraints. For our proposed HARQ-IR protocol, we define $a_{k,m}(N_l)$ as the amount of bits generated at the end of N_lth HARQ-IR retransmission from the kth AP to mobile user m and $s_{k,m}(N_l)$ as the instantaneous data transmission rate over wireless channels at the end of N_lth HARQ-IR retransmission from the k AP to mobile user m. Define $A_m(N_l)=\sum_{k\in\mathcal{K}(m)}\sum_{j=0}^{N_l-1}a_{k,m}(j)$ as the accumulated source rate at the end of N_lth HARQ-IR retransmission to mobile user m and $S_m(N_l)=\sum_{k\in\mathcal{K}(m)}\sum_{j=0}^{N_l-1}s_{k,m}(j)$ as the accumulated data transmission rate over wireless channels at the end of N_lth HARQ-IR retransmission to mobile user m. Define $Q_m(N_l)$ as the dynamics of queue-length process at the end of N_lth HARQ-IR retransmission to mobile user m, which is given as in the following equation:

$$Q_m(N_l)=\max\left\{A_m(N_l)-S_m(N_l),0\right\}. \qquad (45)$$

Define $U_m(N_l) \triangleq A_m(N_l) - S_m(N_l)$. We can rewrite the queue-length process at the end of N_lth HARQ-IR retransmission to mobile user m as in the following equation:

$$Q_m(N_l) = \max\left\{0, U_m(N_l), U_m(N_l) + U_m(N_l-1), \dots\right\}. \tag{46}$$

Assume that the average data arrival rate, denoted by $\overline{\mu}_{k,m}$, is a constant. Then, given the decode error probability function ϵ_{m,N_l} in Eq. (27) in Section IV, the delay-bounded QoS constraint for mobile user m at the end of N_lth HARQ-IR retransmission can be rewritten as follows:

$$\begin{aligned}\Pr\left\{\bigcup_{N_l}\left\{Q_m(N_l) > Q_{m,\text{th}}\right\}\right\} &= \Pr\left\{\bigcup_{N_l}\left\{\sum_{j=1}^{N_l} U_m(j) > Q_{m,\text{th}}\right\}\right\} \\ &\approx \eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right) e^{-\theta_m^{\text{con}}\left(\mu_{k,m},\epsilon_{m,N_l}\right) Q_{m,\text{th}}}\end{aligned} \tag{47}$$

where $\bigcup$ is the "or" operation, $\theta_m^{\text{con}}\left(\mu_{k,m}, \epsilon_{m,N_l}\right)$ is defined as the QoS exponent function considering the constant average arrival rate scenario, and $\eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right)$ is the probability that queue is non-empty. Note that the pair of functions $\left\{\eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right), \theta_m^{\text{con}}\left(\mu_{k,m}, \epsilon_{m,N_l}\right)\right\}$ are functions of the source rate $\mu_{k,m}$ and the decoding error probability ϵ_{m,N_l}, which depend on the channel condition and decoding processes.

On the other hand, we can characterize the queuing delay of the buffer at the end of N_lth HARQ-IR retransmission as $D_m(N_l)$. First, using Eq. (47), we can derive the bound on the steady-state delay distribution in terms of the delay violation probability, denoted by $\epsilon_{\text{q},m}$, given the decode error probability function ϵ_{m,N_l} in Eq. (27) in Section IV as follows:

$$\begin{aligned}\epsilon_{\text{q},m} &= \Pr\left\{\bigcup_{N_l}\left\{D_m(N_l) > D_{m,\text{th}}\right\}\right\} \\ &\approx \eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right) e^{-\theta_m^{\text{con}}\left(\mu_{k,m},\epsilon_{m,N_l}\right) R_{m,N_l} D_{m,\text{th}}}\end{aligned} \tag{48}$$

where $D_{m,\text{th}}$ is the delay bound for mobile user m and R_{m,N_l} denotes data transmission rate at the end of N_lth HARQ-IR retransmission for mobile user m, specified by Eq. (19). ***Remarks:*** *Comparing Eq. (47) with its equivalent Eq. (48), we obtain the following relationships:* $Q_{m,\text{th}} = R_{m,N_l} D_{m,\text{th}}$. Then, based on the derivations in [36], we can characterize the estimated analytical relationship between the functions $\left\{\eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right), \theta_m^{\text{con}}\left(\mu_{k,m}, \epsilon_{m,N_l}\right)\right\}$ as follows:

$$\begin{cases}\dfrac{\eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right)}{\theta_m^{\text{con}}\left(\mu_{k,m}, \epsilon_{m,N_l}\right)} = \mathbb{E}\left[D_m(N_l)\right]; \\ \eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right) = \Pr\left\{Q_m(N_l) > 0\right\}.\end{cases} \tag{49}$$

Accordingly, the expectation of the delay process can be derived as in the following equation:

$$\mathbb{E}\left[D_m(N_l)\right] = \mathbb{E}\left[\frac{A_m(N_l) - S_m(N_l)}{\sum\limits_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}}\right]. \tag{50}$$

Since $A_m(N_l)$ and $S_m(N_l)$ are independent of each other, we can determine the value of $\mathbb{E}\left[A_m(N_l)\right]$ when $N_l \to \infty$. Using the Central Limit Theorem, we get

$$\mathbb{E}\left[A_m(N_l)\right] = \sum_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}. \tag{51}$$

In order to derive the expected value of the accumulated process $S_m(N_l)$, using Eqs. (18) and (19), we have

$$\mathbb{E}\left[S_m(N_l)\right] = \widehat{n}\mathbb{E}\left[\sum_{j=0}^{N_l-1} R_{m,j}\right] = \widehat{n}\mathbb{E}\left[\sum_{j=0}^{N_l-1} \frac{R_{m,\text{in}}}{j}\right] \tag{52}$$

