Statistical Delay and Error-Rate Bounded QoS Provisioning for mURLLC Over 6G CF M-MIMO Mobile Networks in the Finite Blocklength Regime

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

***Abstract*— In supporting the new 6G standard traffic services–massive ultra-reliable low-latency communications (mURLLC), several advanced techniques, including statistical delay-bounded quality-of-service (QoS) provisioning theory and finite blocklength coding (FBC), have been developed to upper-bound both delay and error-rate for time-sensitive multimedia applications. On the other hand, cell-free (CF) massive multi-input multi-output (m-MIMO), where a large number of distributed access points (APs) jointly serve a massive number of mobile devices using the same time-frequency resources, has emerged as one of the 6G key promising techniques to significantly improve various QoS performances for supporting mURLLC. However, it is challenging to statistically guarantee stringent mURLLC QoS-requirements for transmitting multimedia traffics over CF m-MIMO and FBC based 6G wireless networks. To overcome these problems, we develop analytical models to precisely characterize the delay and error-rate bounded QoS performances while considering non-vanishing decode-error probability for CF m-MIMO based schemes. In particular, we develop FBC based system models and apply the Mellin transform to characterize arrival/service processes for our proposed CF m-MIMO modeling schemes. Then, we formulate and solve the delay violation probability minimization problem and obtain the closed-form solution of the optimal rate adaptation policy for each mobile user over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime. Our simulation results validate and evaluate our proposed schemes for statistical delay and error-rate bounded QoS provisioning.**

***Index Terms*— Statistical delay and error-rate bounded QoS, 6G, mURLLC, CF m-MIMO, FBC, Mellin transform, SNC.**

I. Introduction

DUE TO the stochastic nature of wireless fading channels, it is challenging to guarantee both reliability and low-latency requirements for delay-sensitive wireless multimedia services over 6G mobile wireless networks. Traditionally, researchers have developed a deterministic network calculus to derive explicit guarantees on the maximum delay for wireless data transmissions. However, deterministic network calculus is not sufficient for characterizing the wireless traffic due to time-varying and stochastic natures of wireless fading channels. Towards this end, the delay-bounded QoS theory [1] [2] and stochastic network calculus (SNC) [3] have been developed to analyze queueing behaviors based on the theories of large deviations and effective capacity. Particularly, the concept of *statistical QoS provisioning* [1] [4], in terms of effective capacity and delay-bounded violation probabilities, has been proposed to support time-sensitive wireless communications over 6G mobile wireless networks.

In addition, as a new and dominating 6G mobile-networks' service class for time-sensitive traffic, massive ultra-reliable and low latency communications (mURLLC), which integrates URLLC with massive access, also known as massive machine type communications (mMTC), requires massive short-packet data communications to support time-sensitive 6G wireless multimedia services with high resource efficiency and low access latency [5]. This implies that the traditional Shannon's theorem with infinite blocklength is no longer applicable in this regime. Towards this end, finite blocklength coding (FBC) [6] has been proposed to guarantee both latency and reliability requirements using short-packet data communications for supporting 6G wireless multimedia services. The maximum achievable coding rate using FBC over additive white Gaussian noise (AWGN) channels has been derived in [7]. The authors of [8] have derived the goodput over AWGN channels and the energy-efficiency spectral-efficiency tradeoff by using recent results on the non-asymptotic coding rate. The maximum achievable data rates using FBC over quasi-static MIMO based wireless fading channels with and without the knowledge of channel state information (CSI) have been derived in [9]. The authors of [10] have investigated different properties of channel codes for a given memoryless wireless channel with a non-vanishing error probability.

On the other hand, various advanced promising 6G techniques, such as the cell-free (CF) massive multi-input multi-output (m-MIMO) [11], have been designed to play a critically important role in supporting mURLLC as well as mMTC in terms of connecting massive numbers of mobile devices without imposing congestion. In particular, CF m-MIMO [12], where the geographically distributed APs jointly

Manuscript received January 27, 2020; revised June 10, 2020; accepted July 17, 2020. Date of publication September 3, 2020; date of current version February 17, 2021. 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 also supported in part by the U.S. National Science Foundation under Grant CCF-0939370 and Grant CCF-1908308. *(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 figures in this article are available at https://doi.org/10.1109/JSAC.2020.3020109.

Digital Object Identifier 10.1109/JSAC.2020.3020109

serve a massive number of mobile devices using the same time-frequency resources, has been developed as a promising 6G network architecture for improving the reliability of massive access while reducing the co-channel interference caused by standard m-MIMO systems. Traditionally, optimal power control is performed at the central processing unit (CPU). However, centralized power-control strategies may jeopardize the system scalability and violate mURLLC requirements as the numbers of APs and mobile users grow significantly. The authors of [13] have proposed scalable and distributed power control policies for CF m-MIMO systems to achieve system scalability and mURLLC as the number of mobile users goes to infinity. The authors of [14] have developed a new framework for scalable CF m-MIMO systems, where the complexity and signalling at each AP is finite when connecting a massive number of mobile devices. The system scalability aspects of CF m-MIMO systems are analyzed in [15] and a solution is proposed for data processing, network topology, and power control. However, it is challenging to characterize the stochastic networking/queueing behaviors when being integrated with CF m-MIMO schemes while guaranteeing mURLLC under statistical delay and error-rate bounded QoS constraints in the finite blocklength regime. As a result, how to accurately upper-bound the delay violation probability while guaranteeing statistical delay and error-rate bounded QoS provisionings for supporting mURLLC still remains as a challenging and open problem for CF m-MIMO mobile wireless networks in the finite blocklength regime.

To effectively overcome the above-mentioned challenges, in this paper we propose to apply the Mellin transform to analytically model and characterize stochastic QoS performances in terms of both delay and error-rate for CF m-MIMO modeling schemes in the finite blocklength regime. In particular, we develop CF m-MIMO based system models across Rician wireless fading channels in the finite blocklength regime. Furthermore, we propose and analyze the delay-violation probability function by applying the Mellin transform over both arrival and service processes, while taking into account the statistical delay and error-rate bounded QoS constraints. We also formulate and solve the delay-violation probability minimization problem for our proposed CF m-MIMO modeling schemes in the finite blocklength regime. Also conducted is a set of simulations to validate and evaluate our proposed schemes for statistical delay and error-rate bounded QoS provisioning over 6G CF m-MIMO mobile wireless networks.

The rest of this paper is organized as follows: Section II establishes FBC based CF m-MIMO based system models across Rician wireless channels. Section III derives and analyzes the Mellin transform over arrival/service processes as well as an upper bound on the delay violation probability in the finite blocklength regime. Section IV analyzes the delay performance and formulates and solves the delay violation probability minimization problem for statistical delay and error-rate bounded QoS provisioning in the finite blocklength regime. Section V evaluates and analyzes the system performance for our proposed FBC based CF m-MIMO schemes. The paper concludes with Section VI.

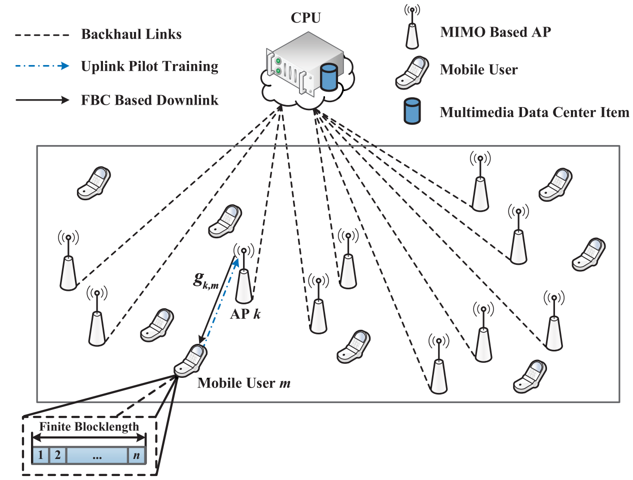


Fig. 1. The system architecture model for 6G CF m-MIMO mobile wireless networks in the finite blocklength regime.

II. The System Models

Consider a CF m-MIMO network model, where each mobile user is served by coherent joint transmissions from all APs, as shown in Fig. 1. Assume that there are $K_{\rm a}$ randomly located APs over a large area and $K_{\rm u}$ mobile users. Assume that each AP is equipped with $N_{\rm T}$ antennas, while each mobile user is equipped with a single antenna. In addition, time division duplexing (TDD) mode is assumed to be operated over our proposed 6G CF m-MIMO mobile wireless networks. All APs are connected to a CPU through backhaul links. Define $n_{\rm p}$ as the number of channel uses for uplink pilot training symbols and $n_{\rm d}$ as the number of channel uses reserved for downlink data transmissions. Defining n as the total number of channel uses for both uplink pilot training and downlink data transmission phases, we have $n = n_{\rm p} + n_{\rm d}$.

A. Massive MIMO Based Rician Wireless Fading Channel Model

Considering the propagation effects, the channel's impulse response vector, denoted by $\boldsymbol{g}_{k,m} \in \mathbb{C}^{N_{\rm T}\times 1}$, between mobile user m and AP k over Rician massive MIMO based wireless fading channel model can be characterized as follows:

$$\boldsymbol{g}_{k,m} = \boldsymbol{h}_{k,m}\sqrt{\beta_{k,m}} \tag{1}$$

where $\beta_{k,m}$ represents the large-scale propagation that includes pathloss and shadowing effects and $\boldsymbol{h}_{k,m}$ represents the small-scale multipath fading effect which can be modeled using Rician distribution [16] as follows:

$$\boldsymbol{h}_{k,m} = \sqrt{\frac{\kappa}{\kappa+1}}\overline{\boldsymbol{h}}_{k,m} + \sqrt{\frac{1}{\kappa+1}}\widetilde{\boldsymbol{h}}_{k,m} \tag{2}$$

where $\boldsymbol{h}_{k,m}$ consists of the component $\overline{\boldsymbol{h}}_{k,m}$ representing the line of sight (LOS) signals and a Rayleigh distributed random component $\widetilde{\boldsymbol{h}}_{k,m}$ representing the non-line-of-sight (NLOS) signals, and $\kappa > 0$ is the Rician factor. Note that $\kappa = 0$ corresponds to a Rayleigh fading channel, while $\kappa \to \infty$ corresponds to non-fading channel. We can rewrite the

Rician massive MIMO based wireless fading channel model as follows:

$$\boldsymbol{g}_{k,m} = \sqrt{\frac{\kappa}{\kappa+1}}\overline{\boldsymbol{g}}_{k,m} + \sqrt{\frac{1}{\kappa+1}}\widetilde{\boldsymbol{g}}_{k,m} \tag{3}$$

where $\overline{\boldsymbol{g}}_{k,m} \triangleq \overline{\boldsymbol{h}}_{k,m}\beta_{k,m}$ and $\widetilde{\boldsymbol{g}}_{k,m} \triangleq \widetilde{\boldsymbol{h}}_{k,m}\beta_{k,m}$.

B. Uplink Pilot Training and Channel Estimation

Define the pilot training sequence for mobile user m as $\boldsymbol{\phi}_m^{n_\text{p}} = \left[\phi_m^{(1)}, \ldots, \phi_m^{(n_\text{p})}\right] \in \mathbb{C}^{1\times n_\text{p}}$ and $\|\boldsymbol{\phi}_m^{n_\text{p}}\|^2 = 1$, where $\|\cdot\|$ denotes the Euclidean norm. During the uplink pilot training phase, we derive the received signal, denoted by $\mathbf{Y}_k^{n_\text{p}} \in \mathbb{C}^{N_\text{T}\times n_\text{p}}$, for transmitting n_p pilot data blocks from mobile user m to AP k as in the following equation:

$$\mathbf{Y}_k^{n_\text{p}} = \sum_{m=1}^{K_\text{u}} \sqrt{n_\text{p}\mathcal{P}_\text{p}}\boldsymbol{g}_{k,m}\boldsymbol{\phi}_m^{n_\text{p}} + \mathbf{N}_k \tag{4}$$

where $\mathcal{P}_\text{p}$ is the uplink pilot transmit power at the mobile users and $\mathbf{N}_k \in \mathbb{C}^{N_\text{T}\times n_\text{p}}$ is the AWGN matrix with zero mean and covariance $\mathbf{I}_{N_\text{T}}$ where $\mathbf{I}_{N_\text{T}}$ is the identity matrix of size N_T. We assume that the LOS component $\overline{\boldsymbol{g}}_{k,m}$ given in Eq. (3) is perfectly known at both the APs and mobile users. Accordingly, we only need to estimate the NLOS Rayleigh-distributed random component. As a result, the received matrix, denoted by $\widetilde{\mathbf{Y}}_k^{n_\text{p}}$, for the NLOS channel estimation can be derived as follows:

$$\widetilde{\mathbf{Y}}_k^{n_\text{p}} = \sum_{m=1}^{K_\text{u}} \sqrt{\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}}\widetilde{\boldsymbol{g}}_{k,m}\boldsymbol{\phi}_m^{n_\text{p}} + \mathbf{N}_k. \tag{5}$$

Then, by projecting the received signal $\widetilde{\mathbf{Y}}_k^{n_\text{p}}$ onto $\boldsymbol{\phi}_m^{n_\text{p}}$, we obtain:

$$\widetilde{\boldsymbol{y}}_k^{n_\text{p}} = \widetilde{\mathbf{Y}}_k^{n_\text{p}}\left(\boldsymbol{\phi}_m^{n_\text{p}}\right)^H = \sqrt{\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}}\widetilde{\boldsymbol{g}}_{k,m} + \sum_{\substack{m'=1\\ m'\neq m}}^{K_\text{u}} \sqrt{\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}}\widetilde{\boldsymbol{g}}_{k,m'} + \widetilde{\boldsymbol{n}}_k \tag{6}$$

where $(\cdot)^H$ represents the conjugate transpose of a vector and $\widetilde{\boldsymbol{n}}_k \triangleq \mathbf{N}_k\left(\boldsymbol{\phi}_m^{n_\text{p}}\right)^H$ is an independent and identically distributed (i.i.d.) Gaussian vector with zero mean and covariance $\mathbf{I}_{N_\text{T}}$. Denote by $\widetilde{\mathbf{G}}_k \triangleq \left[\widetilde{\boldsymbol{g}}_{k,1}, \ldots, \widetilde{\boldsymbol{g}}_{k,K_\text{u}}\right]$ the NLOS component of the channel's impulse response matrix from all AP k to all mobile users. Define $\mathbf{R}_{\widetilde{\mathbf{G}}_k} \triangleq \mathbb{E}\left[\widetilde{\mathbf{G}}_k\left(\widetilde{\mathbf{G}}_k\right)^H\right] = \text{diag}\left(\beta_{k,1}, \ldots, \beta_{k,K_\text{u}}\right)$ as the covariance matrix of $\widetilde{\mathbf{G}}_k$, where $\mathbb{E}[\cdot]$ is the expectation operation and diag$(\cdot)$ represents the diagonal matrix. Then, considering the Rician wireless fading channels, we can derive the channel estimation for our proposed CF m-MIMO schemes as detailed in the following lemma.

Lemma 1: The minimum mean-squared error (MMSE) estimator, denoted by $\widehat{\mathbf{G}}_k$, for the NLOS component of the Rician massive MIMO based wireless fading channel $\widetilde{\mathbf{G}}_k$ from AP k to all mobile users is derived as follows:

$$\widehat{\mathbf{G}}_k = \frac{\sqrt{n_\text{p}\mathcal{P}_\text{p}}}{\kappa+1}\mathbf{R}_{\widetilde{\mathbf{G}}_k}\left(\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}\mathbf{R}_{\widetilde{\mathbf{G}}_k} + \mathbf{I}_{N_\text{T}}\right)^{-1}\widetilde{\boldsymbol{y}}_k^{n_\text{p}}. \tag{7}$$

Proof: Applying the MMSE estimator of $\widetilde{\mathbf{G}}_k$ based on the observation of $\widetilde{\boldsymbol{y}}_k^{n_\text{p}}$, we can obtain the following equation [17]:

$$\begin{aligned}\widehat{\mathbf{G}}_k &= \mathbb{E}\left[\widetilde{\mathbf{G}}_k | \widetilde{\boldsymbol{y}}_k^{n_\text{p}}\right]\\ &= \mathbf{R}_{\widetilde{\mathbf{G}}_k,\widetilde{\boldsymbol{y}}_k^{n_\text{p}}}\left(\mathbf{R}_{\widetilde{\boldsymbol{y}}_k^{n_\text{p}}}\right)^{-1}\left(\widetilde{\boldsymbol{y}}_k^{n_\text{p}} - \mathbb{E}\left[\widetilde{\boldsymbol{y}}_k^{n_\text{p}}\right]\right) + \mathbb{E}\left[\widetilde{\mathbf{G}}_k\right]\end{aligned} \tag{8}$$

where $\mathbf{R}_{\widetilde{\mathbf{G}}_k,\widetilde{\boldsymbol{y}}_k^{n_\text{p}}}$ and $\mathbf{R}_{\widetilde{\boldsymbol{y}}_k^{n_\text{p}}}$ represent the covariance matrices, respectively, as follows:

$$\begin{cases}\mathbf{R}_{\widetilde{\mathbf{G}}_k,\widetilde{\boldsymbol{y}}_k^{n_\text{p}}} = \mathbb{E}\left[\widetilde{\mathbf{G}}_k\left(\widetilde{\boldsymbol{y}}_k^{n_\text{p}}\right)^H\right] = \sqrt{\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}}\mathbf{R}_{\widetilde{\mathbf{G}}_k};\\ \mathbf{R}_{\widetilde{\boldsymbol{y}}_k^{n_\text{p}}} = \mathbb{E}\left[\widetilde{\boldsymbol{y}}_k^{n_\text{p}}\left(\widetilde{\boldsymbol{y}}_k^{n_\text{p}}\right)^H\right] = \frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}\mathbf{R}_{\widetilde{\mathbf{G}}_k} + \mathbf{I}_{N_\text{T}}.\end{cases} \tag{9}$$

Since $\mathbb{E}\left[\widetilde{\boldsymbol{y}}_k^{n_\text{p}}\right]$ and $\mathbb{E}\left[\widetilde{\mathbf{G}}_k\right]$ are equal to zero, we have

$$\widehat{\mathbf{G}}_k = \sqrt{\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}}\mathbf{R}_{\widetilde{\mathbf{G}}_k}\left(\frac{n_\text{p}\mathcal{P}_\text{p}}{\kappa+1}\mathbf{R}_{\widetilde{\mathbf{G}}_k} + \mathbf{I}_{N_\text{T}}\right)^{-1}\widetilde{\boldsymbol{y}}_k^{n_\text{p}} \tag{10}$$

which completes the proof of Lemma 1. ■

As a result, the channel estimation of Rician m-MIMO based wireless fading channel model, denoted by $\widehat{\widehat{\mathbf{G}}}_k$, can be derived as follows:

$$\widehat{\widehat{\mathbf{G}}}_k = \sqrt{\frac{\kappa}{\kappa+1}}\overline{\mathbf{G}}_k + \sqrt{\frac{1}{\kappa+1}}\widehat{\mathbf{G}}_k. \tag{11}$$

C. Downlink Finite-Blocklength Data Transmission

Define P_m as the multiplexing order for mobile user m, i.e., the number of information symbols simultaneous transmitted at the transmitter. As a result, we define a P_m-dimensional beamformer $\mathbf{L}_m \triangleq \mathbf{I}_{P_m} \otimes \mathbf{1}_{N_\text{R}/P_m}$ for mobile user m, where $\otimes$ is the Kronecker product, $\mathbf{I}_{P_m}$ is the identity matrix of size P_m, and $\mathbf{1}_{N_\text{R}/P_m}$ is the all one vector with length N_R/P_m. We define the transmit signal matrix as $\mathbf{X}_k^{n_\text{d}} \triangleq \left[\boldsymbol{x}_k^{(1)}, \ldots, \boldsymbol{x}_k^{(n_\text{d})}\right]$ and receive signal vector $\boldsymbol{y}_m^{n_\text{d}} \triangleq \left[y_m^{(1)}, \ldots, y_m^{(n_\text{d})}\right]$. Based on the MMSE estimator matrix, denoted by $\widehat{\widehat{\mathbf{G}}}_k = \left[\widehat{\widehat{\boldsymbol{g}}}_{k,1}, \ldots, \widehat{\widehat{\boldsymbol{g}}}_{k,K_\text{u}}\right]$, obtained during the uplink pilot training phase, we can derive the transmitted signal, denoted by $\boldsymbol{x}_k^{(l)}$, for transmitting data block l from AP k to mobile user m by employing conjugate beamforming [11] as follows:

$$\boldsymbol{x}_k^{(l)} = \mathbf{W}_k\left(\boldsymbol{\Sigma}_k\right)^{\frac{1}{2}} s_m^{(l)}, \quad l = 1, \ldots, n_\text{d} \tag{12}$$

where $s_m^{(l)}$ represents the lth transmitted data block to mobile user m, $\boldsymbol{\Sigma}_k \triangleq \text{diag}\left(\eta_{k,1}, \ldots, \eta_{k,K_\text{u}}\right)$ is the power allocation coefficient matrix, where $\eta_{k,m}$ $(m = 1, \ldots, K_\text{u})$ is the power allocation coefficient for transmitting finite-blocklength data block l from AP k to mobile user m, and $\mathbf{W}_k$ is the downlink precoder from AP k to all K_u mobile users, which is given by

$$\mathbf{W}_k \triangleq \widehat{\widehat{\mathbf{G}}}_k\left[\left(\widehat{\widehat{\mathbf{G}}}_k\right)^H\widehat{\widehat{\mathbf{G}}}_k\right]^{-1}\mathbf{L}_m\left(\boldsymbol{\Xi}_k\right)^{\frac{1}{2}} \tag{13}$$

where $\boldsymbol{\Xi}_k = \text{diag}\left(\chi_1, \ldots, \chi_{K_\text{u}}\right)$ is the normalization matrix such that the columns of $\mathbf{W}_k$ have unit norm and the normalization variable χ_k with $k \in \{1, \ldots, K_\text{u}\}$ follows the central chi-square distribution with (2ℓ) degrees of freedom, where $\ell = K_{\mathrm{a}} - K_{\mathrm{u}} + 1$. The probability density function (PDF) of χ_k is given as follows [18]:

$$f_\ell(\chi_k) = \frac{1}{\Gamma(\ell)} \chi_k^{\ell-1} e^{-\chi_k} \tag{14}$$

where $\Gamma(\cdot)$ denotes the Gamma function. We have $\mathbf{W}_k = [\boldsymbol{w}_{k,1}, \ldots, \boldsymbol{w}_{k,K_{\mathrm{u}}}]$, where $\boldsymbol{w}_{k,m}$ is the downlink precoder from AP k to mobile user m $(m \in \{1, \ldots, K_{\mathrm{u}}\})$. In addition, the power control coefficients need to satisfy the following power constraint at each AP:

$$\frac{1}{n_{\mathrm{d}}} \sum_{l=1}^{n_{\mathrm{d}}} \mathbb{E}\left[\left\|\boldsymbol{x}_k^{(l)}\right\|^2\right] \leq \overline{\mathcal{P}}_{\mathrm{d}} \tag{15}$$

where $\overline{\mathcal{P}}_{\mathrm{d}}$ represents the average transmit power at each AP and $\boldsymbol{x}_k^{(l)}$ is given by Eq. (12). Furthermore, as the number of APs K_{a} grows sufficiently large, the system will experience only small variations (relative to the average) in the achievable data transmission rate, which is known as the channel hardening [19]. As a result, although the instantaneous CSI is not available at the mobile users, $\mathbb{E}\left[\left(\boldsymbol{g}_{k,m}\right)^T \boldsymbol{w}_{k,m}\right]$ can be used to calculate the channel gain, where $(\cdot)^T$ represents the transpose of a vector. Considering Rician wireless fading channels, we derive the received signal, denoted by $y_m^{(l)}$, from the kth AP to the mth mobile user for transmitting the lth finite-blocklength data block as follows [11]:

$$\begin{aligned} y_m^{(l)} = & \underbrace{\sum_{k=1}^{K_{\mathrm{a}}} \sqrt{\overline{\mathcal{P}}_{\mathrm{d}} \eta_{k,m}}\, \mathbb{E}\left[\left(\boldsymbol{g}_{k,m}\right)^T \boldsymbol{w}_{k,m}\right] s_m^{(l)}}_{\mathrm{DS}_m} \\ & + \underbrace{\sqrt{\overline{\mathcal{P}}_{\mathrm{d}}} \Bigg\{ \sum_{k=1}^{K_{\mathrm{a}}} \sqrt{\eta_{k,m}} \left(\boldsymbol{g}_{k,m}\right)^T \boldsymbol{w}_{k,m} - \mathbb{E}\Bigg[\sum_{k=1}^{K_{\mathrm{a}}} \sqrt{\eta_{k,m}}}_{\mathrm{BU}_m} \\ & \times \underbrace{\left(\boldsymbol{g}_{k,m}\right)^T \boldsymbol{w}_{k,m}\Bigg]\Bigg\} s_m^{(l)}}_{\mathrm{BU}_m} \\ & + \underbrace{\sum_{\substack{m'=1 \\ m' \neq m}}^{K_{\mathrm{u}}} \sum_{k=1}^{K_{\mathrm{a}}} \sqrt{\overline{\mathcal{P}}_{\mathrm{d}} \eta_{k,m'}} \left(\boldsymbol{g}_{k,m}\right)^T \boldsymbol{w}_{k,m'} s_{m'}^{(l)}}_{\mathrm{UI}_{m'}} + n_m^{(l)} \end{aligned} \tag{16}$$

where $s_m^{(l)}$ and $s_{m'}^{(l)}$ are the signals sent to mobile user m and mobile user m', respectively; $\eta_{k,m}$ and $\eta_{k,m'}$ are the power allocation parameters for transmitting from AP k to mobile user m and mobile user m', respectively; $\boldsymbol{g}_{k,m} \in \mathbb{C}^{1 \times N_{\mathrm{T}}}$ represents the channel's impulse response vector from the kth AP to mobile user m; $n_m^{(l)}$ is the AWGN with zero mean and unit variance; and DS_m, BU_m, and $\mathrm{UI}_{m'}$ represent the strength of the desired signal (DS), the beamforming gain uncertainty (BU), and the interference caused by the m'th mobile user (UI), respectively. Correspondingly, we can derive the signal-to-noise-plus-interference ratio (SINR), denoted by γ_m, from the APs to mobile user m as follows:

$$\gamma_m = \frac{\left\|\mathrm{DS}_m\right\|^2}{\mathbb{E}\left[\left\|\mathrm{BU}_m\right\|^2\right] + \sum\limits_{\substack{m'=1 \\ m' \neq m}}^{K_{\mathrm{u}}} \mathbb{E}\left[\left\|\mathrm{UI}_{m'}\right\|^2\right] + 1}. \tag{17}$$

III. Statistical Delay and Error-Rate Bounded QoS Provisioning in the Finite Blocklength Regime

In this section, we derive the Mellin transform over arrival and service processes by using SNC for our proposed CF m-MIMO schemes given the non-vanishing error probability.