Then, as $N_l \to \infty$, we can obtain the long-term average transmission rate (LATR), denoted by $\mathbb{E}\left[N_l\right]$. Correspondingly, we derive the probability that the number of HARQ-IR retransmission rounds when $N_l = j$ as in the following equation:

$$\Pr\{N_l = j\} = \begin{cases}\Pr\left\{\overline{\mathcal{A}}_0\right\} - \Pr\left\{\overline{\mathcal{A}}_1\right\}, & \text{for } j=1; \\ \Pr\left\{\bigcap\limits_{\iota=1}^{j-1}\left\{\overline{\mathcal{A}}_\iota\right\}\right\} - \Pr\left\{\bigcap\limits_{\iota=1}^{j}\left\{\overline{\mathcal{A}}_\iota\right\}\right\}, & \text{for } 1<j<L; \\ \Pr\left\{\bigcap\limits_{j=1}^{L-1}\left\{\overline{\mathcal{A}}_j\right\}\right\}, & \text{for } j=L,\end{cases} \tag{53}$$

where $\bigcap$ is the "and" operation and $\overline{\mathcal{A}}_j$ $(j = 1, \dots, L)$ denotes the event that the received data packet cannot be decoded at the end of jth HARQ-IR retransmission round. Correspondingly, the LATR can be upper-bounded as follows:

$$\begin{aligned}\mathbb{E}[N_l] &= \sum_{j=1}^{L} j\Pr\left\{N_l = j\right\} = \Pr\left\{\overline{\mathcal{A}}_0\right\} + \sum_{j=1}^{L-1}\Pr\left\{\bigcap_{j=1}^{L-1}\left\{\overline{\mathcal{A}}_j\right\}\right\} \\ &\leq 1 + \sum_{j=1}^{L-1}\Pr\left\{\overline{\mathcal{A}}_j\right\} \approx 1 + \sum_{j=1}^{L-1}\epsilon_{m,j}\end{aligned} \tag{54}$$

where $\Pr\left\{\overline{\mathcal{A}}_0\right\} = 1$ and $\epsilon_{m,j} \approx \Pr\left\{\overline{\mathcal{A}}_j\right\}$ is the approximate decoding error probability after jth HARQ-IR retransmission round to mobile user m given by Eq. (27).

The probability that buffer is non-empty is similar to the probability that the received SINR falls below a certain specified threshold, i.e., the decoding error probability at the receiver [37]. Due to the fact that the non-empty buffer probability also considers the effect of packet accumulation in the queue, the non-empty buffer probability is larger than the decoding error probability, i.e.,

$$\eta_m\left(\mu_{k,m}, \epsilon_{m,N_l}\right) \geq \epsilon_{m,L} \tag{55}$$

where $\epsilon_{m,L}$ is the decoding error probability after L HARQ-IR retransmission rounds for mobile user m, specified by Eq. (27) when $N_l = L$. Using Eqs. (49), (50), (54) and (55), we can obtain the following equation that characterizes the analytical relationship between the decoding error probability and the QoS exponent function considering the constant average data arrival rate scenario:

$$\theta_m^{\text{con}}(\mu_{k,m}, \epsilon_{m,L}) \approx \frac{\epsilon_{m,L}}{1 - R_{m,\text{in}}\left[\sum\limits_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}\left(1 + \sum\limits_{j=1}^{L-1}\epsilon_{m,j}\right)\right]^{-1}} \tag{56}$$

where $R_{m,\text{in}}$ is the initial transmission rate for mobile user m given by Eq. (18). Note that the above Eq. (56) implies that there exists an analytical relationship between the decoding error probability function and the QoS exponent function using FBC-HARQ-IR protocol in the finite blocklength, i.e., given the error-rate constraints, we can them characterize the delay-bounded QoS exponent function $\theta_m^{\text{con}}(\mu_{k,m}, \epsilon_{m,L})$ for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime.

B. Statistical Delay and Error-Rate Bounded QoS Metrics Under Random Arrival Rate in the Finite Blocklength Regime

1) Discrete-Time Markov Model: Consider a two-state discrete-time Markov model for which the transition probability matrix, denoted by $\mathbf{J}_m$, is given as in the following equation:

$$\mathbf{J}_m = \begin{bmatrix} p_{m,11} & p_{m,12} \\ p_{m,21} & p_{m,22} \end{bmatrix} \tag{57}$$

where $p_{m,11}$ and $p_{m,22}$ represent the probabilities that data source remains in the ON state and OFF state, respectively, in the next time slot and $p_{m,12}$ and $p_{m,21}$ are the probabilities of transitioning to a different state in the next time slot. In the OFF state of the discrete-time Markov model, no data arrives from the source, while in the ON state, data arrives at rate $\mu_{k,m}$ for mobile user m. Using the transition probability matrix $\mathbf{J}_m$, we can derive the steady state probability of ON state, denoted by $p_{m,\text{ON}}$, as follows [38]:

$$p_{m,\text{ON}} = \frac{1 - p_{m,11}}{2 - p_{m,11} - p_{m,22}}. \tag{58}$$

Accordingly, we can characterize the average source rate as follows:

$$\mathbb{E}\left[A_m(l)\right] = \frac{\sum\limits_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}(1 - p_{m,11})}{2 - p_{m,11} - p_{m,22}}. \tag{59}$$

Similar to Eq. (56), we can derive the QoS exponent function, denoted by $\theta_m^{\text{DM}}(\mu_{k,m}, \epsilon_{m,L})$, for discrete-time Markov model as follows:

$$\theta_m^{\text{DM}}(\mu_{k,m}, \epsilon_{m,L}) \approx \frac{\epsilon_{m,L}}{1 - \frac{R_{m,\text{in}}(2 - p_{m,11} - p_{m,22})}{\sum\limits_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}(1 - p_{m,11})\left(1 + \sum\limits_{j=1}^{L-1} \epsilon_{m,j}\right)}}. \tag{60}$$

2) Markov Fluid Model: Consider data arrival as a continuous-time Markov process. We can derive the transition rate matrix, denoted by $\mathbf{G}_m$, of data arrival process as follows:

$$\mathbf{G}_m = \begin{bmatrix} -\upsilon_{m,1} & \upsilon_{m,1} \\ \upsilon_{m,2} & -\upsilon_{m,2} \end{bmatrix} \tag{61}$$

where $\upsilon_{m,1} > 0$ and $\upsilon_{m,2} > 0$ are the transition rates between ON state and OFF state. Then, we can derive the steady state probability of ON state as follows:

$$p_{m,\text{ON}} = \frac{\upsilon_{m,1}}{\upsilon_{m,1} + \upsilon_{m,2}}. \tag{62}$$