A. (n_d, M_m, ϵ_m)-Code

Definition 1 ($(n_d, M_m, \epsilon_m)-Code$): We define a codebook consisting of M_m codewords, denoted by $(c_1, \ldots, c_{M_m})$, with length n_{d}. We define a message set $\mathcal{M}_m = \{1, \ldots, M_m\}$ and a message $W_m \in \mathcal{M}_m$ which is uniformly distributed on $\mathcal{M}_m$. We define an $(n_{\mathrm{d}}, M_m, \epsilon_m)$-code $(\epsilon_m \in [0, 1))$ as follows:

- An encoder Υ: $\{1, \ldots, M_m\} \mapsto \mathbb{C}^{N_{\mathrm{T}} \times n_{\mathrm{d}}}$ that maps the message $W_m \in \mathcal{M}_m$ to a codeword $\mathbf{X}_m^{n_{\mathrm{d}}}$ with length n_{d}.
- A decoder $\mathcal{D}$: A decoder $\left\{\mathcal{D}_{\boldsymbol{g}_{k,m}}\right\}_{\boldsymbol{g}_{k,m}}$: $\mathbb{C}^{1 \times N_{\mathrm{T}}} \times \mathbb{C}^{N_{\mathrm{T}} \times n_{\mathrm{d}}} \mapsto \{1, \ldots, M_m\} \bigcup \{e\}$, where e represents the error event.

In [6], the accurate approximation of the maximum achievable data rate, denoted by R_m (bits per channel use), with error probability, denoted by ϵ_m with $0 \leq \epsilon_m < 1$, and coding blocklength, denoted by n_{d}, for mobile user m in the finite blocklength regime can be determined as follows:

$$R_m(n_{\mathrm{d}}, \epsilon_m) \approx C(\gamma_m) - \sqrt{\frac{V(\gamma_m)}{n_{\mathrm{d}}}} Q^{-1}(\epsilon_m) \tag{18}$$

where $Q^{-1}(\cdot)$ is the inverse of Q-function and $C(\gamma_m)$ and $V(\gamma_m)$ are the channel capacity and channel dispersion, respectively, which are given as follows [6]:

$$\begin{cases} C(\gamma_m) = \log_2\left(1 + \gamma_m\right); \\ V(\gamma_m) = 1 - \dfrac{1}{\left(1 + \gamma_m\right)^2}. \end{cases} \tag{19}$$

B. Stochastic Network Calculus

Consider that each AP is equipped with a QoS-driven first-in-first-out (FIFO) buffer. We define $a_m(l)$ as the source rate for transmitting the lth data block to mobile user m and $s_m(l)$ as the instantaneous service rate over wireless channels for transmitting the lth data block to mobile user m. Define $A_m(l) \triangleq \sum_{j=0}^{l-1} a_m(j)$ as the accumulated source rate for transmitting l data blocks to mobile user m and $S_m(l) \triangleq \sum_{j=0}^{l-1} s_m(j)$ as the accumulated service rate over wireless channels for transmitting l data blocks to mobile user m. Define $Q_m(l)$ as the dynamics of queuing process for transmitting l data blocks to mobile user m, which is given as in the following equation:

$$Q_m(l) = \max\left\{A_m(l) - S_m(l), 0\right\}. \tag{20}$$

However, in practical scenarios, it is difficult to obtain the statistical characteristics of random arrival and service processes. As a result, by taking the exponential of arrival and service processes, we can transform the arrival and service processes, denoted by $A_m(l)$ and $S_m(l)$, respectively, in the bit domain into the exponential domain, i.e., *signal-to-noise ratio (SNR)-domain* [20] by using the exponential function given as follows:

$$\begin{cases} \mathcal{A}_m(l) \triangleq e^{A_m(l)}; \\ \mathcal{S}_m(l) \triangleq e^{S_m(l)}. \end{cases} \tag{21}$$

Define the Mellin transform, denoted by $\mathcal{M}_{\mathcal{X}}(\theta_m)$, of a non-negative random variable $\mathcal{X}$ as follows [21]:

$$\mathcal{M}_{\mathcal{X}}(\theta_m) \triangleq \mathbb{E}\left[\mathcal{X}^{(\theta_m-1)}\right] \tag{22}$$

where $\theta_m > 0$ is defined as the QoS exponent. Denoting d_{th} a target delay, we can define the kernel function $\mathcal{K}(\theta_m, d_{\text{th}})$ as follows [20]:

$$\mathcal{K}(\theta_m, d_{\text{th}}) \triangleq \frac{\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)^{d_{\text{th}}}}{1-\mathcal{M}_{\mathcal{A}_m}(1+\theta_m)\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}, \tag{23}$$

if the following stability condition holds:

$$\mathcal{M}_{\mathcal{A}_m}(1+\theta_m)\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) < 1. \tag{24}$$

Correspondingly, an upper bound on the delay violation probability, denoted by $p_m(d_{\text{th}})$, can be obtained using the Mellin transform over the arrival and service processes $A_m(l)$ and $S_m(l)$ in the SNR-domain as follows:

$$p_m(d_{\text{th}}) \leq \inf_{\theta_m>0} \{\mathcal{K}(\theta_m, d_{\text{th}})\}. \tag{25}$$

C. Statistical Delay and Error-Rate Bounded QoS Provisioning for CF m-MIMO in the Finite Blocklength Regime

Statistical delay-bounded QoS guarantees [22] [23] have been extensively studied to analyze queuing behavior for time-varying arrival and service processes. The PDF of SINR γ_m over Rician wireless fading channels is given as follows [24]:

$$f_{\gamma_m}(\gamma_m) = 2(1+\kappa)\gamma_m e^{-(1+\kappa)\gamma_m-\kappa} I_0\left[2\sqrt{\kappa(1+\kappa)}\gamma_m\right] \tag{26}$$

where $I_0[\cdot]$ is the 0th order modified Bessel function of the first kind.

1) Mellin Transform Over Arrival Process: Assume that the arrivals at all time slots are independent and i.i.d. for each mobile user m, i.e., the accumulated source rate $A_m(l)$ has i.i.d. increments, denoted by $a_m(l)$, or equivalently $a_m = a_m(l)$ due to $a_m(l)$'s being i.i.d. We can derive the Mellin transform over accumulated arrival process, denoted by $\mathcal{M}_{\mathcal{A}_m}(\theta_m)$, as follows:

$$\begin{aligned} \mathcal{M}_{\mathcal{A}_m}(\theta_m) &= \mathbb{E}\left[\left(\prod_{j=1}^{l} e^{a_m(j)}\right)^{\theta_m-1}\right] = \left(\mathbb{E}\left[e^{a_m(\theta_m-1)}\right]\right)^l \\ &= \left[\mathcal{M}_{\alpha_m}(\theta_m)\right]^l \end{aligned} \tag{27}$$

where $\alpha_m = e^{a_m}$. Assume that the arrival process follows a Poisson distribution with average rate λ_m. We can derive the Mellin transform of α_m as follows:

$$\mathcal{M}_{\alpha_m}(\theta_m) = \sum_{i=1}^{\infty} e^{i(\theta_m-1)} \frac{(\lambda_m)^i}{i!} e^{-\lambda_m} = e^{\lambda_m\left(e^{\theta_m-1}-1\right)}. \tag{28}$$

2) Mellin Transform Over Service Process: Equations (23) and (25) show that deriving the closed-form expression of Mellin transform over service process at mobile user m is important for analyzing the delay violation probability, which motivates the following theorem.

Theorem 1: Given the statistical delay and error-rate bounded QoS provisioning $\{\theta_m, \epsilon_m\}$, the Mellin transform over service process, denoted by $\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)$, of mobile user m over Rician wireless fading channels in the high-end SNR region can be derived as follows:

$$\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) = (1-\epsilon_m)\left[F_1(\gamma_0) + F_0(\gamma_0)\right] + \epsilon_m \tag{29}$$

where $\gamma_0 \triangleq e^{\sqrt{\frac{V(\gamma_m)}{n_{\text{d}}}} Q^{-1}(\epsilon_m)}$ and

$$\begin{cases} F_0(\gamma_0) \triangleq \dfrac{2e^{-\kappa}}{\kappa+1} \displaystyle\sum_{i=0}^{\infty} \frac{\kappa^i}{(i!)^2} \gamma\left(i+2, (1+\kappa)\gamma_0\right); \\ F_1(\gamma_0) \triangleq 2e^{-\kappa}\left[e^{-\sqrt{n_{\text{d}}}} Q^{-1}(\epsilon_m)\right]^{\frac{\theta_m n_{\text{d}}}{\log 2}} \displaystyle\sum_{i=0}^{\infty} \frac{\kappa^i (\kappa+1)^{\frac{\theta_m n_{\text{d}}}{\log 2}-1}}{(i!)^2} \\ \qquad \times \Gamma\left(i+2-\dfrac{\theta_m n_{\text{d}}}{\log 2}, (1+\kappa)\gamma_0\right), \end{cases} \tag{30}$$

where $\log(\cdot)$ represents $\log_e(\cdot)$ and $\gamma(a,b)$ and $\Gamma(a,b)$ are the lower and upper incomplete Gamma functions, respectively.

Proof: Considering the decoding error at the receiver, we can derive the Mellin transform over service process as follows:

$$\begin{aligned} &\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) \\ &\quad = \mathbb{E}_{\gamma_m}\left[\epsilon_m + (1-\epsilon_m)e^{-\theta_m n_{\text{d}} R_m(n_{\text{d}},\epsilon_m)}\right] \\ &\quad = \int_0^{\infty} \left[\epsilon_m + (1-\epsilon_m)e^{-\theta_m n_{\text{d}} R_m(n_{\text{d}},\epsilon_m)}\right] f_{\gamma_m}(\gamma_m) d\gamma_m \end{aligned} \tag{31}$$

where $\mathbb{E}_{\gamma_m}[\cdot]$ is the expectation operation with respect to γ_m. We define

$$f_m(n_{\text{d}}, \epsilon_m) \triangleq \frac{1+\gamma_m}{\exp\left(\sqrt{\frac{V(\gamma_m)}{n_{\text{d}}}} Q^{-1}(\epsilon_m)\right)}. \tag{32}$$

Given the decoding error probability ϵ_m, the data rate $R_m(n_{\text{d}}, \epsilon_m)$ could become smaller than zero when the SINR is below a certain threshold γ_{th} [25]. As a result, the achievable data rate can be rewritten as follows:

$$R_m(n_{\text{d}}, \epsilon_m) = \max\left\{\log_2\left(f_m(n_{\text{d}}, \epsilon_m)\right), 0\right\}. \tag{33}$$

Accordingly, we can obtain the following equation:

$$\begin{aligned} &\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) \\ &\quad = (1-\epsilon_m)\Bigg\{ \int_0^{\infty} \left[f_m(n_{\text{d}}, \epsilon_m)\right]^{-\frac{\theta_m n_{\text{d}}}{\log 2}} \end{aligned}$$

$$\left.\times f_{\gamma_m}(\gamma_m)d\gamma_m + F_0(\gamma_m)\right\} + \epsilon_m$$

$$= (1-\epsilon_m)\left\{\int_{\gamma_0}^{\infty}\left[\frac{1+\gamma_m}{\exp\left\{\sqrt{\frac{V(\gamma_m)}{n_\mathrm{d}}}Q^{-1}(\epsilon_m)\right\}}\right]^{-\frac{\theta_m n_\mathrm{d}}{\log 2}}\right.$$
$$\left.\times f_{\gamma_m}(\gamma_m)d\gamma_m + F_0(\gamma_0)\right\} + \epsilon_m \tag{34}$$

where

$$F_0(\gamma_0) \triangleq \int_0^{\gamma_0} 2(1+\kappa)\gamma_m e^{-(1+\kappa)\gamma_m-\kappa} I_0\left[2\sqrt{\kappa(1+\kappa)}\gamma_m\right] d\gamma_m. \tag{35}$$

Then, we can expand the Bessel function into an infinite series and obtain the following equation [26]:

$$F_0(\gamma_0) = 2e^{-\kappa}\int_0^{\gamma_0}\gamma_m e^{-(1+\kappa)\gamma_m}\sum_{i=0}^{\infty}\frac{\kappa^i(\kappa+1)^{i+1}(\gamma_m)^i}{(i!)^2}d\gamma_m$$
$$= \frac{2e^{-\kappa}}{\kappa+1}\sum_{i=0}^{\infty}\frac{\kappa^i}{(i!)^2}\gamma\left(i+2,(1+\kappa)\gamma_0\right). \tag{36}$$

We define the following equation:

$$F_1(\gamma_0) \triangleq \int_{\gamma_0}^{\infty}\left[\frac{1+\gamma_m}{e^{\sqrt{\frac{V(\gamma_m)}{n_\mathrm{d}}}Q^{-1}(\epsilon_m)}}\right]^{-\frac{\theta_m n_\mathrm{d}}{\log 2}} 2(1+\kappa)\gamma_m$$
$$\times e^{-(1+\kappa)\gamma_m-\kappa} I_0\left[2\sqrt{\kappa(1+\kappa)}\gamma_m\right] d\gamma_m$$
$$= 2e^{-\kappa}\int_{\gamma_0}^{\infty}\left[\frac{1+\gamma_m}{e^{\sqrt{\frac{V(\gamma_m)}{n_\mathrm{d}}}Q^{-1}(\epsilon_m)}}\right]^{-\frac{\theta_m n_\mathrm{d}}{\log 2}}\gamma_m$$
$$\times e^{-(1+\kappa)\gamma_m}\sum_{i=0}^{\infty}\frac{\kappa^i(\kappa+1)^{i+1}(\gamma_m)^i}{(i!)^2}d\gamma_m. \tag{37}$$