Accordingly, we can characterize the average source rate as follows:

$$\mathbb{E}\left[A_m(l)\right] = \frac{\sum\limits_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}\upsilon_{m,1}}{\upsilon_{m,1} + \upsilon_{m,2}}. \tag{63}$$

Similarly, we can derive the QoS exponent function, denoted by $\theta_m^{\text{MF}}(\mu_{k,m}, \epsilon_{m,L})$, for Markov Fluid model as follows:

$$\theta_m^{\text{MF}}(\mu_{k,m}, \epsilon_{m,L}) \approx \frac{\epsilon_{m,L}}{1 - \frac{R_{m,\text{in}}(\upsilon_{m,1} + \upsilon_{m,2})}{\sum\limits_{k\in\mathcal{K}(m)} \overline{\mu}_{k,m}\upsilon_{m,1}\left(1 + \sum\limits_{j=1}^{L-1} \epsilon_{m,j}\right)}}. \tag{64}$$

C. Effective Capacity Under HARQ-IR Protocol in the Finite Blocklength Regime

We define the asymptotic log-moment generating function [39], denoted by $\Lambda_{U_m}(\theta_m)$, of $U_m(N_l)$ as follows:

$$\Lambda_{U_m}(\theta_m) \triangleq \lim_{N_l\to\infty} \frac{1}{N_l} \log\left\{\mathbb{E}\left[e^{\theta_m U_m(N_l)}\right]\right\}. \tag{65}$$

Since $a_{k,m}(N_l)$ and $s_{k,m}(N_l)$ are independent of each other, we have $\Lambda_{U_m}(\theta_m) = \Lambda_{A_m}(\theta_m) + \Lambda_{S_m}(-\theta_m)$, where $\Lambda_{A_m}(\theta_m)$ and $\Lambda_{S_m}(\theta_m)$ are the asymptotic log-moment generating functions of the accumulated source process $A_m(N_l)$ and the accumulated channel process $S_m(N_l)$, respectively. For a given QoS exponent θ_m in Eq. (44), the processes $S_m(N_l)$ and $A_m(N_l)$ need to satisfy the following equation:

$$\Lambda_{A_m}(\theta_m) = -\Lambda_{S_m}(-\theta_m). \tag{66}$$

The effective capacity [2] is defined as the maximum constant arrival rate that a given service process can support in order to guarantee a QoS requirement specified by θ_m. Given a service process S_m, the effective capacity of the service process, denoted by $EC_m(\theta_m)$, where $\theta_m > 0$, is defined as follows [40]:

$$EC_m(\theta_m) \triangleq -\frac{\Lambda_{S_m}(-\theta_m)}{\theta_m}. \tag{67}$$

Considering the non-vanishing error probability, it is challenging to derive the closed-form expression of the effective capacity for our proposed mmWave user-centric cell-free m-MIMO schemes using the HARQ-IR protocol. The theorem that follows bellow derives the closed-form expression of the effective capacity $EC_m(\theta_m)$ for our proposed mmWave user-centric cell-free m-MIMO schemes under statistical delay and error-rate bounded QoS constraints in the finite blocklength regime.

Theorem 3: **If** the statistical delay and error-rate bounded QoS constraints are specified by Eqs. (55)-(65), **then** the effective capacity $EC_m(\theta_m)$ for our proposed mmWave user-centric cell-free m-MIMO and FBC-HARQ-IR based 6G mobile wireless networks is given as follows:

$$\begin{aligned} EC_m(\theta_m) = {} & \frac{R_{m,\text{in}}}{2}\left(1+\sum_{j=1}^{L-1}\epsilon_{m,j}\right)^{-1} + \frac{R_{m,\text{in}}}{2} \\ & \times\left\{\left(1+\sum_{j=1}^{L-1}\epsilon_{m,j}\right)^{-2} + \frac{2\widehat{n}\log\left(\epsilon_{\text{q},m}\right)}{D_{m,\text{th}}}\right. \\ & \left.\times\left[\sum_{j=1}^{L-1}(2j-1)\,\epsilon_{m,j} - \left(\sum_{j=1}^{L-1}\epsilon_{m,j}\right)^{2}\right]^{-1}\right\}^{\frac{1}{2}} \end{aligned} \tag{68}$$

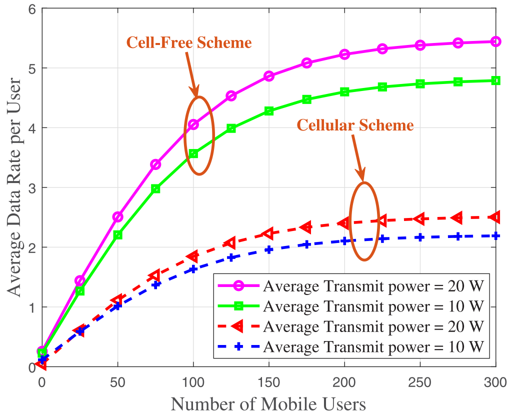


Fig. 4. The average data transmission rate per user vs. number of mobile users K_u over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks in the finite blocklength regime.

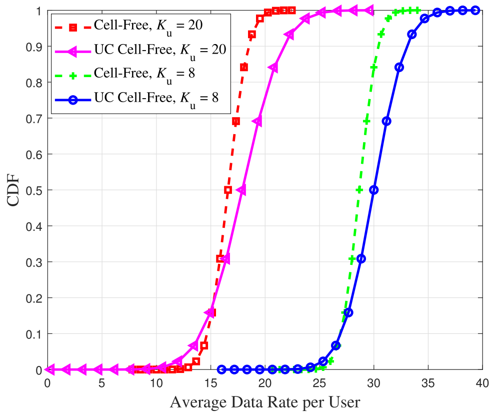


Fig. 5. The CDFs of downlink data transmission rate per user over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks in the finite blocklength regime.

where $R_{m,\text{in}}$ is the initial transmission rate specified by Eq. (18), $\epsilon_{\text{q},m}$ represents the delay violation probability given by Eq. (48), $\epsilon_{m,j}$ is the probability that the received data packet cannot be decoded at the end of jth HARQ-IR retransmission round for mobile user m, which is given in Eq. (54), and $D_{m,\text{th}}$ is the delay bound for mobile user m.