In the high-end SNR region ($\gamma_m \gg 1$), the channel dispersion $V(\gamma_m) \to 1$. Correspondingly, we can rewrite Eq. (37) as follows:

$$F_1(\gamma_0)$$
$$= 2e^{-\kappa}\sum_{i=0}^{\infty}\frac{\kappa^i(\kappa+1)^{i+1}}{(i!)^2}\int_{\gamma_0}^{\infty}(\gamma_m)^{i+1-\frac{\theta_m n_\mathrm{d}}{\log 2}}e^{-(1+\kappa)\gamma_m}d\gamma_m$$
$$= 2e^{-\kappa}\left[e^{-\sqrt{n_\mathrm{d}}}Q^{-1}(\epsilon_m)\right]^{\frac{\theta_m n_\mathrm{d}}{\log 2}}\sum_{i=0}^{\infty}\frac{\kappa^i(\kappa+1)^{\frac{\theta_m n_\mathrm{d}}{\log 2}-1}}{(i!)^2}$$
$$\times\Gamma\left(i+2-\frac{\theta_m n_\mathrm{d}}{\log 2},(1+\kappa)\gamma_0\right). \tag{38}$$

Therefore, by substituting Eqs. (36) and (38) back into Eq. (34), we can obtain the results in Eq. (29), which completes the proof of Theorem 1. ■

Correspondingly, substituting Eqs. (27), (28), and (29) back into Eq. (23), we can derive the closed-form expression of the steady-state kernel $\mathcal{K}(\theta_m, d_\mathrm{th})$ as follows:

$$\mathcal{K}(\theta_m, d_\mathrm{th}) = \frac{\left\{(1-\epsilon_m)\left[F_1(\gamma_0)+F_0(\gamma_0)\right]+\epsilon_m\right\}^{d_\mathrm{th}}}{1-e^{\lambda_m l(e^{\theta_m}-1)}(1-\epsilon_m)\left[F_1(\gamma_0)+F_0(\gamma_0)\right]+\epsilon_m} \tag{39}$$

under the stability condition $\mathcal{M}_{\mathcal{A}_m}(1+\theta_m)\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) < 1$. As a result, we can derive the upper-bound on the delay violation probability by plugging Eq. (39) back into Eq. (25).

IV. Delay Analyses for Statistical Delay and Error-Rate Bounded QoS Provisioning in the Finite Blocklength Regime

In the previous Section III, we have investigated an upper on the delay violation probability using the Mellin transform for a given decoding error probability ϵ_m. In this section, assuming the decoding error probability is a function of $\{n_\mathrm{d}, \gamma_m\}$, we derive the delay violation probability in terms of the average decoding error probability function over Rayleigh wireless fading channels ($\kappa = 0$).

A. Upper-Bound on the Average Decoding Error Probability Function for CF m-MIMO in the Finite Blocklength Regime

Consider the case where $\kappa = 0$, i.e., the Rayleigh fading channel model. We define [11]

$$c_{k,m} \triangleq \frac{\sqrt{n_\mathrm{p}\mathcal{P}_\mathrm{p}}\beta_{k,m}}{\sum_{m'=1}^{K_\mathrm{u}} n_\mathrm{p}\mathcal{P}_\mathrm{p}\beta_{k,m'}\left\|\boldsymbol{\phi}_{m'}^{n_\mathrm{p}}\left(\boldsymbol{\phi}_{m'}^{n_\mathrm{p}}\right)^H\right\|^2+1}. \tag{40}$$

Considering the massive access scenario with vary large K_a, we can obtain the following equations by applying the Tcheby-shev's theorem for our proposed CF m-MIMO schemes across Rayleigh fading channel model [11]:

$$\begin{cases}\frac{1}{K_\mathrm{a}}\sum_{k=1}^{K_\mathrm{a}}(\boldsymbol{g}_{k,m})^T\boldsymbol{w}_{k,m} - \frac{N_\mathrm{T}\sqrt{n_\mathrm{p}\mathcal{P}_\mathrm{p}\overline{\mathcal{P}}_\mathrm{d}}}{K_\mathrm{a}}\sum_{k=1}^{K_\mathrm{a}}\eta_{k,m}(c_{k,m}\beta_{k,m}\chi_k)^2 \xrightarrow[K_\mathrm{a}\to\infty]{p} 0; \\ \frac{1}{K_\mathrm{a}}\sum_{k=1}^{K_\mathrm{a}}(\boldsymbol{g}_{k,m})^T\boldsymbol{w}_{k,m'} \xrightarrow[K_\mathrm{a}\to\infty]{p} 0, \qquad \forall m \neq m',\end{cases} \tag{41}$$

where the symbol $\xrightarrow[K_\mathrm{a}\to\infty]{p}$ represents the convergence in probability as $K_\mathrm{a} \to \infty$. The results given by Eq. (41) imply that as $K_\mathrm{a} \to \infty$, we only need to consider the desired parts of the received signal $y_m^{(l)}$ and ignore the noise and interference in the asymptotic analysis. As a result, we can show that the SINR γ_m follows the following distribution:

$$\gamma_m \sim (N_\mathrm{T})^2 n_\mathrm{p}\mathcal{P}_\mathrm{p}\overline{\mathcal{P}}_\mathrm{d}\sum_{k=1}^{K_\mathrm{a}}\eta_{k,m}\left(c_{k,m}\beta_{k,m}\chi_k\right)^2 \sim \sum_{k=1}^{K_\mathrm{a}}\mathcal{E}\left(\xi_{k,m}\right) \tag{42}$$

where $\mathcal{E}(\xi_{k,m})$ is the exponential distribution with its parameter equal to $\xi_{k,m}$, which is given as follows:

$$\xi_{k,m} \triangleq \frac{1}{2(N_\mathrm{T})^2 n_\mathrm{p}\mathcal{P}_\mathrm{p}\overline{\mathcal{P}}_\mathrm{d}\eta_{k,m}\left(c_{k,m}\beta_{k,m}\right)^2} \tag{43}$$

where $k \in \{1,\ldots,K_\mathrm{a}\}$ and $m \in \{1,\ldots,K_\mathrm{u}\}$. We can obtain the decoding error probability function, denoted by $\epsilon_m(n_\mathrm{d},\gamma_m)$, for mobile user m as follows [6]:

$$\epsilon_m(n_\mathrm{d},\gamma_m) \approx Q\left(\frac{C(\gamma_m)-R_m}{\sqrt{V(\gamma_m)/n_\mathrm{d}}}\right) \tag{44}$$

where $Q(\cdot)$ is the Q-function, R_m (bits per channel use) is the achievable data rate, and $C(\gamma_m)$ and $V(\gamma_m)$ denote the channel capacity and channel dispersion, respectively, given in Eq. (19). Similar to Eq. (31), given the achievable finite-blocklength coding rate R_m and the decoding error probability function $\epsilon_m(n_\mathrm{d}, \gamma_m)$, we can derive the Mellin transform over service process $S_m(l)$ at mobile user m as follows:

$$\begin{aligned}
&\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) \\
&= \mathbb{E}_{\gamma_m}\Big[\epsilon_m(n_\mathrm{d},\gamma_m) + [1-\epsilon_m(n_\mathrm{d},\gamma_m)]\, e^{-\theta_m n_\mathrm{d} R_m}\Big] \\
&= \mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)] + \{1-\mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)]\}\, e^{-\theta_m n_\mathrm{d} R_m}.
\end{aligned} \tag{45}$$

Equation (45) shows that deriving the average decoding error probability function $\mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)]$ is important to obtain the closed-form expression for Mellin transform over service process at mobile user m, motivating the theorem that follows.

Theorem 2: Given the achievable finite-blocklength coding rate R_m, the average decoding error probability function $\mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)]$ for mobile user m over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime is determined as follows:

$$\begin{aligned}
&\mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)] \\
&\approx 1-\sum_{k=1}^{K_\mathrm{a}}\left[1-e^{-\xi_{k,m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right]+\left[\frac{1}{2}+\vartheta_m\sqrt{n_\mathrm{d}}\right. \\
&\quad\left.\times\left(e^{R_m}-1\right)\right]\left[\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right. \\
&\quad\left.-\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right]-\vartheta_m\sqrt{n_\mathrm{d}}\Bigg\{\Bigg(2^{R_m-1} \\
&\quad-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\Bigg)\left[\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right]-\Bigg(2^{R_m-1} \\
&\quad+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\Bigg)\left[\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right] \\
&\quad+\sum_{k=1}^{K_\mathrm{a}}\left[\mathrm{Ei}\left(-\xi_{k,m}\left[2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right]\right)\right. \\
&\quad\left.-\mathrm{Ei}\left(-\xi_{k,m}\left[2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right]\right)\right]\Bigg\}
\end{aligned} \tag{46}$$

where $\mathrm{Ei}(x) \triangleq -\int_{-x}^{\infty}\frac{e^{-t}}{t}dt$ represents the exponential integral function and $\vartheta_m \triangleq \frac{1}{2\pi\sqrt{2^{2R_m}-1}}$.

Proof: To derive the average decoding error probability function, first we introduce an approximation of the Q-function as follows:

$$Q\left(\frac{C(\gamma_m)-R_m}{\sqrt{V(\gamma_m)/n_\mathrm{d}}}\right)\approx\Psi(\gamma_m) \tag{47}$$

where the function $\Psi(\gamma_m)$ is given as follows [27]:

$$\Psi(\gamma_m)=\begin{cases}1, & \gamma_m\le\zeta_{m,\mathrm{l}};\\ \frac{1}{2}-\vartheta_m\sqrt{n_\mathrm{d}}\left(\gamma_m-2^{R_m-1}\right), & \zeta_{m,\mathrm{l}}<\gamma_m<\zeta_{m,\mathrm{u}};\\ 0, & \gamma_m\ge\zeta_{m,\mathrm{u}},\end{cases} \tag{48}$$

where $\vartheta_m \triangleq \frac{1}{2\pi\sqrt{2^{2R_m}-1}}$, $\zeta_{m,\mathrm{l}} \triangleq 2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}$, and $\zeta_{m,\mathrm{u}} \triangleq 2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}$. Taking expectation over Eqs. (44) and (48), we can obtain the following equation:

$$\begin{aligned}
&\mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)] \\
&\approx F_{\gamma_m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)+\left[\frac{1}{2}+\vartheta_m\sqrt{n_\mathrm{d}}\left(e^{R_m}-1\right)\right] \\
&\quad\times\left[F_{\gamma_m}\left(2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)-F_{\gamma_m}\left(2^{R_m-1}\right.\right. \\
&\quad\left.\left.-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)\right]-\vartheta_m\sqrt{n_\mathrm{d}}\int_{2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}}^{2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}}\gamma_m f_{\gamma_m}(\gamma_m)d\gamma_m
\end{aligned} \tag{49}$$

where $F_{\gamma_m}(\gamma_m)$ is the cumulative probability function (CDF) of SINR γ_m. Using Eqs. (42) and (43), we can derive the CDF of SINR as follows:

$$F_{\gamma_m}(\gamma_m)=1-\sum_{k=1}^{K_\mathrm{a}}\left(1-e^{-\xi_{k,m}\gamma_m}\right). \tag{50}$$

Plugging Eq. (50) back into Eq. (49), we obtain:

$$\begin{aligned}
&\mathbb{E}_{\gamma_m}[\epsilon_m(n_\mathrm{d},\gamma_m)] \\
&\approx 1-\sum_{k=1}^{K_\mathrm{a}}\left[1-e^{-\xi_{k,m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right]+\left[\frac{1}{2}+\vartheta_m\sqrt{n_\mathrm{d}}\right. \\
&\quad\left.\times\left(e^{R_m}-1\right)\right]\left[\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right. \\
&\quad\left.-\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right]-\vartheta_m\sqrt{n_\mathrm{d}} \\
&\quad\times\Bigg\{\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)\left[\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right] \\
&\quad-\left(2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)\left[\sum_{k=1}^{K_\mathrm{a}}e^{-\xi_{k,m}\left(2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}\right)}\right] \\
&\quad+\sum_{k=1}^{K_\mathrm{a}}\int_{2^{R_m-1}-\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}}^{2^{R_m-1}+\frac{1}{2\vartheta_m\sqrt{n_\mathrm{d}}}}\frac{e^{-\xi_{k,m}\gamma_m}}{\gamma_m}d\gamma_m\Bigg\}
\end{aligned} \tag{51}$$

which, by employing the exponential integral function defined at the bottom of Eq. (46), leads to Eq. (46), completing the proof of Theorem 2. ■

In the high-end SNR regime, we have $V(\gamma_m) = 1-(1+\gamma_m)^{-2} \to 1$. Using Eq. (44), we define the following function:

$$\widetilde{\Phi}(n_\mathrm{d},\gamma_m)\triangleq[C(\gamma_m)-R_m]\sqrt{n_\mathrm{d}}. \tag{52}$$

Considering the high-end SNR regime, we can derive the average decoding error probability function as detailed in the following lemma.