Proof: Given statistical delay and error-rate bounded QoS constraints, we can exploit the definition of the effective capacity given in Eq. (67) and obtain the following equation:

$$\begin{aligned} EC_m(\theta_m) &= -\frac{\Lambda_{S_m}(-\theta_m)}{\theta_m} \\ &= -\lim_{N_l\to\infty}\frac{1}{\theta_m N_l}\log\left\{\mathbb{E}\left[e^{-\theta_m \sum\limits_{k\in\mathcal{K}(m)}\sum\limits_{j=0}^{N_l-1} s_{k,m}(j)}\right]\right\}. \end{aligned} \tag{69}$$

Then, using the Central Limit Theorem, we can rewrite the effective capacity $EC_m(\theta_m)$ when $N_l \to \infty$ as in the following equation [41]:

$$\begin{aligned} EC_m(\theta_m) &= \frac{\mathbb{E}\left[S_m(N_l)\right]}{2\widehat{n}} + \frac{1}{2\widehat{n}} \\ &\times\sqrt{\left(\mathbb{E}\left[S_m(N_l)\right]\right)^2 - \frac{2\widehat{n}\left(-\log\left(\epsilon_{\text{q},m}\right)\right)}{D_{m,\text{th}}}\text{Var}\left[S_m(N_l)\right]} \end{aligned} \tag{70}$$

where

$$\text{Var}\left[S_m(N_l)\right] = \frac{\left(\widehat{n}R_{m,\text{in}}\right)^2}{\text{Var}\left[N_l\right]}. \tag{71}$$

Accordingly, we can derive the variance of N_l when $N_l \to \infty$ as follows:

$$\begin{aligned} \text{Var}[N_l] &= \mathbb{E}\left[N_l^2\right] - \left(\mathbb{E}\left[N_l\right]\right)^2 \\ &\approx \sum_{j=1}^{L} j^2 \text{Pr}\left\{N_l = j\right\} - \left(1 + \sum_{j=1}^{L-1}\epsilon_{m,j}\right)^2 \\ &= 1 + \sum_{j=1}^{L-1}(2j+1)\text{Pr}\left\{\bigcap_{\iota=1}^{j}\left\{\overline{\mathcal{A}}_\iota\right\}\right\} - \left(1 + \sum_{j=1}^{L-1}\epsilon_{m,j}\right)^2 \\ &\le 1 + \sum_{j=1}^{L-1}(2j+1)\text{Pr}\left\{\overline{\mathcal{A}}_j\right\} - \left(1 + \sum_{j=1}^{L-1}\epsilon_{m,j}\right)^2 \\ &\approx 1 + \sum_{j=1}^{L-1}(2j+1)\epsilon_{m,j} - \left(1 + \sum_{j=1}^{L-1}\epsilon_{m,j}\right)^2 \\ &= \sum_{j=1}^{L-1}(2j-1)\epsilon_{m,j} - \left(\sum_{j=1}^{L-1}\epsilon_{m,j}\right)^2. \end{aligned} \tag{72}$$

Using Eqs. (70) and (72), we can obtain the expression for effective capacity $EC_m(\theta_m)$ as given by Eq. (68), which completes the proof of Theorem 3. ■

VI. Performance Evaluations

We use MATLAB-based simulations to validate and evaluate our proposed mmWave user-centric cell-free m-MIMO based schemes in the finite blocklength regime under HARQ-IR protocol. Throughout our simulations, we set the total bits of the required data packet $M_m = 10^8$ bits, the average transmit power $\overline{\mathcal{P}}_m$ can be choose from $[1, 30]$ Watt for each mobile user, the number of APs $K_\text{a} = 1000$, the number of mobile users $K_u \in [10, 300]$, the pilot signal transmit power $\mathcal{P}_\text{p}$ can be choose from $[1, 5]$ Watt for each mobile user, the number of transmit antennas $N_\text{T} \in [100, 800]$, the number of RF chains $L_\text{T} \in [5, 40]$, the number of entries of the over-complete beam steering codebook $M_\text{T} \in [200, 1000]$, and the maximum number of HARQ-IR retransmission rounds $L \in [5, 20]$.

We set the number of transmit antennas $N_\text{T} = 400$, the number of RF chains $L_\text{T} = 10$, and the number of entries of the over-complete beam steering codebook $M_\text{T} = 600$, the maximum number of HARQ-IR retransmission rounds $L = 10$, the blocklength $\widehat{n} = 600$. Compared with the mmWave m-MIMO based cellular schemes, Fig. 4 depicts the average data transmission rate per user with different numbers of mobile users K_u for our proposed mmWave cell-free m-MIMO schemes. We can observe from Fig. 4 that the average data transmission rate per user increases with the number of mobile users for both cellular and cell-free schemes. Fig. 4

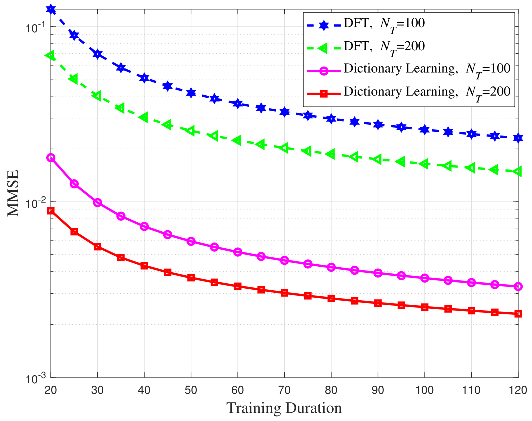


Fig. 6. The MMSE performance vs. beam-training duration for our proposed mmWave user-centric cell-free m-MIMO schemes over 6G mobile wireless networks.

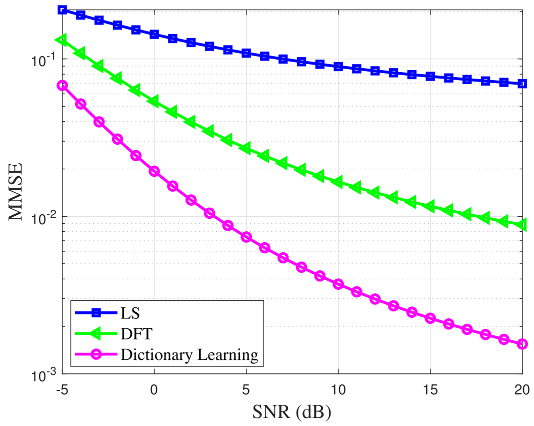


Fig. 7. The MMSE performance vs. SINR (dB) for our proposed mmWave user-centric cell-free m-MIMO schemes over 6G mobile wireless networks.