Lemma 2: The approximate average decoding error probability function $\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\rm d},\gamma_m)]$ for mobile user m in the high-end SNR regime is determined as follows:

$$\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\rm d},\gamma_m)] \approx \frac{\sqrt{\pi}}{2\sqrt{2n_{\rm d}}}\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\exp\left\{\frac{(\xi_{k,m})^2}{2n_{\rm d}}-\xi_{k,m}\nu_m\right\} \times\left[1-\operatorname{erf}\left(\frac{\xi_{k,m}}{\sqrt{2n_{\rm d}}}-\frac{\sqrt{n_{\rm d}}}{\sqrt{2}}\nu_m\right)\right] \tag{53}$$

where $\nu_m \triangleq 2^{R_m}-1$.

Proof: Using the Chernoff bound, we have $Q(x) \leq \frac{1}{2}e^{-\frac{x^2}{2}}$ if $x \geq 0$. We can derive an upper bound on the average decoding error probability function $\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\rm d},\gamma_m)]$ for mobile user m in the high-end SNR regime as follows:

$$\begin{aligned}\mathbb{E}_{\gamma_m}&[\epsilon_m(n_{\rm d},\gamma_m)]\\ &=\int_0^{\infty}\mathbb{E}_{\gamma_m}\left[Q\left(\widetilde{\Phi}(n_{\rm d},\gamma_m)\right)\right]f_{\gamma_m}(\gamma_m)d\gamma_m\\ &\leq\frac{1}{2}\int_0^{\infty}\mathbb{E}_{\gamma_m}\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right]f_{\gamma_m}(\gamma_m)d\gamma_m\end{aligned} \tag{54}$$

where $f_{\gamma_m}(\gamma_m)$ is the PDF of the SINR γ_m over Rayleigh wireless fading channels, which is given as follows:

$$f_{\gamma_m}(\gamma_m)=\sum_{k=1}^{K_{\rm a}}\xi_{k,m}e^{-\xi_{k,m}\gamma_m}. \tag{55}$$

Then, using Eq. (19), we can obtain:

$$\begin{aligned}\mathbb{E}_{\gamma_m}\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right]&=\mathbb{E}_{\gamma_m}\left[e^{-\frac{n_{\rm d}[\log_2(\gamma_m+1)-R_m]^2}{2}}\right]\\ &=\int_0^{\infty}e^{-\frac{n_{\rm d}[\log_2(\gamma_m+1)-R_m]^2}{2}}f_{\gamma_m}(\gamma_m)d\gamma_m.\end{aligned} \tag{56}$$

Since we can easily show that $\widetilde{\Phi}(n_{\rm d},\gamma_m)$ is concave in γ_m. As a result, we can derive a lower bound on the function $\mathbb{E}_{\gamma_m}\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right]$ when $\widetilde{\Phi}(n_{\rm d},\gamma_m)$ is replaced by its Taylor expansion at any point. Using the Taylor expansion of $\widetilde{\Phi}(n_{\rm d},\gamma_m)$ at $\gamma_m=\nu_m$, we can obtain the following equation:

$$\mathbb{E}_{\gamma_m}\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right] \geq \int_0^{\infty}\sum_{k=1}^{K_{\rm a}}e^{-\frac{n_{\rm d}}{2}(\gamma_m-\nu_m)^2}\xi_{k,m}e^{-\xi_{k,m}\gamma_m}d\gamma_m. \tag{57}$$

Letting $t_m \triangleq \sqrt{\frac{n_{\rm d}}{2}}(\gamma_m-\nu_m)$, we have

$$\begin{aligned}\mathbb{E}_{\gamma_m}&\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right]\\ &\geq\sqrt{\frac{2}{n_{\rm d}}}\sum_{k=1}^{K_{\rm a}}\xi_{k,m}e^{-\xi_{k,m}\nu_m}\int_{-\frac{\sqrt{n_{\rm d}}}{\sqrt{2}}\nu_m}^{\infty}e^{-(t_m)^2}e^{-\frac{\sqrt{2}\xi_{k,m}}{\sqrt{n_{\rm d}}}t_m}dt_m\\ &=\sqrt{\frac{2}{n_{\rm d}}}\sum_{k=1}^{K_{\rm a}}\xi_{k,m}e^{-\xi_{k,m}\nu_m}\int_{-\frac{\sqrt{n_{\rm d}}}{\sqrt{2}}\nu_m}^{\infty}e^{\frac{(\xi_{k,m})^2}{2n_{\rm d}}}\\ &\quad\times e^{-\left(t_m+\frac{\xi_{k,m}}{\sqrt{2n_{\rm d}}}\right)^2}dt_m.\end{aligned} \tag{58}$$

Letting $\widetilde{t}_m \triangleq t_m+\frac{\xi_{k,m}}{\sqrt{2n_{\rm d}}}$, we can derive a lower bound on the function $\mathbb{E}_{\gamma_m}\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right]$ as follows:

$$\begin{aligned}\mathbb{E}_{\gamma_m}&\left[e^{-\frac{\widetilde{\Phi}^2(n_{\rm d},\gamma_m)}{2}}\right]\\ &\geq\sqrt{\frac{2}{n_{\rm d}}}\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\exp\left\{\frac{(\xi_{k,m})^2}{2n_{\rm d}}-\xi_{k,m}\nu_m\right\}\\ &\quad\times\int_{\frac{\xi_{k,m}}{\sqrt{2n_{\rm d}}}-\frac{\sqrt{n_{\rm d}}}{\sqrt{2}}\nu_m}^{\infty}e^{-(\widetilde{t}_m)^2}d\widetilde{t}_m\\ &=\frac{\sqrt{\pi}}{2\sqrt{2n_{\rm d}}}\sum_{k=1}^{K_{\rm a}}\left\{\xi_{k,m}\exp\left\{\frac{(\xi_{k,m})^2}{2n_{\rm d}}-\xi_{k,m}\nu_m\right\}\right.\\ &\quad\left.\times\left[1-\operatorname{erf}\left(\frac{\xi_{k,m}}{\sqrt{2n_{\rm d}}}-\frac{\sqrt{n_{\rm d}}}{\sqrt{2}}\nu_m\right)\right]\right\}\end{aligned} \tag{59}$$

where $\operatorname{erf}(x) \triangleq \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the error function. As a result, by plugging Eq. (59) back into Eq. (54), we can obtain Eq. (53), which completes the proof of Lemma 2. ∎

B. Performance Analyses and Rate Adaptation for Statistical Delay and Error-Rate Bounded QoS Provisioning

Substituting Eqs. (28), (29), and (46) back into Eq. (23), we can derive the closed-form expression for the kernel function $\mathcal{K}(\theta_m, d_{\rm th})$. Correspondingly, we formulate the delay violation probability minimization problem as follows:

$$\mathbf{P_1}: R_m^{\rm opt}=\arg\min_{R_m}\{p_m(d_{\rm th})\}=\arg\min_{R_m}\{\mathcal{K}_m(\theta_m,d_{\rm th})\}. \tag{60}$$

Using Eq. (23), we can convert $\mathbf{P_1}$ into an equivalent minimization problem $\mathbf{P_2}$ as in the following equation:

$$\begin{aligned}\mathbf{P_2}: R_m^{\rm opt}&=\arg\min_{R_m}\{\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)\}\\ &=\arg\min_{R_m}\left\{\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\rm d},\gamma_m)]+\left[1-e^{-\theta_m n_{\rm d}R_m}\right.\right.\\ &\quad\left.\left.\times\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\rm d},\gamma_m)]\right]\right\}.\end{aligned} \tag{61}$$

The monotonicity of decoding error probability function $\epsilon_m(n_{\rm d},\gamma_m)$ plays an important role in analyzing the convexity of $\mathbf{P_2}$ given in Eq. (60) as detailed in the following lemma.

Lemma 3: The decoding error probability function $\epsilon_m(n_{\rm d},\gamma_m)$ is a monotonically increasing function with respect to the achievable data rate R_m for our proposed CF m-MIMO modeling schemes.

Proof: To prove the monotonicity of the decoding error probability function $\epsilon_m(n_{\rm d},\gamma_m)$, using Eq. (44), we can take the first-order derivative of $\epsilon_m(n_{\rm d},\gamma_m)$ with respect to R_m as in the following equation:

$$\frac{\partial\epsilon_m(n_{\rm d},\gamma_m)}{\partial R_m}=-\frac{1}{\sqrt{2\pi}}e^{-\frac{\Phi^2(n_{\rm d},\gamma_m)}{2}}\left[\frac{\partial\Phi(n_{\rm d},\gamma_m)}{\partial R_m}\right] \tag{62}$$

where $\Phi\left(n_{\mathrm{d}}, \gamma_m\right) \triangleq \frac{C(\gamma_m)-R_m}{\sqrt{V(\gamma_m)/n_{\mathrm{d}}}}$ due to Eq. (44) and thus its first-order derivative yields the following equations:

$$\frac{\partial \Phi\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m}=\frac{-\sqrt{n_{\mathrm{d}}}}{\sqrt{1-\frac{1}{(1+\gamma_m)^2}}}<0. \tag{63}$$

Therefore, we have $\frac{\partial \epsilon_m(n_{\mathrm{d}}, \gamma_m)}{\partial R_m}>0$, which implies that the decoding error probability function $\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)$ is a monotonically increasing function of R_m, completing the proof. ∎

Lemma 3 and its proof can help further investigate the convexity of $\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)$ as shown in the following lemma.

Lemma 4: The block error probability function $\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)$ is convex with respect to the achievable data rate R_m for each mobile user m.

Proof: Applying Eq. (63), we can derive the second-order derivative of the function $\Phi\left(n_{\mathrm{d}}, \gamma_m\right)$ with respect to R_m as follows:

$$\frac{\partial^2 \Phi\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m^2}=0. \tag{64}$$

Using Eqs. (44), (63), and (64) and the fact that $\Phi\left(n_{\mathrm{d}}, \gamma_m\right)>0$ due to $C\left(\gamma_m\right)>R_m$, we obtain the following equations:

$$\frac{\partial^2 \epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m^2}=\frac{1}{\sqrt{2\pi}}e^{-\frac{\Phi^2(n_{\mathrm{d}}, \gamma_m)}{2}}\left\{\Phi(n_{\mathrm{d}}, \gamma_m)\left[\frac{\partial \Phi\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m}\right]^2 -\frac{\partial^2 \Phi\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m^2}\right\}>0 \tag{65}$$

which implies that $\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)$ is a convex function with respect to R_m. Therefore, we complete the proof. ∎

Combining and extending Theorem 2, Lemma 3, and Lemma 4 yield the following theorem.

Theorem 3: If the statistical delay and error-rate bounded QoS constraints are characterized by $\{\theta_m, \epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\}$ with $\theta_m>0$, then the following claims hold for each mobile user m.

Claim 1. The delay upper-bound violation probability minimization problem $\mathbf{P_2}$ given by Eq. (61) is *strictly convex* with respect to the achievable data rate R_m for our proposed CF m-MIMO modeling schemes.

Claim 2. Our obtained Mellin transform function $\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)$ over service process satisfies the following condition:

$$\frac{\partial^2 \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial R_m^2}>0. \tag{66}$$

Claim 3. The unique optimal rate adaptation policy, denoted by R_m^{opt}, for each mobile user m in the high-end SNR region is given by the following equation:

$$\begin{aligned}R_m^{\mathrm{opt}}\approx\log_2\Bigg\{&\frac{\theta_m n_{\mathrm{d}}}{(\log 2)\sum_{k=1}^{K_{\mathrm{a}}}\xi_{k,m}}\Bigg[\frac{(\log 2)}{\theta_m n_{\mathrm{d}}}\left(\sum_{k=1}^{K_{\mathrm{a}}}\xi_{k,m}\right)\\&\times\left(1-\frac{\theta_m\sqrt{2\pi n_{\mathrm{d}}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\mathrm{d}}}}-\mathcal{W}\Bigg(-\frac{(\log 2)\sum_{k=1}^{K_{\mathrm{a}}}\xi_{k,m}}{\theta_m n_{\mathrm{d}}}\\&\times\left[\frac{2\theta_m n_{\mathrm{d}}}{\sum_{k=1}^{K_{\mathrm{a}}}\xi_{k,m}\exp\left\{\frac{(\xi_{k,m})^2}{2n_{\mathrm{d}}}+\xi_{k,m}\right\}}\right]^{\frac{\log 2}{\theta_m n_{\mathrm{d}}}}\\&\times\exp\left\{\frac{(\log 2)}{\theta_m n_{\mathrm{d}}}\left(\sum_{k=1}^{K_{\mathrm{a}}}\xi_{k,m}\right)\right.\\&\left.\times\left(1-\frac{\theta_m\sqrt{2\pi n_{\mathrm{d}}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\mathrm{d}}}}\right\}\Bigg)\Bigg]\Bigg\}\end{aligned} \tag{67}$$

where $\mathcal{W}(\cdot)$ is the Lambert W function.

Proof: For presenting convenience, we start with Claim 2.

Claim 2. To analyze the convexity of the minimization problem $\mathbf{P_2}$, we apply the chain rule for second-order derivative and obtain the following equation:

$$\begin{aligned}\frac{\partial^2 \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial R_m^2}=&\frac{\partial^2 \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial \epsilon_m^2\left(n_{\mathrm{d}}, \gamma_m\right)}\left[\frac{\partial \epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m}\right]^2\\&+\frac{\partial \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial \epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)}\frac{\partial^2 \epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)}{\partial R_m^2}.\end{aligned} \tag{68}$$

To analyze the convexity of the Mellin transform over service process $\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)$, first we need to investigate the following equations when $\theta_m>0$:

$$\begin{cases}\dfrac{\partial \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial \epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)}=1-e^{-\theta_m n_{\mathrm{d}} R_m}>0;\\[2ex]\dfrac{\partial^2 \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial \epsilon_m^2\left(n_{\mathrm{d}}, \gamma_m\right)}=0.\end{cases} \tag{69}$$

Then, using Eq. (65) and (69), we can show that $\frac{\partial^2 \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial R_m^2}>0$, completing the proof of Claim 2 in Theorem 3.