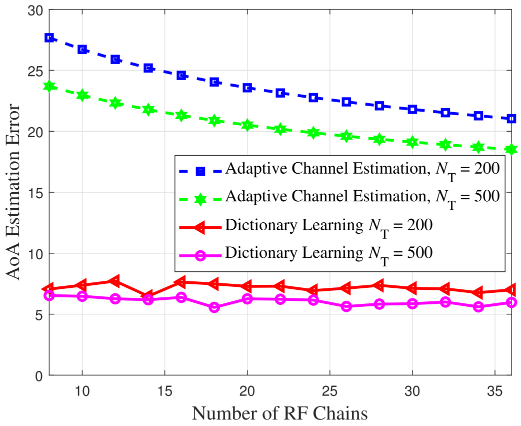


Fig. 8. The AoA estimation error vs. number of RF chains at the AP $L_{\rm T}$ for our proposed mmWave user-centric cell-free m-MIMO schemes over 6G mobile wireless networks.

also shows that with a higher average transmit power at the APs, a better average data transmission rate per user can be achieved. It is shown in Fig. 4 that our proposed mmWave cell-free m-MIMO schemes outperform the traditional mmWave m-MIMO based cellular schemes in terms of the average data transmission rate per user.

Using the same settings as in Fig. 4, Fig. 5 plots the cumulative distribution functions (CDFs) of the downlink data transmission rates per user for our proposed mmWave user-centric cell-free m-MIMO schemes compared with the traditional cell-free m-MIMO schemes. We can observe from Fig. 5 that there is always a crossing point between the CDF curves corresponding to the user-centric (UC) cell-free approach and the traditional cell-free approach. As shown in Fig. 5, the Y-coordinate of the crossing point is far below 0.5 in both $K_{\rm u} = 20$ and $K_{\rm u} = 8$ scenarios. This implies that for the majority of mobile users, our proposed mmWave user-centric cell-free m-MIMO and FBC-HARQ-IR based schemes outperform the traditional mmWave m-MIMO based cellular schemes over 6G mobile wireless networks.

We set the number of mobile users $K_{\rm u} = 20$, the number of RF chains $L_{\rm T} = 10$, and the number of entries of the over-complete beam steering codebook $M_{\rm T} = 600$. Compared with the discrete Fourier transform (DFT) based processing scheme [42], Fig. 6 depicts the MMSE performance with respect to the beam-training duration for our proposed mmWave user-centric cell-free m-MIMO scheme. As shown in Fig. 6, the DFT based processing scheme requires much more training time compared with our proposed dictionary learning based beam-training algorithm for achieving the same MMSE performance. We can observe from Fig. 6 that for our proposed dictionary learning based beam-training algorithm and DFT based processing scheme, we can achieve better MMSE performance with more transmit antennas $N_{\rm T}$ over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks.

Setting the number of mobile users $K_{\rm u} = 20$, the number of transmit antennas $N_{\rm T} = 100$, the number of RF chains $L_{\rm T} = 10$, and the number of entries of the over-complete beam steering codebook $M_{\rm T} = 600$, Fig. 7 plots the MMSE performance with different values of SINR for our proposed mmWave user-centric cell-free m-MIMO schemes in comparison with the Least Square (LS) based channel estimation scheme and DFT based processing scheme. As shown in Fig. 7, our proposed dictionary learning based beam-training algorithm outperforms the LS based channel estimation scheme and DFT based processing scheme in terms of the MMSE performance even in very noisy environment (small SINR environment). Fig. 7 also shows that when the value of SINR is small, i.e., the noise is large, the gaps among different curves are relatively small compared with large SINR environment. This indicates that since the accuracy of channel estimation is limited mostly by noise in low SINR environment, different channel estimation methods only has a small influence on the MMSE performance.

We set the number of mobile users $K_{\rm u} = 20$ and the number of entries of the over-complete beam steering codebook $M_{\rm T} = 600$. Fig. 8 depicts the average AoA estimation error with varying numbers of RF chains at the AP $L_{\rm T}$ for our proposed mmWave user-centric cell-free m-MIMO schemes in comparison with the hybrid design based adaptive channel estimation scheme proposed in [43]. As shown in Fig. 8, the average AoA estimation error for our proposed dictionary

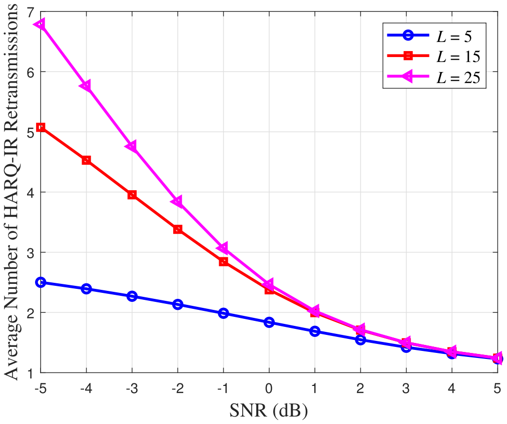


Fig. 9. The LATR $\mathbb{E}[N_l]$ vs. SINR (dB) under HARQ-IR protocol over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks in the finite blocklength regime.

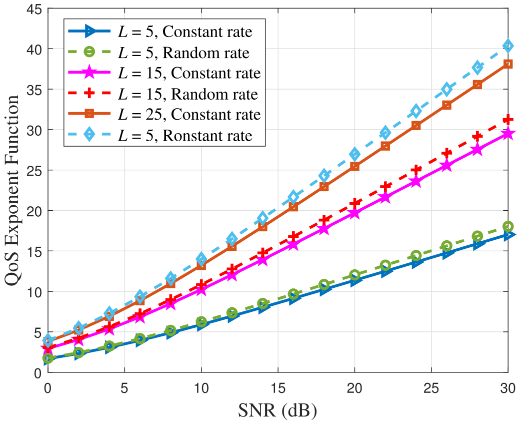


Fig. 11. The QoS exponent function vs. SINR (dB) under HARQ-IR protocol over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks in the finite blocklength regime.