Claim 1. Applying the sufficient conditions for convexity, we can prove that the Mellin transform function $\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)$ is strictly convex with respect to the achievable data rate R_m for $\theta_m>0$, completing the proof of Claim 1 in Theorem 3.

Claim 3. Due to the property of strict convexity and uniqueness of optimal solutions [28], there exists the unique optimal rate adaptation policy R_m^{opt} that minimizes problem $\mathbf{P_2}$ given by Eq. (61) for each mobile user m when $\theta_m>0$. Taking the first-order derivative of the Mellin transform function $\mathcal{M}_{\mathcal{S}_m}$ with respect to R_m, we can obtain the following equation:

$$\begin{aligned}\frac{\partial \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial R_m}=&\frac{\partial \mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\right]}{\partial R_m}-\frac{\partial \mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\right]}{\partial R_m}\\&\times e^{-\theta_m n_{\mathrm{d}} R_m}-\left\{1-\mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\right]\right\}\\&\times\theta_m n_{\mathrm{d}} e^{-\theta_m n_{\mathrm{d}} R_m}.\end{aligned} \tag{70}$$

Setting $\frac{\partial \mathcal{M}_{\mathcal{S}_m}(1-\theta_m)}{\partial R_m}=0$, we have

$$\begin{aligned}\frac{\partial \mathbb{E}_{\gamma_m}[\epsilon_m(n_{\mathrm{d}}, \gamma_m)]}{\partial R_m}&\left(1-e^{-\theta_m n_{\mathrm{d}} R_m}\right)-\left\{1-\mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\right]\right\}\\&\times\theta_m n_{\mathrm{d}} e^{-\theta_m n_{\mathrm{d}} R_m}=0.\end{aligned} \tag{71}$$

We can rewrite Eq. (71) as follows:

$$e^{\theta_m n_{\mathrm{d}} R_m}-1=\frac{\theta_m n_{\mathrm{d}}\left(1-\mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\right]\right)}{\frac{\partial \mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\mathrm{d}}, \gamma_m\right)\right]}{\partial R_m}}. \tag{72}$$

Using Eqs. (62) and (63), we can obtain the following equation:

$$R_m = \frac{1}{\theta_m n_{\rm d}} \log\left\{1+\frac{\theta_m n_{\rm d}\left(1-\mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\rm d},\gamma_m\right)\right]\right)}{\frac{\partial \mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\rm d},\gamma_m\right)\right]}{\partial R_m}}\right\}$$
$$= \frac{1}{\theta_m n_{\rm d}} \log\left\{1+\frac{\theta_m\sqrt{2\pi n_{\rm d}}\left\{1-\mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\rm d},\gamma_m\right)\right]\right\}}{\mathbb{E}_{\gamma_m}\left[e^{-\frac{\Phi^2\left(n_{\rm d},\gamma_m\right)}{2}}\left(\frac{1}{\sqrt{1-\frac{1}{(1+\gamma_m)^2}}}\right)\right]}\right\} \tag{73}$$

where $\mathbb{E}_{\gamma_m}\left[\epsilon_m\left(n_{\rm d},\gamma_m\right)\right]$ is given by Eq. (46). In the high-end SNR regime, we have $V\left(\gamma_m\right)=1-\left(1+\gamma_m\right)^{-2}\to 1$. Using the Chernoff bound, we have $Q(x)\le \frac{1}{2}e^{-\frac{x^2}{2}}$ when $x\ge 0$. Applying Eq. (73), we can obtain the following equation:

$$R_m=\frac{1}{\theta_m n_{\rm d}}\log\left\{1+\frac{\theta_m\sqrt{2\pi n_{\rm d}}\left\{1-\mathbb{E}_{\gamma_m}\left[Q\left(\widetilde{\Phi}\left(n_{\rm d},\gamma_m\right)\right)\right]\right\}}{\mathbb{E}_{\gamma_m}\left[e^{-\frac{\tilde{\Phi}^2\left(n_{\rm d},\gamma_m\right)}{2}}\right]}\right\}$$
$$\ge\frac{1}{\theta_m n_{\rm d}}\log\left\{1+\frac{\theta_m\sqrt{2\pi n_{\rm d}}\left(1-\mathbb{E}_{\gamma_m}\left[\frac{1}{2}e^{-\frac{\tilde{\Phi}^2\left(n_{\rm d},\gamma_m\right)}{2}}\right]\right)}{\mathbb{E}_{\gamma_m}\left[e^{-\frac{\tilde{\Phi}^2\left(n_{\rm d},\gamma_m\right)}{2}}\right]}\right\}$$
$$=\frac{1}{\theta_m n_{\rm d}}\log\left\{1+\theta_m\sqrt{2\pi n_{\rm d}}\left\{\frac{1}{\mathbb{E}_{\gamma_m}\left[e^{-\frac{\tilde{\Phi}^2\left(n_{\rm d},\gamma_m\right)}{2}}\right]}-\frac{1}{2}\right\}\right\}. \tag{74}$$

Then, plugging Eq. (59) back into Eq. (74), we have:

$$R_m \approx \frac{1}{\theta_m n_{\rm d}}\log\left\{1+\theta_m\sqrt{2\pi n_{\rm d}}\left\{-\frac{1}{2}+\frac{\sqrt{2n_{\rm d}}}{\sqrt{\pi}}\right.\right.$$
$$\times\left[\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}-\xi_{k,m}\nu_m\right\}\right.$$
$$\left.\left.\left.\times\left\{1-\operatorname{erf}\left(\frac{\xi_{k,m}}{\sqrt{2n_{\rm d}}}-\frac{\sqrt{n_{\rm d}}}{\sqrt{2}}\nu_m\right)\right\}\right]^{-1}\right\}\right\}. \tag{75}$$

Since the error function $-1\le \operatorname{erf}(x)\le 1$, we can remove the error function from Eq. (75) and obtain the following equation:

$$e^{\theta_m n_{\rm d} R_m}\approx 1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}+\frac{2\theta_m n_{\rm d}}{\sum_{k=1}^{K_{\rm a}}\xi_{k,m}}$$
$$\times\left\{\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}-\xi_{k,m}\nu_m\right\}\right\}^{-1}. \tag{76}$$

Substituting $\nu_m=2^{R_m}-1$ back into Eq. (76), we have

$$e^{\theta_m n_{\rm d} R_m}-\sum_{k=1}^{K_{\rm a}}\left[\frac{2\theta_m n_{\rm d} e^{\xi_{k,m}2^{R_m}}}{\xi_{k,m}\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}+\xi_{k,m}\right\}}\right]\approx 1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2} \tag{77}$$

where $\xi_{k,m}$ is given by Eq. (43). Let $z_m \triangleq \sum_{k=1}^{K_{\rm a}}\xi_{k,m}2^{R_m}$. We have

$$\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)^{-\frac{\theta_m n_{\rm d}}{(\log 2)}}\left(z_m\right)^{\frac{\theta_m n_{\rm d}}{(\log 2)}}-2\theta_m n_{\rm d} e^{z_m}\left\{\sum_{k=1}^{K_{\rm a}}\left[\xi_{k,m}\right.\right.$$
$$\left.\left.\times\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}+\xi_{k,m}\right\}\right]\right\}^{-1}\approx 1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}. \tag{78}$$

Then, solving for the $\frac{\theta_m n_{\rm d}}{\log 2}$th root on both sides of Eq. (78), we can obtain:

$$\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)^{-1}z_m - e^{z_m\frac{\log 2}{\theta_m n_{\rm d}}}$$
$$\times\left[\frac{2\theta_m n_{\rm d}}{\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}+\xi_{k,m}\right\}}\right]^{\frac{\log 2}{\theta_m n_{\rm d}}}$$
$$\approx\left[1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}\right]^{\frac{\log 2}{\theta_m n_{\rm d}}}. \tag{79}$$

Let $\widetilde{z}_m \triangleq z_m\frac{\log 2}{\theta_m n_{\rm d}}$. After some algebra manipulations, we have

$$\frac{(\log 2)\sum_{k=1}^{K_{\rm a}}\xi_{k,m}}{\theta_m n_{\rm d}}\left[\frac{2\theta_m n_{\rm d}}{\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}+\xi_{k,m}\right\}}\right]^{\frac{\log 2}{\theta_m n_{\rm d}}} e^{\widetilde{z}_m}$$
$$\approx\widetilde{z}_m-\frac{(\log 2)}{\theta_m n_{\rm d}}\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)\left[1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}\right]^{\frac{\log 2}{\theta_m n_{\rm d}}}. \tag{80}$$

Then, multiplying the following expression on both sides of Eq. (80):

$$\left[e^{-\widetilde{z}_m}+\frac{(\log 2)}{\theta_m n_{\rm d}}\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)\left[1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2(\log 2)}\right]^{\frac{\log 2}{\theta_m n_{\rm d}}}\right], \tag{81}$$

we can obtain:

$$\frac{(\log 2)\sum_{k=1}^{K_{\rm a}}\xi_{k,m}}{\theta_m n_{\rm d}}\left[\frac{2\theta_m n_{\rm d}}{\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\exp\left\{\frac{\left(\xi_{k,m}\right)^2}{2n_{\rm d}}+\xi_{k,m}\right\}}\right]^{\frac{\log 2}{\theta_m n_{\rm d}}}$$
$$\times\exp\left\{\frac{(\log 2)}{\theta_m n_{\rm d}}\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)\left(1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\rm d}}}\right\}$$
$$\approx\left[\widetilde{z}_m-\frac{(\log 2)}{\theta_m n_{\rm d}}\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)\left(1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\rm d}}}\right]$$
$$\times\exp\left\{-\widetilde{z}_m+\frac{(\log 2)}{\theta_m n_{\rm d}}\left(\sum_{k=1}^{K_{\rm a}}\xi_{k,m}\right)\right.$$
$$\left.\times\left(1-\frac{\theta_m\sqrt{2\pi n_{\rm d}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\rm d}}}\right\}. \tag{82}$$

Therefore, using the Lambert W function [29], we can derive the approximate optimal rate adaptation policy R_m^{opt} for our proposed FBC based CF m-MIMO schemes in the high-end SNR region as follows:

$$
\begin{aligned}
R_m^{\text{opt}} \approx \log_2 &\left\{ \frac{\theta_m n_{\text{d}}}{(\log 2) \sum_{k=1}^{K_{\text{a}}} \xi_{k,m}} \left[\frac{(\log 2)}{\theta_m n_{\text{d}}} \left(\sum_{k=1}^{K_{\text{a}}} \xi_{k,m} \right) \right.\right. \\
&\times \left(1 - \frac{\theta_m \sqrt{2\pi n_{\text{d}}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\text{d}}}} - \mathcal{W}\left(- \frac{(\log 2) \sum_{k=1}^{K_{\text{a}}} \xi_{k,m}}{\theta_m n_{\text{d}}} \right. \\
&\times \left[\frac{2\theta_m n_{\text{d}}}{\sum_{k=1}^{K_{\text{a}}} \xi_{k,m} \exp\left\{ \frac{(\xi_{k,m})^2}{2n_{\text{d}}} + \xi_{k,m} \right\}} \right]^{\frac{\log 2}{\theta_m n_{\text{d}}}} \\
&\times \exp\left\{ \frac{(\log 2)}{\theta_m n_{\text{d}}} \left(\sum_{k=1}^{K_{\text{a}}} \xi_{k,m} \right) \right. \\
&\left.\left.\left.\left. \times \left(1 - \frac{\theta_m \sqrt{2\pi n_{\text{d}}}}{2}\right)^{\frac{\log 2}{\theta_m n_{\text{d}}}} \right\} \right) \right] \right\}. \qquad (83)
\end{aligned}
$$

Thus, we complete the proof of Claim 3 in Theorem 3. ■

Remarks on Theorem 3: Claim 1 guarantees the existence of the optimal solution to the optimization problem $\mathbf{P_2}$ given by Eq. (61) for our proposed CF m-MIMO modeling schemes when $\theta_m > 0$. Claim 2 shows that our obtained Mellin transform function over service process can characterize the convexity of $\mathbf{P_2}$. Claim 3 derives the closed-form expression of the optimal rate adaptation policy R_m^{opt} as the function of $\{\theta_m, \epsilon_m(n_{\text{d}}, \gamma_m)\}$ for each mobile user m in the high-end SNR region, which plays the important roles in the system designs and performance analyses for statistical delay and error-rate bounded QoS provisioning over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime.