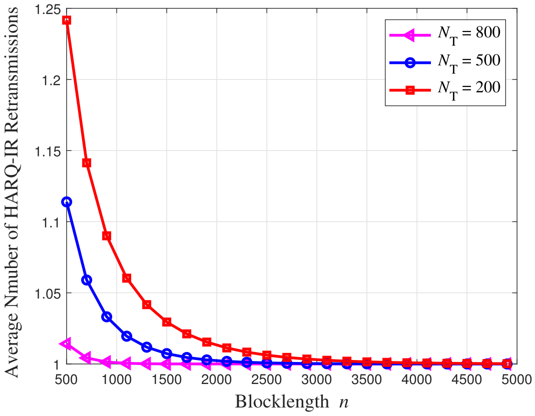


Fig. 10. The LATR $\mathbb{E}[N_l]$ vs. blocklength $\widehat{n}$ under HARQ-IR protocol for mmWave user-centric cell-free m-MIMO schemes.

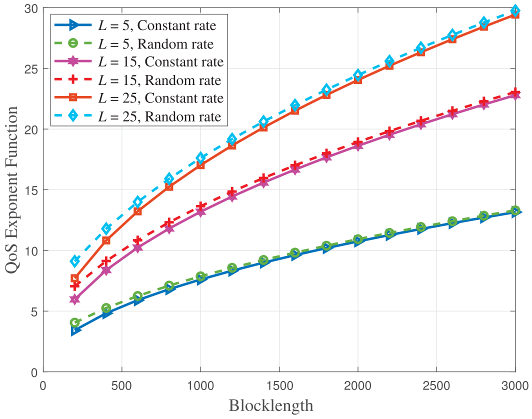


Fig. 12. The QoS exponent function vs. blocklength $\widehat{n}$ under HARQ-IR protocol over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks in the finite blocklength regime.

learning based beam-training algorithm is always less than 10°, which is independent from the number of RF chains at the AP L_{T} and mobile users L_{R}. Fig. 8 also shows that our proposed dictionary learning based beam-training algorithm outperforms the hybrid design based adaptive channel estimation scheme in terms of the average AoA estimation error.

We set the number of mobile users $K_{\mathrm{u}} = 20$, the number of entries of the over-complete beam steering codebook $M_{\mathrm{T}} = 1000$, the blocklength $\widehat{n} = 600$, the number of transmit antennas $N_{\mathrm{T}} = 800$, and the number of RF chains $L_{\mathrm{T}} = 20$. Using the function of LATR $\mathbb{E}[N_l]$ derived in Eq. (54), Fig. 9 plots the LATR $\mathbb{E}[N_l]$ with different values of SINR under HARQ-IR protocol for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime. We can observe from Fig. 9 that for a given maximum number of HARQ-IR retransmission rounds L, the LATR $\mathbb{E}[N_l]$ is a decreasing function of the SINR. This implies that as SINR increases, the average decoding error probability decreases, which results in the decreased value of LATR $\mathbb{E}[N_l]$. Also, as shown in Fig. 9, we can achieve a higher value of LATR $\mathbb{E}[N_l]$ with a larger number of the maximum HARQ-IR retransmission rounds L, which validates the analytical results specified in Eq. (54).

Setting the number of mobile users $K_{\mathrm{u}} = 20$, the number of RF chains $L_{\mathrm{T}} = 20$, the maximum number of HARQ-IR retransmission rounds $L = 25$ and SINR to be 5 dB, Fig. 10 depicts the LATR $\mathbb{E}[N_l]$ with varying values of the blocklengths $\widehat{n}$ under HARQ-IR protocol for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime. We can observe from Fig. 10 that for a given number of transmit antennas N_{T}, the LATR $\mathbb{E}[N_l]$ decreases as the blocklength $\widehat{n}$ increases, and will finally converge to one as $\widehat{n} \to \infty$, which means that as $\widehat{n} \to \infty$, we only need one HARQ-IR retransmission round for a successful decoded message at the receiver. This observation obtained from Fig. 10 implies that as the codeword of length $\widehat{n}$ gets larger, the average decoding error probability decreases, which results in the decreased value of LATR $\mathbb{E}[N_l]$, verifying the analytical results specified in Eqs. (27) and (54).

Then, we set the number of mobile users $K_{\mathrm{u}} = 20$, the blocklength $\widehat{n} = 600$, and the number of transmit antennas $N_{\mathrm{T}} = 500$. Compared with the discrete-time Markov arrival rate model, Fig. 11 plots the QoS exponent function with different values of SINR under HARQ-IR protocol using the average constant arrival rate model and the random arrival rate model in the finite blocklength regime. Fig. 11 shows that the QoS exponent function increases with higher value of the SINR. We can observe from Fig. 11 that for a given maximum number of HARQ-IR retransmission rounds L, we can achieve

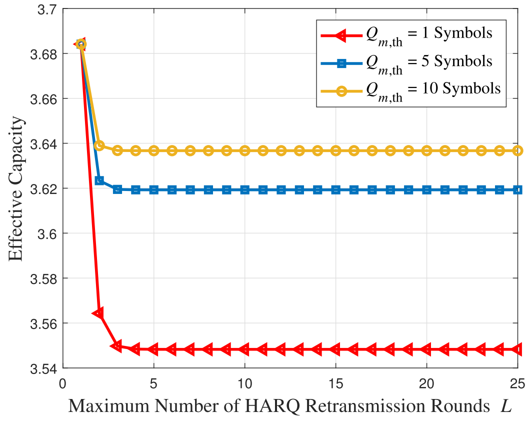


Fig. 13. The effective capacity $EC_m(\theta_m)$ vs. number of the maximum HARQ-IR retransmission rounds L under HARQ-IR protocol for mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime.

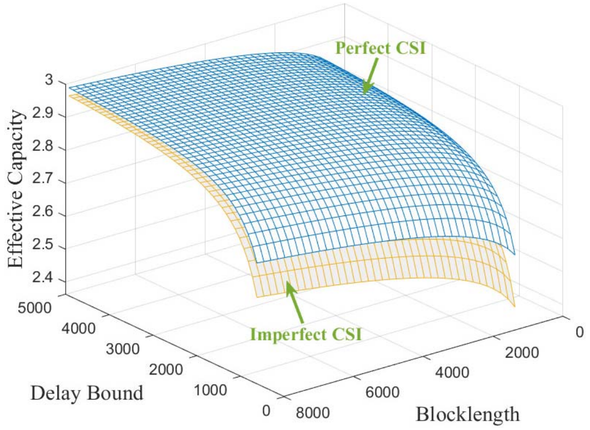


Fig. 14. The effective capacity $EC_m(\theta_m)$ vs. delay bound $D_{m,\text{th}}$ and blocklength $\widehat{n}$ under HARQ-IR protocol for our proposed mmWave user-centric cell-free schemes in the finite blocklength regime.

a higher value of the QoS exponent function by setting a larger number of the maximum HARQ-IR retransmission rounds L, which validates the analytical results specified in Eq. (56).

Now we set the number of mobile users $K_{\text{u}} = 20$, the number of entries of the over-complete beam steering codebook $M_{\text{T}} = 700$, the maximum number of HARQ-IR retransmission rounds $L \in \{5, 15, 20\}$, SINR to be 10 dB, and the number of transmit antennas $N_{\text{T}} = 500$. Using the QoS exponent function derived in Eqs. (56) and (60), Fig. 12 depicts the QoS exponent function with varying values of the blocklengths $\widehat{n}$ under HARQ-IR protocol for our proposed mmWave cell-free m-MIMO schemes in the finite blocklength regime. Fig. 12 shows that for a given value of L, the QoS exponent function is an increasing function of the blocklength $\widehat{n}$, which is consistent with the analytical results specified in Eq. (56).

We set the number of mobile users $K_{\text{u}} = 20$, the number of entries of the over-complete beam steering codebook $M_{\text{T}} = 800$, the blocklength $\widehat{n} = 600$, the number of transmit antennas $N_{\text{T}} = 600$, and the number of RF chains $L_{\text{T}} = 20$. Fig. 13 plots the effective capacity $EC_m(\theta_m)$ with different numbers of the maximum HARQ-IR retransmission rounds L for our proposed mmWave user-centric cell-free m-MIMO schemes in the finite blocklength regime. We can observe from Fig. 13 that as the number of the maximum HARQ-IR retransmission rounds L increases, the effective capacity $EC_m(\theta_m)$ decreases and will finally converge to a certain value. In addition, Fig. 13 shows that with a large/loose delay bound $D_{m,\text{th}}$, or equivalently a large/loose buffer-size overflow threshold $Q_{m,\text{th}}$ which is equal to $R_{m,N_l} D_{m,\text{th}}$ due to the ***Remarks*** as described in the paragraph following Eq. (48), we can achieve a larger value of the effective capacity $EC_m(\theta_m)$, which verifies the analytical results specified in Eq. (68) in Theorem 3.

Setting the number of mobile users $K_{\text{u}} = 20$, the number of entries of the over-complete beam steering codebook $M_{\text{T}} = 800$, the number of transmit antennas $N_{\text{T}} = 600$, the number of RF chains $L_{\text{T}} = 20$, the maximum HARQ-IR retransmission rounds $L = 10$, and SINR to be 15 dB, Fig. 14 depicts the effective capacity $EC_m(\theta_m)$ with varying delay bounds $D_{m,\text{th}}$ and blocklengths $\widehat{n}$ under HARQ-IR protocol for our proposed mmWave cell-free m-MIMO schemes in the finite blocklength regime considering both perfect CSI and imperfect CSI scenarios. We can observe from Fig. 14 that the effective capacity $EC_m(\theta_m)$ increases as the blocklength $\widehat{n}$ gets larger and will enventually converge to a certain value, which is consistent with the analytical results specified in Eq. (68) in Theorem 3.

VII. Conclusions

We have proposed system models that efficiently integrate the HARQ-IR protocol with FBC over mmWave user-centric cell-free m-MIMO based 6G mobile wireless networks. In particular, we have established mmWave user-centric cell-free m-MIMO-based system models. Then, we have designed dictionary learning based beam-training algorithm for solving the low-complex beamforming optimization problem. We also have characterized the channel capacity, channel dispersion, and block error probability under HARQ-IR protocol using FBC. Based on the information theoretic results in QoS theory, we have derived QoS metrics in terms of the error probability and corresponding effective capacity function for our proposed FBC-HARQ-IR based mmWave cell-free m-MIMO schemes. We also have conducted a set of simulations to validate and evaluate our proposed mmWave user-centric cell-free m-MIMO schemes by implementing statistical delay and error-rate bounded QoS provisioning in the finite blocklength regime.

References

[1] X. Zhang, J. Tang, H.-H. Chen, S. Ci, and M. Guizani, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Commun. Mag.*, vol. 44, no. 1, pp. 100–106, 2006.

[2] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, 2007.

[3] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 118–129, Jan. 2008.

[4] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, Jun. 2008.

[5] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous statistical QoS provisioning over 5G mobile wireless networks," *IEEE Netw.*, vol. 28, no. 6, pp. 46–53, Nov. 2014.

[6] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[7] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2823–2835, Aug. 2015.

[8] P. Mary, J.-M. Gorce, A. Unsal, and H. V. Poor, "Finite blocklength information theory: What is the practical impact on wireless communications?" in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[9] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[10] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with nonvanishing error probability," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 5–21, Jan. 2014.

[11] S. Lin and P. Yu, "A hybrid ARQ scheme with parity retransmission for error control of satellite channels," *IEEE Trans. Commun.*, vol. COM-30, no. 7, pp. 1701–1719, Jul. 1982.

[12] D. To, H. X. Nguyen, Q.-T. Vien, and L.-K. Huang, "Power allocation for HARQ-IR systems under QoS constraints and limited feedback," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1581–1594, Mar. 2015.

[13] J.-F. Cheng, "Coding performance of hybrid ARQ schemes," *IEEE Trans. Commun.*, vol. 54, no. 6, pp. 1017–1029, Jun. 2006.

[14] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with chase combining and incremental redundancy for HSDPA," in *Proc. IEEE 54th Veh. Technol. Conf. VTC Fall*, vol. 3, Oct. 2001, pp. 1829–1833.