C. Maximizing Effective Capacity for Statistical Delay and Error-Rate Bounded QoS Constraints in the Finite Blocklength Regime

The proposed upper bound on the delay violation probability $p_m(d_{\text{th}})$ using SNC characterizes the small values of the target delay d_{th}. For analyzing fairly long delays, i.e., the tail of delay distribution, researchers have proposed the concept of *effective capacity* to proximate the delay violation probability $p_m(d_{\text{th}})$. The *effective capacity* [30] is defined as the maximum constant arrival rate for a given service process subject to statistical delay-bounded QoS constraints. We can derive the effective capacity, denoted by $EC_m(\theta_m)$, in terms of the accumulated service process $S_m(l)$ for mobile user m in the bit domain as in the following equation:

$$EC_m(\theta_m) \triangleq -\lim_{l\to\infty} \frac{1}{l\theta_m} \log\left(\mathbb{E}_\gamma \left[e^{-\theta S_m(l)} \right] \right). \qquad (84)$$

On the other hand, considering the SNC, we can redefine the effective capacity, denoted by $EC_m(\theta_m)$, using the Mellin transform over service process in the SNR-domain as follows:

$$EC_m(\theta_m) \triangleq -\frac{1}{\theta_m} \log\left(\mathcal{M}_{\mathcal{S}_m}(1-\theta_m) \right). \qquad (85)$$

Accordingly, we can formulate the optimization problem $\mathbf{P_3}$ for statistical delay and error-rate bounded QoS provisioning $\{\theta_m, \epsilon_m(n_{\text{d}}, \gamma_m)\}$ to maximize the downlink effective capacity $EC_m(\theta_m)$ at mobile user m for our proposed CF m-MIMO schemes in the finite blocklength regime as follows:

$$\mathbf{P_3}: \arg\max_{R_m} \{EC_m(\theta_m)\}. \qquad (86)$$

Since $\log(\cdot)$ is monotonically increasing, the above maximization problem $\mathbf{P_3}$ can be converted into an equivalent minimization problem $\mathbf{P_4}$ as follows:

$$\mathbf{P_4}: \arg\min_{R_m} \{\mathcal{M}_{\mathcal{S}_m}(1-\theta_m)\}. \qquad (87)$$

which is equivalent to the minimization problem $\mathbf{P_2}$ given in Eq. (61). Consequently, we can show that our derived optimal rate adaptation policy R_m^{opt} given in Eq. (67) also maximizes the effective capacity $EC_m(\theta_m)$ given in Eq. (85) for mobile user m in the finite blocklength regime considering the high-end SNR region. Therefore, using Eq. (45), (53), and (67), we can derive the maximum effective capacity, denoted by $EC_m^{\max}(\theta_m)$, for statistical delay and error-rate bounded QoS provisioning in supporting mURLLC over 6G CF m-MIMO and FBC mobile wireless networks in the high-end SNR region as follows:

$$
\begin{aligned}
&EC_m^{\max}(\theta_m) \\
&\approx -\frac{1}{\theta_m} \log \left\{ \frac{\sqrt{\pi}}{2\sqrt{2n_{\text{d}}}} \sum_{k=1}^{K_{\text{a}}} \left\{ \xi_{k,m} \exp\left\{ \frac{(\xi_{k,m})^2}{2n_{\text{d}}} - \xi_{k,m}\nu_m \right\} \right.\right. \\
&\quad \left. \times \left[1 - \text{erf}\left(\frac{\xi_{k,m}}{\sqrt{2n_{\text{d}}}} - \sqrt{\frac{n_{\text{d}}}{2}} \nu_m \right) \right] \right\} \\
&\quad + \left(1 - \frac{\sqrt{\pi}}{2\sqrt{2n_{\text{d}}}} \sum_{k=1}^{K_{\text{a}}} \left\{ \xi_{k,m} \exp\left\{ \frac{(\xi_{k,m})^2}{2n_{\text{d}}} - \xi_{k,m}\nu_m \right\} \right.\right. \\
&\quad \left.\left. \times \left[1 - \text{erf}\left(\frac{\xi_{k,m}}{\sqrt{2n_{\text{d}}}} - \sqrt{\frac{n_{\text{d}}}{2}} \nu_m \right) \right] \right\} \right) \\
&\quad \times \left\{ \frac{\theta_m n_{\text{d}}}{(\log 2) \sum_{k=1}^{K_{\text{a}}} \xi_{k,m}} \left[\frac{(\log 2)}{\theta_m n_{\text{d}}} \left(\sum_{k=1}^{K_{\text{a}}} \xi_{k,m} \right) \right.\right. \\
&\quad \times \left[1 - \frac{\theta_m \sqrt{2\pi n_{\text{d}}}}{2} \right]^{\frac{\log 2}{\theta_m n_{\text{d}}}} - \mathcal{W}\left(- \frac{(\log 2) \sum_{k=1}^{K_{\text{a}}} \xi_{k,m}}{\theta_m n_{\text{d}}} \right. \\
&\quad \times \left[\frac{2\theta_m n_{\text{d}}}{\sum_{k=1}^{K_{\text{a}}} \xi_{k,m} \exp\left\{ \frac{(\xi_{k,m})^2}{2n_{\text{d}}} + \xi_{k,m} \right\}} \right]^{\frac{\log 2}{\theta_m n_{\text{d}}}}
\end{aligned}
$$

$$
\times \exp\left[\frac{(\log 2)}{\theta_m n_{\mathrm{d}}}\left(\sum_{k=1}^{K_{\mathrm{a}}}\xi_{k,m}\right)\right.
$$
$$
\left.\left.\left.\left.\times\left(1-\frac{\theta_m\sqrt{2\pi n_{\mathrm{d}}}}{2}\right)^{\frac{(\log 2)}{\theta_m n_{\mathrm{d}}}}\right]\right)\right]\right\}^{-\frac{\theta_m n_{\mathrm{d}}}{\log 2}}\Bigg\}. \tag{88}
$$

V. Performance Evaluations

We use MATLAB-based simulations to validate and evaluate our proposed CF m-MIMO based schemes for statistical delay and error-rate bounded QoS provisioning in supporting mURLLC in the finite blocklength regime. Throughout our simulations, we set the number of APs $K_{\mathrm{a}} \in [50, 900]$, the number of mobile users $K_{\mathrm{u}} \in [10, 700]$, the number of transmit antennas $N_{\mathrm{T}} \in [2, 10]$, the uplink pilot transmit power $\overline{\mathcal{P}}_{\mathrm{p}}$ from $[1, 10]$ Watt for each mobile user, the average downlink transmit power $\overline{\mathcal{P}}_{\mathrm{d}}$ from $[1, 40]$ Watt for each mobile user, and the Rician factor κ from $[0, 30]$.

We set the number of Rician factor $\kappa = 4$, the number of uplink channel uses $n_{\mathrm{p}} = 100$, the number of transmit antennas $N_{\mathrm{T}} = 10$, and the decoding error probability $\epsilon_m = 10^{-6}$. Compared with the classical least-square (LS) channel estimator, Fig. 2 plots the achievable data transmission rate with varying numbers of APs K_{a} for our proposed 6G CF m-MIMO mobile wireless networks over Rician wireless fading channels in the finite blocklength regime. We can observe from Fig. 2 that the achievable data transmission rate increases with the number of APs. It is shown in Fig. 2 that the MMSE channel estimator performs better than the LS estimator over Rician wireless fading channels in terms of the achievable data transmission rate. Fig. 2 also shows that the gap between the MMSE estimator and LS estimator increases with K_{a}, which is because of the channel hardening effect.

Setting the number of transmit antennas $N_{\mathrm{T}} = 10$ and the decoding error probability $\epsilon_m = 10^{-6}$, Fig. 3 depicts the achievable data transmission rate with different Rician factors κ for our proposed 6G CF m-MIMO mobile wireless networks in the finite blocklength regime. We can observe from Fig. 3 that the achievable data rate increases as the Rician factor κ increases. Traditionally, the CSI estimation is not good enough when K_{a} is small, which leads to a low data rate. For our proposed CF m-MIMO scheme, the channel estimation quality can be significantly improved with large number of APs K_{a}.

Now we set the number of APs $K_{\mathrm{a}} = 100$, the number of downlink channel uses $n_{\mathrm{d}} = 800$, the average downlink transmit power $\overline{\mathcal{P}}_{\mathrm{d}} = 20$ Watt for each AP, and the Rician factor $\kappa = 2$. Fig. 4 plots the CDFs of the downlink data transmission rate per user for our proposed CF m-MIMO schemes in the finite blocklength regime. As shown in Fig. 4, the downlink data transmission rate per user increases with the number of transmit antennas N_{T}. In addition, Fig. 4 shows that a higher multiplexing order P_m is more beneficial for our proposed CF m-MIMO schemes with larger antenna arrays.

We set the number of APs $K_{\mathrm{a}} = 100$, the number of transmit antennas $N_{\mathrm{T}} = 10$, the multiplexing order $P_m = 2$, and decoding error probability $\epsilon_m = 10^{-6}$. Compared with the

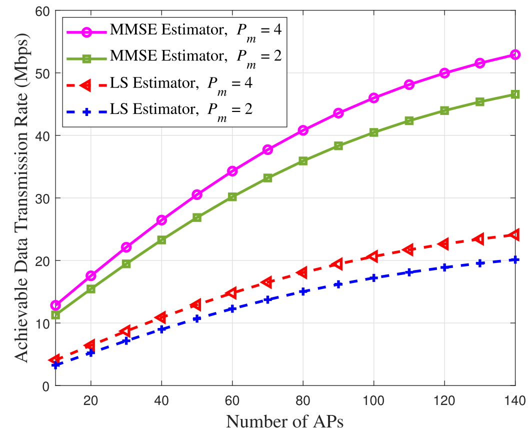


Fig. 2. The achievable data transmission rate vs. number of APs for our proposed CF m-MIMO scheme in the finite blocklength regime.

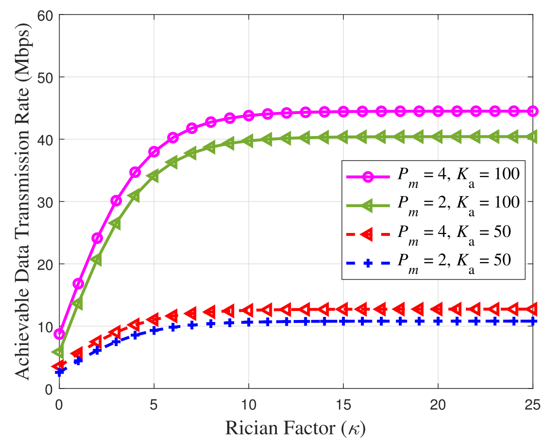


Fig. 3. The achievable data transmission rate vs. Rician factor κ for our proposed CF m-MIMO scheme in the finite blocklength regime.

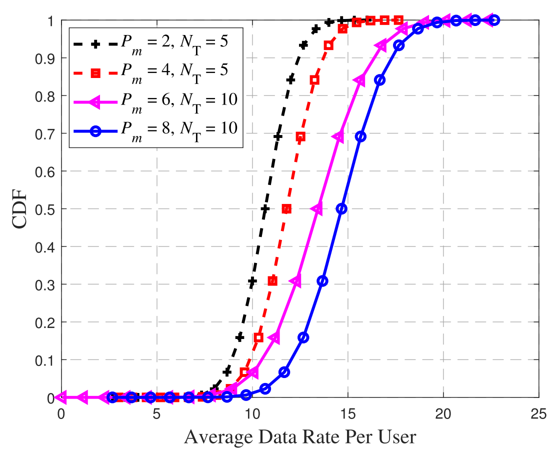


Fig. 4. The CDFs of downlink data transmission rate per user for CF m-MIMO schemes in the finite blocklength regime.

traditional Shannon's theorem which requires infinite blocklength, Fig. 5 plots the delay violation probability $p_m(d_{\mathrm{th}})$ with different target delays d_{th} over Rician wireless fading channels for our proposed CF m-MIMO scheme in the finite blocklength regime. It is shown in Fig. 5 that the delay violation probability $p_m(d_{\mathrm{th}})$ decreases as the target delay d_{th} increases. Fig. 5 also shows the delay violation probability increases with the increased average data arrival rate λ_m. This

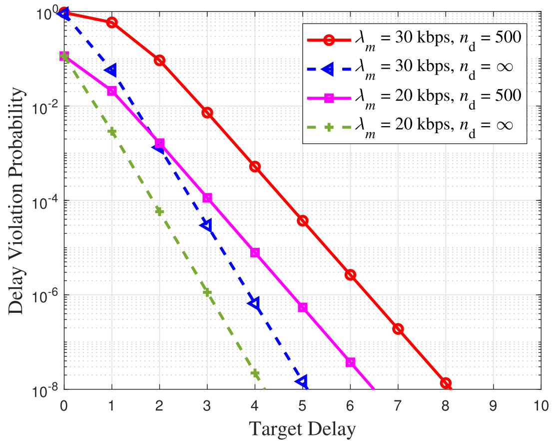


Fig. 5. The delay violation probability $p_m(d_{\text{th}})$ vs. target delay d_{th} for our proposed CF m-MIMO scheme in the finite blocklength regime.

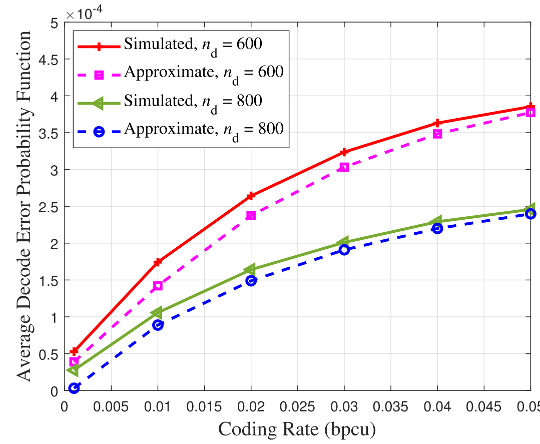


Fig. 7. The average decoding error probability function $\epsilon_m(n_{\text{d}}, \gamma_m)$ vs. achievable coding rate R_m for our proposed CF m-MIMO scheme in the finite blocklength regime.