[15] Y. Li, M. C. Gursoy, and S. Velipasalar, "On the throughput of hybrid-ARQ under statistical queuing constraints," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2725–2732, Jun. 2015.

[16] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.

[17] 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.

[18] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[19] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[20] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[21] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Limited feedback hybrid precoding for multiuser millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[22] Y. Ding and B. D. Rao, "Dictionary learning-based sparse channel representation and estimation for FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5437–5451, Aug. 2018.

[23] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[24] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive MIMO with zero-forcing precoding design," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1871–1874, Aug. 2017.

[25] M. Alonzo and S. Buzzi, "Cell-free and user-centric massive MIMO at millimeter wave frequencies," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.

[26] M. Alonzo, S. Buzzi, and A. Zappone, "Energy-efficient downlink power control in mmWave cell-free and user-centric massive MIMO," in *Proc. IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 493–496.

[27] O. E. Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 3724–3729.

[28] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[29] S. L. Fong, V. Y. F. Tan, and J. Yang, "Non-asymptotic achievable rates for energy-harvesting channels using save-and-transmit," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3499–3511, Dec. 2016.

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

[31] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2, 2nd ed. New York, NY, USA: Wiley, 1971.

[32] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.

[33] X. Li, E. Bjornson, S. Zhou, and J. Wang, "Massive MIMO with multi-antenna users: When are additional user antennas beneficial?" in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, May 2016, pp. 1–6.

[34] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[35] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. Berlin, Germany: Springer-Verlag, 2006.

[36] B. L. Mark and G. Ramamurthy, "Real-time estimation and dynamic renegotiation of UPC parameters for arbitrary traffic sources in ATM networks," *IEEE/ACM Trans. Netw.*, vol. 6, no. 6, pp. 811–827, Dec. 1998.

[37] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 630–643, May 2003.

[38] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1375–1395, Mar. 2016.

[39] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[40] W. Cheng, X. Zhang, and H. Zhang, "Heterogeneous statistical QoS provisioning for downlink transmissions over mobile wireless cellular networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4757–4763.

[41] B. Soret, M. Aguayo-torres, and J. Entrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated Rayleigh channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 1901–1911, Jun. 2010.

[42] W. Tan, M. Matthaiou, S. Jin, and X. Li, "Spectral efficiency of DFT-based processing hybrid architectures in massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 6, no. 5, pp. 586–589, Oct. 2017.

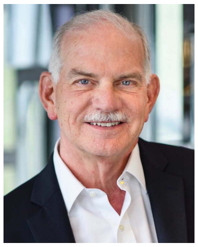
[43] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

Xi Zhang (Fellow, IEEE) received the B.S. degree and M.S. degree from Xidian University, Xi'an, China, the M.S. degree from Lehigh University, Bethlehem, PA, USA, all in electrical engineering and computer science, and the Ph.D. degree in electrical engineering and computer science (Electrical Engineering – Systems) from The University of Michigan, Ann Arbor, MI, USA.

He is currently a Full Professor and the Founding Director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is a Fellow of the IEEE for contributions to quality-of-services (QoS) theory in mobile wireless networks. He was with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hill, NJ, USA, and AT&T Laboratories Research, Florham Park, NJ, USA, in 1997. He was a Research Fellow with the School of Electrical Engineering, University of Technology, Sydney, Australia, and the Department of Electrical and Computer Engineering, James Cook University, Australia. He has published more than 370 research articles on wireless networks and communications systems, network protocol design and modeling, statistical communications, random signal processing, information theory, and control theory and systems. He received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He received five Best Paper Awards at IEEE ICC 2018, IEEE GLOBECOM 2014, IEEE GLOBECOM 2009, IEEE GLOBECOM 2007, and IEEE WCNC 2010, respectively. One of his IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS articles has been listed as the IEEE Best Readings Paper (receiving the highest citation rate among all IEEE Transactions/Journal papers in the area) on "Wireless Cognitive Radio Networks" and "Statistical QoS Provisioning over Mobile Wireless Networking". He is an IEEE Distinguished Lecturer of both IEEE Communications Society and IEEE Vehicular Technology Society. He also received the TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering, Texas A&M University, in 2006.

Professor Xi Zhang is serving or has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, twice as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for two Special Issues on "Broadband Wireless Communications for High Speed Vehicles" and "Wireless Video Transmissions", an Associate Editor for IEEE COMMUNICATIONS LETTERS, twice as a Lead Guest Editor for *IEEE Communications Magazine* for two Special Issues on "Advances in Cooperative Wireless Networking" and "Underwater Wireless Communications and Networks: Theory and Applications", a Guest Editor for *IEEE Wireless Communications Magazine* for Special Issue on "Next Generation CDMA Versus OFDMA for 4G Wireless Applications", an Editor for *Journal on Wireless Communications and Mobile Computing* (Wiley), the *Journal of Computer Systems, Networking, and Communications*, and *Journal on Security and Communications Networks* (Wiley), and an Area Editor for *Journal on Computer Communications* (Elsevier), among many others. He is serving or has served as the TPC Chair for IEEE GLOBECOM 2011, TPC Vice-Chair for IEEE INFOCOM 2010, TPC Area Chair for IEEE INFOCOM 2012, Panel/Demo/Poster Chair for ACM MobiCom 2011, General Chair for IEEE WCNC 2013, and TPC Chair for IEEE INFOCOM 2017–2019 Workshops on "Integrating Edge Computing, Caching, and Offloading in Next Generation Networks", and so on.

Jingqing Wang received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China. She is currently pursuing the Ph.D. degree with Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, under the supervision of Professor Xi Zhang. Her research interests include big data-based 5G wireless networks technologies, statistical QoS provisioning, and cognitive radio networks. She won the Best Paper Award at the IEEE GLOBECOM 2014 and also received the Hagler Institute for Advanced Study Heep Graduate Fellowship Award from Texas A&M University in 2018.



H. Vincent Poor (Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign.

Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests include information theory, signal processing and machine learning, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Multiple Access Techniques for 5G Wireless Networks and Beyond* (Springer, 2019).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the Marconi and Armstrong Awards of the IEEE Communications Society, in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, a D.Sc. *honoris causa* from Syracuse University awarded in 2017, and a D.Eng. *honoris causa* from the University of Waterloo awarded in 2019.