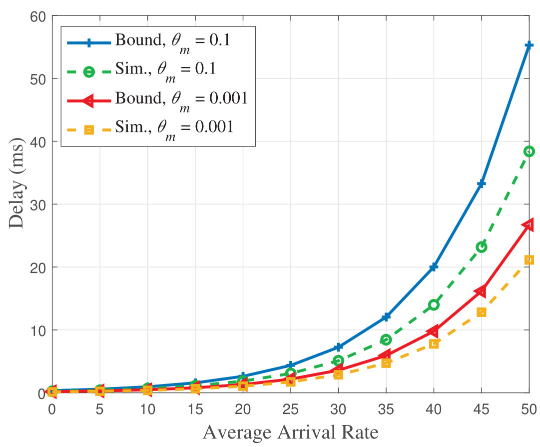


Fig. 6. The delay (ms) vs. average arrival rate λ_m for our proposed CF m-MIMO scheme in the finite blocklength regime.

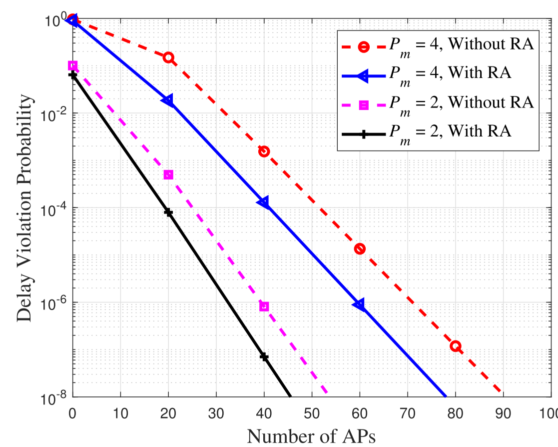


Fig. 8. The delay violation probability $p_m(d_{\text{th}})$ vs. number of APs K_{a} for our proposed CF m-MIMO scheme in the finite blocklength regime.

is because that the queues can be built up more quickly with a larger arrival rate.

We set the target delay $d_{\text{th}} = 5$ ms, Rician factor $\kappa = 10$, the number of downlink channel uses $n_{\text{d}} = 800$, the average downlink transmit power $\overline{\mathcal{P}}_{\text{d}} = 20$ Watt for AP, and decoding error probability $\epsilon_m = 10^{-6}$. Fig. 6 depicts the delay in millisecond (ms) with varying average arrival rates λ_m over Rician wireless fading channels for our proposed CF m-MIMO scheme in the finite blocklength regime, which implies the potential to support massive number of mobile users. We can observe from Fig. 6 that the queuing delay increases as the average arrival rate λ_m increases. Fig. 6 also shows that the analytical results provide a reasonable upper bound for the actual delay as obtained from simulations.

We set the number of APs $K_{\text{a}} = 100$, the number of transmit antennas $N_{\text{T}} = 10$, the multiplexing order $P_m = 2$, Rician factor $\kappa = 0$, and the average downlink transmit power $\overline{\mathcal{P}}_{\text{d}} = 20$ Watt for each AP. Using Eq. (46), Fig. 7 depicts the average decoding error probability function $\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\text{d}}, \gamma_m)]$ with different achievable finite-blocklength coding rates R_m for our proposed CF m-MIMO scheme across Rayleigh wireless fading channels in the finite blocklength regime. We can observe from Fig. 7 that the average decoding error probability function $\mathbb{E}_{\gamma_m}[\epsilon_m(n_{\text{d}}, \gamma_m)]$ increases as the achievable finite-blocklength coding rate R_m increases. Fig. 7 also shows that the gap between the simulated average decoding error function and the approximate average decoding error function is reasonably small.

We set the number of transmit antennas $N_{\text{T}} = 10$, Rician factor $\kappa = 0$, the number of downlink channel uses $n_{\text{d}} = 800$, and the average downlink transmit power $\overline{\mathcal{P}}_{\text{d}} = 20$ Watt. Compared with the scheme without the optimal rate adaptation (RA), Fig. 8 depicts the delay violation probability $p_m(d_{\text{th}})$ with different numbers of APs K_{a} for our proposed CF m-MIMO scheme in the finite blocklength regime. We can observe from Fig. 8 that the delay violation probability $p_m(d_{\text{th}})$ decreases as the numbers of AP K_{a} increases. Fig. 8 also shows that our proposed schemes with optimal RA outperform the schemes without applying the optimal RA in terms of the delay violation probability over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime.

Given different numbers of mobile users K_{u}, Fig. 9 depicts the block error probability function $\epsilon_m(n_{\text{d}}, \gamma_m)$ with varying blocklengths n_{d} for our proposed CF m-MIMO scheme in the finite blocklength regime. We can observe from Fig. 9 that the performance degradation in terms of block error probability function $\epsilon_m(n_{\text{d}}, \gamma_m)$ with the increasing number of mobile users K_{u} is mild, implying the remarkable potential as well as

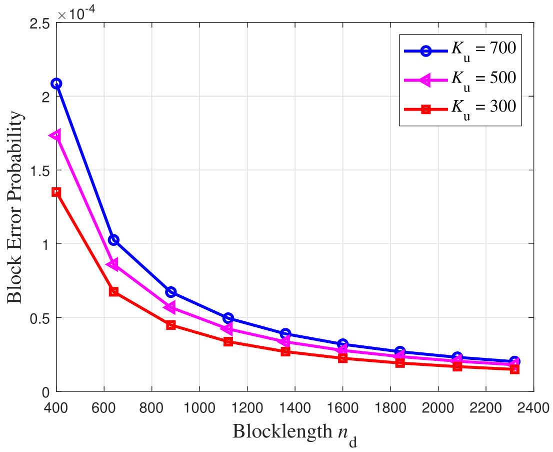


Fig. 9. The block error probability function $\epsilon_m(n_\mathrm{d}, \gamma_m)$ vs. blocklength n_d for our proposed CF m-MIMO scheme in the finite blocklength regime.

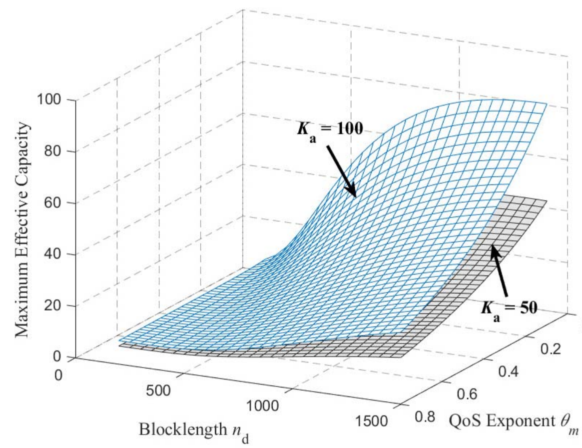


Fig. 11. The maximum effective capacity $EC_m^{\max}(\theta_m)$ vs. blocklength n_d and QoS exponent θ_m for our proposed CF m-MIMO scheme in the finite blocklength regime.

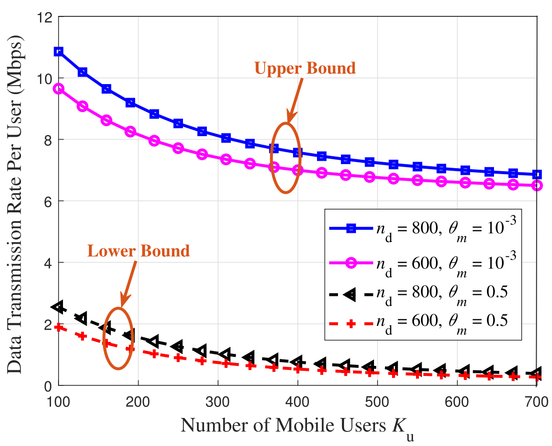


Fig. 10. The data transmission rate per user vs. number of mobile users K_u for our proposed CF m-MIMO scheme in the finite blocklength regime.

the strong and robust scalability in supporting massive access by vast mobile devices over our proposed 6G CF m-MIMO mobile wireless networks.

Setting the number of transmit antennas $N_\mathrm{T} = 2$, Rician factor $\kappa = 0$, and the average downlink transmit power $\overline{\mathcal{P}}_\mathrm{d} = 20$ Watt, Fig. 10 plots the data transmission rate per user with different numbers of mobile users K_u for our proposed CF m-MIMO scheme in the finite blocklength regime. We can observe from Fig. 10 that the data transmission rate per user decreases as the number of mobile users K_u increases and will finally converge to a certain value, which implies the potential to support massive number of mobile users. Fig. 10 also shows that loose QoS constraint $(\theta_m = 10^{-3})$ and stringent QoS constraint $(\theta_m = 0.5)$ set the upper bound and lower bound on the data transmission rate per user, receptively.

Figure 11 plots the maximum effective capacity $EC_m^{\max}(\theta_m)$ with different blocklengths n_d and QoS exponents θ_m for our proposed CF m-MIMO scheme in the finite blocklength regime. We can observe from Fig. 11 that the maximum effective capacity $EC_m^{\max}(\theta_m)$ decreases as the QoS exponent θ_m increases. Fig. 11 also shows that the maximum effective capacity $EC_m^{\max}(\theta_m)$ is an increasing function of the blocklength n_d over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime.

VI. Conclusion

We have developed an analytical model to quantitatively characterize the performance for statistical delay and error-rate bounded QoS provisioning in supporting mURLLC over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime. In particular, we have developed CF m-MIMO based system models. Then, we have applied the Mellin transform to model and characterize both arrival and service processes, derived closed-form expressions for the delay violation probability function, and formulated and solved the delay violation probability minimization problem for our proposed CF m-MIMO modeling schemes in the finite blocklength regime. Furthermore, applying our developed system modeling techniques, we have derived a closed-form solution for the optimal rate adaptation policy, which plays an important role in system design and performance analyses for statistical delay and error-rate bounded QoS provisioning over 6G CF m-MIMO mobile wireless networks in the finite blocklength regime. We also have conducted a set of simulations to validate and evaluate our proposed CF m-MIMO schemes and show that our proposed schemes outperform the other existing schemes for statistical delay and error-rate bounded QoS provisioning in the finite blocklength regime.

References

[1] 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.

[2] J. Tang and X. Zhang, "Cross-layer resource allocation over wireless relay networks for quality of service provisioning," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 645–656, May 2007.

[3] Y. Liu and Y. Jiang, *Stochastic Network Calculus*. London, U.K.: Springer, 2008.

[4] J. Wang and X. Zhang, "Heterogeneous QoS-driven resource adaptation over full-duplex relay networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[5] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," 2020, *arXiv:2002.03491*. [Online]. Available: http://arxiv.org/abs/2002.03491

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

[7] 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.

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

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

[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] 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.

[12] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 651–663, Sep. 2019.

[13] 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.

[14] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[15] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. ICC-IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[16] M. Kang and M. S. Alouini, "Capacity of MIMO Rician channels," *IEEE Trans. Wireless Commun.*, vol. 5, no. 1, pp. 112–122, Jan. 2006.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall PTR, 1993.

[18] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[19] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sep. 2004.

[20] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[21] B. Davies, *Integral Transforms and Their Applications*. New York, NY, USA: Springer-Verlag, 1978.

[22] 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.

[23] 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.

[24] S. K. Jayaweera and H. V. Poor, "On the capacity of multiple-antenna systems in Rician fading," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1102–1111, May 2005.

[25] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, 2015, pp. 13–22.

[26] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 8th ed. New York, NY, USA: Academic, 2007.

[27] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[29] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.

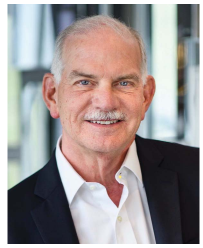
[30] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous statistical QoS provisioning over airborne mobile wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2139–2152, Sep./Oct. 2018.

Xi Zhang (Fellow, IEEE) received the B.S. and M.S. degrees 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 Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He was with the Department of Networks and Distributed Systems Research, AT&T Bell Laboratories, Murray Hill, NJ, USA, and AT&T Laboratories Research, Florham Park, NJ, USA, in 1997. He was a research fellow of 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 400 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 is a fellow of the IEEE for contributions to the Quality of Service (QoS) theory in mobile wireless networks. He is an IEEE Distinguished Lecturer for both the IEEE Communications Society and the 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, College Station, in 2006 and the Outstanding Faculty Award from Department of Electrical and Computer Engineering , Texas A&M University, in 2020. 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 six Best Paper Awards at IEEE GLOBECOME 2020, IEEE ICC 2018, IEEE GLOBECOM 2014, IEEE GLOBECOM 2009, IEEE GLOBECOM 2007, and IEEE WCNC 2010, respectively. One of his articles on the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS has been listed as the IEEE Best Readings Paper (receiving the highest citation rate among all the IEEE TRANSACTIONS/journal articles in the area) on Wireless Cognitive Radio Networks and Statistical QoS Provisioning over Mobile Wireless Networking.

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

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 Prof. 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 from the IEEE GLOBECOM in 2014, the Hagler Institute for Advanced Study Heep Graduate Fellowship Award from Texas A&M University in 2018, and Dr. R.K. Pandey and Christa U. Pandey'84 Fellowship, Texas A&M University, USA, 2020–2021.



H. Vincent Poor (Life 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. Among his publications in these areas is the recent book, *Multiple Access Techniques for 5G Wireless Networks and Beyond* (Springer, 2019). From 2006 to 2016, he served as the Dean of the School of Engineering and Applied Science, Princeton University. 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, machine learning and their applications in wireless networks, and energy systems and their 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 is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. His recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal.