Heterogeneous Statistical-QoS Driven Resource Allocation Over mmWave Massive-MIMO Based 5G Mobile Wireless Networks in the Non-Asymptotic Regime

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

***Abstract*— The statistical delay-bounded quality-of-service (QoS) theory has been developed to efficiently support multimedia transmissions over 5G wireless networks. On the other hand, unlike in Shannon's information-theoretic formalism requiring infinite blocklength, finite blocklength coding (FBC) has recently emerged for error control in the non-asymptotic regime, guaranteeing stringent statistical QoS requirements in terms of both *latency* and *reliability* for ultra-reliable low-latency communications (URLLC) in 5G services. Moreover, integrated with FBC, millimeter wave (mmWave) massive multi-input multi-output (m-MIMO) schemes have been designed to significantly improve the performance in guaranteeing delay/error-rate bounded QoS. However, due to the complexity of modeling and solving the optimization problems over mmWave m-MIMO fading channels in the non-asymptotic error-control regime, it is challenging to derive an optimal resource allocation policy for maximizing the ϵ-effective capacity to guarantee statistical delay/error-rate bounded QoS. To overcome the above problems, in this paper we propose heterogeneous statistical-QoS driven resource allocation policies for mmWave m-MIMO based 5G wireless networks in both asymptotic and non-asymptotic regimes. In particular, we develop an mmWave m-MIMO based 5G wireless networks model to optimize the effective capacity for our proposed schemes. Our simulations show that our proposed schemes outperform the existing schemes in guaranteeing heterogeneous statistical delay/error-rate bounded QoS.**

***Index Terms*— Heterogeneous QoS, mmWave m-MIMO, 5G, FBC, ϵ-effective capacity, D2D, non-asymptotic regime.**

I. Introduction

DELAY-BOUNDED quality-of-service (QoS) guarantees have played a critically important role for supporting the explosive growth in wireless multimedia services over 5G mobile wireless networks. One of the key design issues for multimedia wireless services is how to efficiently guarantee time-sensitive multimedia data transmissions within specified delay bounds. Due to the highly-varying nature of wireless channels, deterministic delay-bounded QoS requirements are usually hard to guarantee. As a result, statistical delay-bounded QoS guarantees [1], [2], in terms of effective capacity and queue-length-bound/delay-bound violation probabilities, have been proposed and proved to be a powerful way of characterizing delay-bounded QoS requirements over wireless fading channels. Since the time sensitivities of various types of wireless services vary from 1 ms to a few seconds across different wireless links, delay-bounded QoS guarantees for diverse types of services over 5G mobile wireless networks demand new heterogeneous statistical delay-bounded QoS provisioning architectures [3], [4], schemes, algorithms, and 5G candidate frameworks, which can be implemented by integrating several 5G candidate techniques. Various advanced 5G promising techniques, such as millimeter wave (mmWave) and massive multiple-input multiple-output (m-MIMO), have been designed to play a critically important role in 5G mobile wireless networks. Accordingly, heterogeneous statistical delay-bounded QoS provisioning over mmWave m-MIMO-based 5G mobile wireless networks still remains as a challenging and open problem due to the complexity of the system design.

Towards this end, there has been a considerable amount of research investigating the integration of various advanced techniques over 5G mobile wireless networks. The authors in [5] have introduced an m-MIMO system with a large number of antennas as an emerging technology that can deliver all the attractive benefits compared with the traditional MIMO system, but at a much larger scale. Such an m-MIMO system can substantially reduce the impacts of noise, fast fading, and interference, and also provide increased system capacity. An m-MIMO full-duplex (FD) relay architecture is considered and a closed-form expression for the achievable rate is derived in [6]. Approximations to achievable rates with several linear precoders and detectors over m-MIMO based wireless fading channels have been developed in [7]. The authors of [8] have investigated a low-complexity hybrid block diagonalization scheme to derive the channel capacity for downlink multiuser MIMO schemes.

Manuscript received March 15, 2019; revised August 10, 2019; accepted August 20, 2019. Date of publication October 23, 2019; date of current version November 27, 2019. This work of X. Zhang and J. Wang was supported in part by the U.S. National Science Foundation under 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. V. Poor was supported in part by the U.S. National Science Foundation under Grant CCF-0939370 and Grant CCF-1513915. *(Corresponding author: Xi Zhang.)*

X. Zhang and J. Wang are with the 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. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

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

Digital Object Identifier 10.1109/JSAC.2019.2947941

In addition, researchers have investigated the integration of mmWave techniques with m-MIMO systems [9] due to its potential to improve the spectral efficiency (SE) while mitigating the self-interference introduced by m-MIMO systems. How to characterize and implement hybrid beamforming models over multi-path wireless fading channels is one of the major design issues for mmWave m-MIMO systems. Accordingly, the authors of [10] investigate how beamforming and precoding are different in mmWave MIMO systems from their lower-frequency counterparts, due to different hardware constraints and channel characteristics. The authors of [11] have proposed a new approach to channel estimation based on training sequences for mmWave m-MIMO systems. However, high implementation costs and energy consumption due to hardware constraints in mmWave m-MIMO systems make it difficult to apply conventional low-complexity MIMO precoding and beamforming techniques. As a result, researchers have been studying multi-user mmWave m-MIMO systems, where the digital precoding layer of hybrid precoding yields more freedom in designing the precoders, which can be exploited to reduce the interference among mobile users.

The authors of [13] have shown that, in order to guarantee the constraints of reliable data transmission for some applications, the codeword blocklength is required to be short (on the order of 100 channel symbols). Accordingly, since Shannon's capacity is based on infinite blocklength and perfect reliability [12], [14], it is not appropriate to characterize the maximum achievable data transmission rate under both constraints of transmission latency and finite data-packet length.

To tackle challenges introduced by finite-length data transmissions, researchers in [15] have proposed methods to characterize the maximum achievable coding rate using finite blocklength coding (FBC) over additive white Gaussian noise (AWGN) channels. Further, the authors of [16] have investigated a framework for cross-layer optimization to support ultra-reliable and low latency communications (URLLC) in radio access networks. The authors in [17] have analyzed the throughput of cognitive radio systems in the non-asymptotic regime under buffer constraints. Approximations of data transmission rates given the error probability and channel capacity in the non-asymptotic regime have been developed in [18]. The authors of [19] have provided new compact integral expressions and associated simple asymptotic approximations for converse and achievability bounds in the non-asymptotic regime. However, due to the design complexity in formulating and solving optimization for mmWave m-MIMO systems with FBC, there are still many new challenges for the statistical QoS theory in the non-asymptotic regime in this setting.

To effectively overcome the aforementioned problems, in this paper we propose heterogeneous statistical QoS driven resource allocation schemes by optimizing multiuser mmWave m-MIMO based 5G mobile wireless networks in both the asymptotic and non-asymptotic regimes. In particular, we develop mmWave m-MIMO based 5G mobile networking system architectures and then establish the corresponding mmWave m-MIMO based 5G mobile wireless networking system models. Further, we propose and analyze a hybrid block diagonalization model over the m-MIMO scheme with imperfect knowledge of the channel state information (CSI). Given the heterogeneous statistical delay-bounded QoS constraints, we design a cross-layer optimization model by developing optimal resource-allocation policies over relay-D2D links for our proposed mmWave m-MIMO scheme in the asymptotic regime. In addition, we also formulate and solve the ϵ-effective capacity maximization problem in the non-asymptotic regime. Also conducted is a set of simulations to validate and evaluate our proposed schemes and show that our proposed schemes outperform the other existing schemes to efficiently implement heterogeneous statistical delay-bounded QoS provisioning over multiuser mmWave m-MIMO based 5G mobile wireless networks.

The rest of this paper is organized as follows. Section II builds the system models for our proposed mmWave m-MIMO based 5G mobile wireless networks. Section III characterizes and analyzes the cross-layer design for maximizing effective capacity over relay-D2D link under heterogeneous statistical delay-bounded QoS constraints in the asymptotic regime. Section IV formulates and solves the ϵ-effective capacity maximization problem in the non-asymptotic regime. Section V evaluates the system performance and shows that our proposed resource allocation policies outperform the other existing schemes in terms of the effective capacity under heterogeneous statistical delay-bounded QoS constraints. The paper concludes with Section VI.

II. The Networking Architectures and System Models for Our Proposed MmWave M-MIMO Based 5G Mobile Wireless Networks With FBC

Fig. 1 depicts the system architecture model for our proposed mmWave m-MIMO based 5G mobile wireless networks with a large number of antennas equipped at the BS in the non-asymptotic regime, where there are $\mathcal{K} = \{1, 2, \ldots, K\}$ randomly distributed mobile devices simultaneously served by one mmWave m-MIMO base station (BS) in each wireless cell. We assume that both the mmWave m-MIMO BS and mobile users are equipped with large antenna arrays. Assume each mobile user has N_{R} receiving antennas and L_{R} radio frequency (RF) chains, while the mmWave m-MIMO BS consists of N_{T} transmit antennas and L_{T} RF chains. For our proposed mmWave m-MIMO scheme in Fig. 1, the mmWave m-MIMO BS applies a low-complex hybrid analog and digital beamforming structure, where the input N_{s} multimedia data streams first pass through the low-dimensional digital baseband precoder, denoted by $\mathbf{F}^{\mathrm{B}}$, which essentially plays the role of power allocation. Then, after the digital precoder, the signals are transmitted through L_{T} RF chains. The transmitter at the BS processes the signals through phase shifters (PSs) for analog precoding, where the analog precoding matrix is denoted by $\mathbf{F}^{\mathrm{R}}$. After the analog precoding, each multimedia data stream is finally transmitted by a sub-antenna array associated with the corresponding RF chain. Due to the power and hardware constraints for large scale MIMO system, the number of RF chains is much smaller than the number of transmit antennas at the mmWave m-MIMO BS to

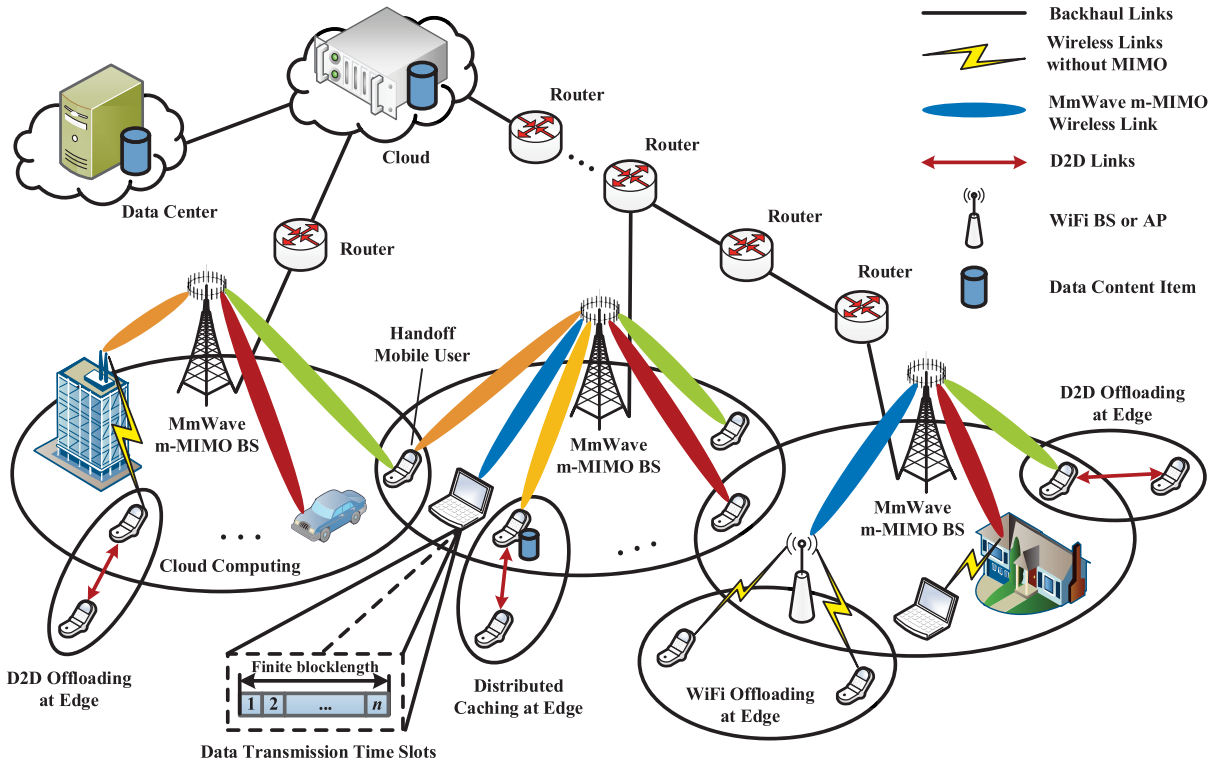


Fig. 1. The system architecture model for our proposed mmWave m-MIMO (m-MIMO) based 5G mobile wireless networks with large antenna arrays in the non-asymptotic regime.

guarantee the effectiveness of the wireless communications, i.e., $KN_{\rm s} \le L_{\rm T} \le N_{\rm T}$ and $N_{\rm s} \le L_{\rm R} \le N_{\rm R}$ [10]. Note that the multimedia data streams are transmitted using FBC, i.e., each data stream is encoded into a codeword with finite length n considering the non-asymptotic regime. The detailed analysis of FBC is in Section IV.

In order to implement the functions of our proposed mmWave m-MIMO based 5G mobile wireless networks, we also develop the networking task control, communication-mode selections, and resource allocation protocols which consist of the following items.

1) Handoff: Since the handoff mechanisms have a significant impact on QoS provisioning as well as the system capacity, continuous and smooth handoffs are necessary to achieve seamless data transmissions.
2) Cloud computing via mmWave m-MIMO: Inspired by the idea of cloud computing, cloud-RAN (C-RAN) transfers, many high-complexity computations are transferred and executed in the cloud through backhaul links.
3) D2D offloading via mmWave m-MIMO: To reduce congestion and support statistical delay-bounded QoS requirements, we apply D2D offloading technique at the edge devices, which reduces the backhaul burden by rerouting cellular traffics from a remote server via cellular mmWave m-MIMO BSs.
4) WiFi offloading via mmWave M-MIMO (based wireless backhaul): Similar to D2D offloading, WiFi offloading strategies can be applied at the edge of the mobile wireless networks to improve the system capacity for mobile devices over the resource-starving 5G mobile wireless networks.
5) Distributed caching via mmWave m-MIMO: To solve the network congestion problem while guaranteeing QoS constraints, researchers have developed distributed caching schemes where each mobile user is equipped with a data storage which caches popular multimedia files from the remote backhaul networks and directly provides the neighboring mobile users with immediately accessible multimedia contents bypassing the mmWave m-MIMO BSs.

A. The mmWave M-MIMO Channel Model With Perfect CSI

We can derive the channel's impulse response matrix, denoted by $\mathbf{H}_k$, over the mmWave m-MIMO based wireless fading channels from the BS to user k as in the following equation [20]:

$$\mathbf{H}_k = \sqrt{\frac{N_{\rm T}N_{\rm R}}{Q\rho}} \sum_{q=1}^{Q} \alpha_k^{(q)} \boldsymbol{a}\left(\phi_k^{(q)}\right)\left(\boldsymbol{b}\left(\psi_k^{(q)}\right)\right)^{\dagger} = \sqrt{\frac{N_{\rm T}N_{\rm R}}{Q\rho}} \mathbf{A}_k \mathrm{diag}\left(\boldsymbol{\alpha}\right)\left(\mathbf{B}_k\right)^{\dagger} \quad (1)$$

where $(\cdot)^{\dagger}$ is the Hermitian transpose of a matrix; $N_{\rm T}$ and $N_{\rm R}$ denote the numbers of transmit antennas at the BS and receive antennas at mobile users, respectively; ρ is the average path loss; Q represents the number of channel paths; $\alpha_k^{(q)}$ is the complex gain of the qth path; $\phi_k^{(q)}$ and $\psi_k^{(q)}$ are the azimuth angles of departure or arrival (AoD/AoA) which are

uniformly distributed between $[0, 2\pi]$ [21], respectively; and $\boldsymbol{a}\left(\phi_k^{(q)}\right) \in \mathbb{C}^{N_{\mathrm{R}}\times 1}$ and $\boldsymbol{b}\left(\psi_k^{(q)}\right) \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$ represent the antenna array's response vectors for the transmitter and the receiver, respectively, which are given by:

$$\begin{cases} \boldsymbol{a}\left(\phi_k^{(q)}\right) = \frac{1}{\sqrt{N_{\mathrm{R}}}} \\ \qquad \times\left[1, e^{\jmath 2\pi d \sin\left(\phi_k^{(q)}\right)\lambda}, \ldots, e^{\jmath 2\pi (N_{\mathrm{R}}-1) d \sin\left(\phi_k^{(q)}\right)\lambda}\right]^T \\ \boldsymbol{b}\left(\psi_k^{(q)}\right) = \frac{1}{\sqrt{N_{\mathrm{R}}}} \\ \qquad \times\left[1, e^{\jmath 2\pi d \sin\left(\psi_k^{(q)}\right)\lambda}, \ldots, e^{\jmath 2\pi (N_{\mathrm{T}}-1) d \sin\left(\psi_k^{(q)}\right)\lambda}\right]^T \end{cases} \tag{2}$$

where $\jmath = \sqrt{-1}$, $(\cdot)^T$ represents the transpose of a matrix, λ represents the wavelength of transmit signal, d is the distance between the antenna array elements, and the matrices $\boldsymbol{\alpha}$, $\mathbf{A}_k$, and $\mathbf{B}_k$ in Eq. (1) are defined by the following equations [22]:

$$\begin{cases} \boldsymbol{\alpha} \;\; = \sqrt{\frac{N_{\mathrm{T}} N_{\mathrm{R}}}{\rho}} \left[\alpha_1, \alpha_2, \ldots, \alpha_Q\right]^T; \\ \mathbf{A}_k = \left[b\left(\phi_1\right), b\left(\phi_2\right), \ldots, b\left(\phi_Q\right)\right]; \\ \mathbf{B}_k = \left[\boldsymbol{b}\left(\psi_1\right), \boldsymbol{b}\left(\psi_2\right), \ldots, \boldsymbol{b}\left(\psi_Q\right)\right]. \end{cases} \tag{3}$$

As a result, the total channel's impulse response, denoted by $\mathbf{H}$, for all K mobile users can be expressed as follows:

$$\mathbf{H} = \left[\left(\mathbf{H}_1\right)^T, \left(\mathbf{H}_2\right)^T, \ldots, \left(\mathbf{H}_K\right)^T\right]^T. \tag{4}$$

Consider a user-centric D2D offloading policy [23] to control the D2D communications: two D2D users can communicate with each other only if their distance is smaller than the collaboration distance threshold d_{th}. We assume that all K mobile users can choose between two modes: *cellular mode* and *D2D mode*. We introduce and define the *binary D2D mode selection indicator variable*, denoted by b_k, for user k subject to the following constraints:

$$\begin{cases} b_k = 1, & \text{if user } k \text{ chooses D2D offloading;} \\ b_k = 0, & \text{otherwise.} \end{cases} \tag{5}$$

The edge mobile user can be considered as a *relay node* for nearby D2D users. Correspondingly, the received signals, denoted by $\boldsymbol{y}_{\mathrm{r},k} \in \mathbb{C}^{N_{\mathrm{R}}\times 1}$ and $\boldsymbol{y}_{\mathrm{d},k} \in \mathbb{C}^{N_{\mathrm{R}}\times 1}$, at relay node k and the corresponding D2D receiver node can be derived as in the following equation:

$$\begin{cases} \boldsymbol{y}_{\mathrm{r},k} = \sqrt{\mathcal{P}_{\mathrm{r},k}}\mathbf{H}_{\mathrm{r},k}\mathbf{F}_{\mathrm{r},k}\boldsymbol{s}_{\mathrm{r},k} + \sum_{j\neq k}^{K}(1-b_j)\sqrt{\mathcal{P}_{\mathrm{r},j}}\mathbf{H}_{\mathrm{r},k}\mathbf{F}_{\mathrm{r},j}\boldsymbol{s}_{\mathrm{r},j} \\ \qquad\qquad + \boldsymbol{n}_{\mathrm{r},k}; \\ \boldsymbol{y}_{\mathrm{d},k} = \sqrt{\mathcal{P}_{\mathrm{d},k}}\mathbf{H}_{\mathrm{d},k}\boldsymbol{s}_{\mathrm{d},k} + \sum_{j\neq k}^{K} b_j\sqrt{\mathcal{P}_{\mathrm{d},j}}\mathbf{H}_{\mathrm{d},j}\boldsymbol{s}_{\mathrm{d},j} + \boldsymbol{n}_{\mathrm{d},k}, \end{cases} \tag{6}$$

where b_j is the binary D2D mode selection indicator variable for user j given by Eq. (5); $\mathcal{P}_{\mathrm{r},k}$ and $\mathcal{P}_{\mathrm{r},j}$ represent the transmit powers from the mmWave m-MIMO BS to edge user k and edge user j, respectively; $\mathcal{P}_{\mathrm{d},k}$ and $\mathcal{P}_{\mathrm{d},j}$ represent the transmit powers from relay node k to the corresponding D2D receiver and from relay node j to the corresponding D2D receiver, respectively; $\mathbf{H}_{\mathrm{r},k}$, $\mathbf{H}_{\mathrm{d},k}$, and $\mathbf{H}_{\mathrm{d},j}$ denote the channel's impulse response matrices for the kth relay link, kth D2D link, and jth D2D link, respectively; $\boldsymbol{n}_{\mathrm{r},k} \sim \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{n}_{\mathrm{d},k} \sim \mathcal{N}(0, \sigma^2)$ denote the AWGN vectors for the kth relay link and kth D2D link, respectively; $\boldsymbol{s}_{\mathrm{r},k}$ and $\boldsymbol{s}_{\mathrm{r},j}$ represent the vectors of the transmit signals sent from the mmWave m-MIMO BS to user to user k and user j, respectively; $\boldsymbol{s}_{\mathrm{d},k}$ is the transmit signal vector from relay node k to the corresponding D2D receiver; and $\mathbf{F}_{\mathrm{r},k}$ and $\mathbf{F}_{\mathrm{r},j}$ are the hybrid precoder matrices at the mmWave m-MIMO BS for user k and user j, respectively. We define the hybrid $\mathbf{F}_{\mathrm{r},k} = \mathbf{F}_{\mathrm{r},k}^{\mathrm{R}}\mathbf{F}_{\mathrm{r},k}^{\mathrm{B}}$ at the mmWave m-MIMO BS for user k, where $\mathbf{F}_{\mathrm{r},k}^{\mathrm{R}} \in \mathbb{C}^{N_{\mathrm{T}}\times L_{\mathrm{T}}}$ is the analog RF precoder and $\mathbf{F}_{\mathrm{r},k}^{\mathrm{B}} \in \mathbb{C}^{L_{\mathrm{T}}\times N_{\mathrm{s}}}$ is the digital baseband precoder. Then, for all K mobile users, we define $\mathbf{F}_{\mathrm{r}} = \left[\mathbf{F}_{\mathrm{r},1}, \mathbf{F}_{\mathrm{r},2}, \ldots, \mathbf{F}_{\mathrm{r},K}\right]$, $\mathbf{F}_{\mathrm{r}}^{\mathrm{R}} = \left[\mathbf{F}_{\mathrm{r},1}^{\mathrm{R}}, \mathbf{F}_{\mathrm{r},2}^{\mathrm{R}}, \ldots, \mathbf{F}_{\mathrm{r},K}^{\mathrm{R}}\right]$, and $\mathbf{F}_{\mathrm{r}}^{\mathrm{B}} = \left[\mathbf{F}_{\mathrm{r},1}^{\mathrm{B}}, \mathbf{F}_{\mathrm{r},2}^{\mathrm{B}}, \ldots, \mathbf{F}_{\mathrm{r},K}^{\mathrm{B}}\right]$. In addition, note that the transmit signal vector need to satisfy $\mathbb{E}\left[\boldsymbol{s}_{\mathrm{r},k}\left(\boldsymbol{s}_{\mathrm{r},k}\right)^{\dagger}\right] = \frac{1}{KN_{\mathrm{s}}}$. Define the transmit signal vector from the mmWave m-MIMO BS to all K users as $\boldsymbol{s}_{\mathrm{r}} = \left[\boldsymbol{s}_{\mathrm{r},1}, \boldsymbol{s}_{\mathrm{r},2}, \ldots, \boldsymbol{s}_{\mathrm{r},K}\right]^T$. Then, we analyze the processed received signal, denoted by $\boldsymbol{z}_{\mathrm{r},k}$, after analog combining at user k can be derived as follows:

$$\begin{aligned} \boldsymbol{z}_{\mathrm{r},k} &= \left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\boldsymbol{y}_{\mathrm{r},k} \\ &= \sqrt{\mathcal{P}_{\mathrm{r},k}}\left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\mathbf{H}_{\mathrm{r},k}\mathbf{F}_{\mathrm{r},k}\boldsymbol{s}_{\mathrm{r},k} + \left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\mathbf{H}_{\mathrm{r},k}\sum_{j\neq k}^{K}(1-b_j) \\ &\quad \times\sqrt{\mathcal{P}_{\mathrm{r},j}}\mathbf{F}_{\mathrm{r},j}\boldsymbol{s}_{\mathrm{r},j} + \left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\boldsymbol{n}_{\mathrm{r},k} \end{aligned} \tag{7}$$

where $\mathbf{W}_{\mathrm{r},k}$ represents the analog RF combiner at the receiver of edge user k for extracting the transmitted data from the received signal. Note that all elements of matrix $\mathbf{W}_{\mathrm{r},k}$ should satisfy the magnitude constraint such that $\left\|W_{\mathrm{r},k}^{(i,j)}\right\| = 1/\sqrt{N_{\mathrm{R}}}$, where $\left\|W_{\mathrm{r},k}^{(i,j)}\right\|$ represents the magnitude of the (i,j)th element of matrix $\mathbf{W}_{\mathrm{r},k}$. Accordingly, we can define an equivalent baseband channel's impulse response matrix, denoted by $\widetilde{\mathbf{H}}_{\mathrm{r},k}$, for edge user k as in the following equation:

$$\widetilde{\mathbf{H}}_{\mathrm{r},k} = \left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\mathbf{H}_{\mathrm{r},k}\mathbf{F}_{\mathrm{r},k}^{\mathrm{R}}. \tag{8}$$

Accordingly, we can derive the total multiuser baseband equivalent channel's impulse response matrix, denoted by $\widetilde{\mathbf{H}}_{\mathrm{r}}$, as follows:

$$\begin{aligned} \widetilde{\mathbf{H}}_{\mathrm{r}} &= \left[\widetilde{\mathbf{H}}_{\mathrm{r},1}\widetilde{\mathbf{H}}_{\mathrm{r},2}, \ldots, \widetilde{\mathbf{H}}_{\mathrm{r},K}\right]^T \\ &= \begin{bmatrix} \left(\mathbf{W}_{\mathrm{r},1}\right)^{\dagger} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{W}_{\mathrm{r},2}\right)^{\dagger} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \left(\mathbf{W}_{\mathrm{r},K}\right)^{\dagger} \end{bmatrix} \mathbf{H}_{\mathrm{r}}\mathbf{F}_{\mathrm{r}}^{\mathrm{R}}. \end{aligned} \tag{9}$$

As a result, the processed received signal can be rewritten as in the following equation:

$$\begin{aligned} \boldsymbol{z}_{\mathrm{r},k} &= \sqrt{\mathcal{P}_{\mathrm{r},k}}\widetilde{\mathbf{H}}_{\mathrm{r},k}\mathbf{F}_{\mathrm{r},k}^{\mathrm{B}}\boldsymbol{s}_{\mathrm{r},k} + \widetilde{\mathbf{H}}_{\mathrm{r},k}\sum_{j\neq k}^{K}(1-b_j)\sqrt{\mathcal{P}_{\mathrm{r},j}}\mathbf{F}_{\mathrm{r},j}^{\mathrm{B}}\boldsymbol{s}_{\mathrm{r},j} \\ &\quad + \left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\boldsymbol{n}_{\mathrm{r},k}. \end{aligned} \tag{10}$$

We can then determine the signal-to-noise radio (SNR), denoted by $\gamma_{\mathrm{r},k}$, from the mmWave m-MIMO BS to edge

user k as follows:

$$\gamma_{\mathrm{r},k}=\frac{\frac{\mathcal{P}_{\mathrm{r},k}}{KN_{\mathrm{s}}}\left\|\widetilde{\mathbf{H}}_{\mathrm{r},k}\mathbf{F}^{\mathrm{B}}_{\mathrm{r},k}\right\|^2_F}{\sum\limits_{j\neq k}^{K}(1-b_j)\frac{\mathcal{P}_{\mathrm{r},j}}{KN_{\mathrm{s}}}\left\|\widetilde{\mathbf{H}}_{\mathrm{r},k}\mathbf{F}^{\mathrm{B}}_{\mathrm{r},j}\right\|^2_F+\sigma^2\left\|\left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\right\|^2_F} \tag{11}$$

where $\|\mathbf{M}\|_F^2$ is the Frobenius norm of matrix $\mathbf{M}$ which is defined as $\sqrt{\mathrm{Tr}\left(\mathbf{M}^{\dagger}\mathbf{M}\right)}$ where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

B. The mmWave M-MIMO Channel Model With Imperfect CSI

In practical scenarios, it is not realistic to assume the perfect CSI at the mmWave m-MIMO BSs. Accordingly, with the imperfect CSI, we can derive the minimum mean-square error (MMSE) channel estimation for the baseband equivalent channel's impulse response matrix $\widetilde{\mathbf{H}}_{\mathrm{r}}$, denoted by $\widehat{\boldsymbol{H}}_{\mathrm{r}}$, as in the following equation:

$$\widehat{\boldsymbol{H}}_{\mathrm{r}}=\widetilde{\mathbf{H}}_{\mathrm{r}}-\widehat{\boldsymbol{\epsilon}}_{\mathrm{r}} \tag{12}$$

where $\widehat{\boldsymbol{\epsilon}}_{\mathrm{r}}$ represents the estimation error matrix for channel's impulse response matrix $\widetilde{\mathbf{H}}_{\mathrm{r}}$. We define $\widehat{\boldsymbol{H}}_{\mathrm{r}}\triangleq\left[\widehat{\boldsymbol{H}}_{\mathrm{r},1},\ldots,\widehat{\boldsymbol{H}}_{\mathrm{r},K}\right]$ and $\widehat{\boldsymbol{\epsilon}}_{\mathrm{r}}\triangleq[\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},1},\ldots,\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},K}]$. Correspondingly, we can rewrite the processed received signal with imperfect CSI as follows:

$$\begin{aligned}\widehat{\boldsymbol{z}}_{\mathrm{r},k}=&\sqrt{\mathcal{P}_{\mathrm{r},k}}\widehat{\mathbf{H}}_{\mathrm{r},k}\mathbf{F}^{\mathrm{B}}_{\mathrm{r},k}\boldsymbol{s}_{\mathrm{r},k}+\sqrt{\mathcal{P}_{\mathrm{r},k}}\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k}\mathbf{F}^{\mathrm{B}}_{\mathrm{r},k}\boldsymbol{s}_{\mathrm{r},k}+\sum_{j\neq k}^{K}(1-b_j)\\&\times\sqrt{\mathcal{P}_{\mathrm{r},j}}\widehat{\mathbf{H}}_{\mathrm{r},k}\mathbf{F}^{\mathrm{B}}_{\mathrm{r},j}\boldsymbol{s}_{\mathrm{r},j}+\sum_{j\neq k}^{K}(1-b_j)\sqrt{\mathcal{P}_{\mathrm{r},j}}\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k}\mathbf{F}^{\mathrm{B}}_{\mathrm{r},j}\boldsymbol{s}_{\mathrm{r},j}\\&+\left(\mathbf{W}_{\mathrm{r},k}\right)^{\dagger}\boldsymbol{n}_{\mathrm{r},k}\end{aligned} \tag{13}$$

where $\widehat{\mathbf{H}}_{\mathrm{r},k}$ denotes the estimation of baseband equivalent channel's impulse response matrix $\widetilde{\mathbf{H}}_{\mathrm{r},k}$.

C. The Hybrid Block Diagonalization Based mmWave M-MIMO System Model

Traditionally, linear processing techniques, such as zero-forcing (ZF) processing technique, have been proven to be able to maximize the achievable data transmission rate or channel capacity for the multiuser MIMO systems. However, due to the significant feedback overhead and high cost of the large number of RF chains/antennas required by ZF precoding, the generalized ZF technique cannot be practically implemented for our proposed mmWave m-MIMO schemes. By reducing the number of RF chains ($L_{\mathrm{T}}\ll N_{\mathrm{T}}$) at both the mmWave m-MIMO BS and the mobile users, a low-complexity processing scheme can be performed at the baseband.

Accordingly, in this section, instead of finding the global optimal value for the joint optimization on the RF and baseband precoders and combiners, researchers have proposed the hybrid processing structure [21] with separated RF and baseband processing designs. The newly developed hybrid processing structure can avoid vast amounts of iterative procedures introduced by the traditional multiuser m-MIMO system. Correspondingly, we investigate and compare two different cases with and without the implementation of hybrid processing structure as follows.

Case 1: *(Single-Path Wireless Channel)* For the mmWave m-MIMO scheme over the single-path wireless fading channel, i.e., $Q=1$, we assume that each mobile user only schedules one data stream through one RF chain, i.e., $N_{\mathrm{s}}=L_{\mathrm{R}}=1$, although the mmWave m-MIMO BS is equipped with $L_{\mathrm{T}}=K$ RF chains. The path gain of channel's impulse response $\mathbf{H}_{\mathrm{r},k}$ and the azimuth angles of AoD and AoA are represented as $\alpha_{\mathrm{r},k}$, $\phi_{\mathrm{r},k}$, and $\psi_{\mathrm{r},k}$, respectively, for simplicity in the single-path wireless channel case. As a result, we can rewrite the single-path channel's impulse response between the mmWave m-MIMO BS and edge user k as in the following equation:

$$\mathbf{H}_{\mathrm{r},k}=\sqrt{N_{\mathrm{T}}N_{\mathrm{R}}}\alpha_{\mathrm{r},k}\boldsymbol{a}\left(\phi_{\mathrm{r},k}\right)\left(\boldsymbol{b}\left(\psi_{\mathrm{r},k}\right)\right)^{\dagger}. \tag{14}$$

Since we assume that there is only one path over the wireless fading channel, the optimal combining and RF precoding vectors can be directly set as $\mathbf{W}^{\mathrm{opt}}_{\mathrm{r},k}=\boldsymbol{a}\left(\phi_{\mathrm{r},k}\right)$ and $\mathbf{F}^{\mathrm{R,opt}}_{\mathrm{r},k}=\boldsymbol{b}\left(\psi_{\mathrm{r},k}\right)$, respectively. Correspondingly, the precoding vector at the mmWave m-MIMO BS for all K mobile users is given as $\mathbf{F}^{\mathrm{R}}_{\mathrm{r}}=[\boldsymbol{b}\left(\psi_{\mathrm{r},1}\right),\boldsymbol{b}\left(\psi_{\mathrm{r},2}\right),\ldots,\boldsymbol{b}\left(\psi_{\mathrm{r},K}\right)]$. With the imperfect knowledge of CSI, the equivalent baseband channel's impulse response for edge user k can be rewritten as follows:

$$\widehat{\mathbf{H}}_{\mathrm{r},k}=\sqrt{N_{\mathrm{T}}N_{\mathrm{R}}}\alpha_{\mathrm{r},k}. \tag{15}$$

Consequently, the baseband equivalent channel's impulse response for all K mobile users can be expressed as in the following equation:

$$\widehat{\mathbf{H}}_{\mathrm{r}}=\sqrt{N_{\mathrm{T}}N_{\mathrm{R}}}\begin{bmatrix}\alpha_{\mathrm{r},1}&0&\cdots&0\\0&\alpha_{\mathrm{r},2}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\alpha_{\mathrm{r},K}\end{bmatrix}. \tag{16}$$

In addition, the rows of the estimation error $\widehat{\boldsymbol{\epsilon}}_{\mathrm{r}}$ are mutually independent and distributed as $\mathcal{CN}(\mathbf{0},\mathbf{D}_{\mathrm{r}}-\widehat{\boldsymbol{D}}_{\mathrm{r}})$, where where $\widehat{\boldsymbol{D}}_{\mathrm{r}}$ is a diagonal matrices whose kth diagonal element can be expressed as

$$\left(\lambda_{\mathrm{r},k}\right)^2\triangleq\frac{\omega\mathcal{P}_{\mathrm{r},k}\left(\alpha_{\mathrm{r},k}\right)^2}{\omega\mathcal{P}_{\mathrm{r},k}\alpha_{\mathrm{r},k}+1} \tag{17}$$

where $\alpha_{\mathrm{r},k}$ denotes the path gain of the mmWave m-MIMO wireless channels, specified by Eq. (16), ω represents the length of the pilot sequences, and $\mathcal{P}_{\mathrm{r},k}$ is the transmit power from the mmWave m-MIMO BS to edge user k. Furthermore, as for the digital baseband precoder $\mathbf{F}^{\mathrm{B}}_{\mathrm{r}}$ at the mmWave m-MIMO BS, the ZF digital precoder can be applied as follows:

$$\mathbf{F}^{\mathrm{B}}_{\mathrm{r}}=\left(\widehat{\mathbf{H}}_{\mathrm{r}}\right)^{\dagger}\left(\widehat{\mathbf{H}}_{\mathrm{r}}\left(\widehat{\mathbf{H}}_{\mathrm{r}}\right)^{\dagger}\right)^{-1}. \tag{18}$$

By applying the baseband precoder at the mmWave m-MIMO BS, we assume that the multiuser interference can be perfectly cancelled with very large arrays, that is, $\mathbf{H}_{\mathrm{r},k}\mathbf{F}_{\mathrm{r},j}=0$ for $i\neq k$. Accordingly, we can derive the data

transmission rate, denoted by $R_{\mathrm{r},k}^{\mathrm{sgl}}$, over the kth relay link as follows:

$$R_{\mathrm{r},k}^{\mathrm{sgl}} = T_f B \log_2 \left(\det \left(\mathbf{I}_{N_\mathrm{s}} + \mathbf{\Gamma}_{\mathrm{r},k}^{\mathrm{sgl}} \right) \right) \tag{19}$$

where $\det(\cdot)$ denotes the determinant of a matrix, $\mathbf{I}_{N_\mathrm{s}}$ is the $N_\mathrm{s} \times N_\mathrm{s}$ identity matrix, T_f represents the frame duration, and

$$\mathbf{\Gamma}_{\mathrm{r},k}^{\mathrm{sgl}} \triangleq \frac{\frac{\mathcal{P}_{\mathrm{r},k}}{KN_\mathrm{s}} \left(\widehat{\mathbf{H}}_{\mathrm{r},k} \right)^\dagger \widehat{\mathbf{H}}_{\mathrm{r},k}}{\sum\limits_{j=1}^{K} \frac{(1-b_j)\mathcal{P}_{\mathrm{r},j}}{KN_\mathrm{s}} \left(\mathbf{F}_{\mathrm{r},j}^{\mathrm{B}} \right)^\dagger (\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k})^\dagger \widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k} \mathbf{F}_{\mathrm{r},j}^{\mathrm{B}} + \sigma^2 \mathbf{W}_{\mathrm{r},k} (\mathbf{W}_{\mathrm{r},k})^\dagger}. \tag{20}$$

Assume that the path gains to be invariant within a frame duration, but vary independently from one frame to another. Intuitively, we have $\left(\boldsymbol{a}\left(\phi_{\mathrm{r},k}\right)\right)^\dagger \boldsymbol{a}\left(\phi_{\mathrm{r},k}\right) = 1$. Since $\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k}$ and the precoding matrix $\mathbf{F}_{\mathrm{r},j}^{\mathrm{B}}$ are uncorrelated, we can obtain the following equation:

$$\mathbb{E}\left[\mathcal{P}_{\mathrm{r},j} \left(\mathbf{F}_{\mathrm{r},j}^{\mathrm{B}} \right)^\dagger (\widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k})^\dagger \widehat{\boldsymbol{\epsilon}}_{\mathrm{r},k} \mathbf{F}_{\mathrm{r},j}^{\mathrm{B}} \right] = \left(\alpha_{\mathrm{r},k} - \lambda_{\mathrm{r},k}^2 \right) \mathcal{P}_{\mathrm{r},k}. \tag{21}$$

Correspondingly, the data transmission rate over the kth relay link for single-path wireless channels can be written as follows:

$$R_{\mathrm{r},k}^{\mathrm{sgl}} = T_f B \log_2 \left(\det \left(\mathbf{I}_{N_\mathrm{s}} + \widetilde{\mathbf{\Gamma}}_{\mathrm{r},k}^{\mathrm{sgl}} \right) \right) \tag{22}$$

where

$$\widetilde{\mathbf{\Gamma}}_{\mathrm{r},k}^{\mathrm{sgl}} \triangleq \frac{\mathcal{P}_{\mathrm{r},k} \left(\widehat{\mathbf{H}}_{\mathrm{r},k} \right)^\dagger \widehat{\mathbf{H}}_{\mathrm{r},k}}{\sum\limits_{j=1}^{K} (1-b_j) \left(\alpha_{\mathrm{r},k} - \lambda_{\mathrm{r},k}^2 \right) \mathcal{P}_{\mathrm{r},k} + KN_\mathrm{s}\sigma^2}. \tag{23}$$

Similarly, we can derive the data transmission rate, denoted by $R_{\mathrm{d},k}^{\mathrm{sgl}}$, over the kth D2D link as follows:

$$\begin{aligned} R_{\mathrm{d},k}^{\mathrm{sgl}} &= T_f B \log_2 \left(1 + \gamma_{\mathrm{d},k} \right) \\ &= T_f B \log_2 \left(\det \left(\mathbf{I}_{N_\mathrm{s}} + \frac{\frac{\mathcal{P}_{\mathrm{d},k}}{KN_\mathrm{s}} \left(\mathbf{H}_{\mathrm{d},k} \right)^\dagger \mathbf{H}_{\mathrm{d},k}}{\sum\limits_{j=1}^{N} \frac{b_j \mathcal{P}_{\mathrm{d},j}}{KN_\mathrm{s}} \left(\mathbf{H}_{\mathrm{d},j} \right)^\dagger \mathbf{H}_{\mathrm{d},j} + \sigma^2} \right) \right) \end{aligned} \tag{24}$$

where $\gamma_{\mathrm{d},k}$ denotes the SNR between the relay node k and the corresponding D2D receiver.

Case 2: *(Multi-Path Wireless Channel)* Over the multi-path wireless fading channels, the hybrid block diagonalization scheme [8] need to be considered for the low-complexity multiuser mmWave m-MIMO system. In order to calculate the RF combining vector $\mathbf{W}_{\mathrm{r},k}$ at each edge user k, we need to proceed with the following steps. First, we can define the effective channel's impulse response matrix, denoted by $\overline{\mathbf{H}}_{\mathrm{r},k}$, excluding edge user k as shown in the following equation:

$$\overline{\mathbf{H}}_{\mathrm{r},k} \triangleq \left[\left(\widehat{\mathbf{H}}_{\mathrm{r},1} \right)^T, \ldots, \widehat{\mathbf{H}}_{\mathrm{r},k-1}^T, \widehat{\mathbf{H}}_{\mathrm{r},k+1}^T, \ldots, \left(\widehat{\mathbf{H}}_{\mathrm{r},K} \right)^T \right]^T \tag{25}$$

where $\widehat{\mathbf{H}}_{\mathrm{r},k}$ is the estimation of the baseband equivalent channel's impulse response matrix at edge user k. Second, the singular-value decomposition (SVD) of the matrix $\overline{\mathbf{H}}_{\mathrm{r},k}$ is defined as follows:

$$\overline{\mathbf{H}}_{\mathrm{r},k} \triangleq \overline{\mathbf{U}}_{\mathrm{r},k} \overline{\mathbf{\Sigma}}_{\mathrm{r},k} \left(\overline{\mathbf{V}}_{\mathrm{r},k} \right)^\dagger = \overline{\mathbf{U}}_{\mathrm{r},k} \overline{\mathbf{\Sigma}}_{\mathrm{r},k} \left[\overline{\mathbf{V}}_{\mathrm{r},k}^{((K-1)L_\mathrm{R})} \overline{\mathbf{V}}_{\mathrm{r},k}^{(L_\mathrm{R})} \right]^\dagger \tag{26}$$

where the columns of $\overline{\mathbf{U}}_{\mathrm{r},k}$ are the left singular vectors of matrix $\overline{\mathbf{H}}_{\mathrm{r},k}$, $\overline{\mathbf{\Sigma}}_{\mathrm{r},k}$ denotes the diagonal matrix containing the singular values of matrix $\overline{\mathbf{H}}_{\mathrm{r},k}$, the rows of $(\overline{\mathbf{V}}_{\mathrm{r},k})^\dagger$ are the right singular vectors for matrix $\overline{\mathbf{H}}_{\mathrm{r},k}$, $\overline{\mathbf{V}}_{\mathrm{r},k}^{((K-1)L_\mathrm{R})}$ represents the first $(K-1)L_\mathrm{R}$ right singular vectors of matrix $\overline{\mathbf{H}}_{\mathrm{r},k}$, and $\overline{\mathbf{V}}_{\mathrm{r},k}^{(L_\mathrm{R})}$ is the rest L_R right singular vectors extracted from $\overline{\mathbf{V}}_{\mathrm{r},k}$ that are the orthogonal bases of the null space for matrix $\overline{\mathbf{H}}_{\mathrm{r},k}$, i.e., $\widehat{\mathbf{H}}_{\mathrm{r},j} \overline{\mathbf{V}}_{\mathrm{r},k}^{(L_\mathrm{R})} = 0$. Third, we can derive the block diagonalization for the baseband equivalent channel matrix $\widehat{\mathbf{H}}_{\mathrm{r},k}$, denoted by $\widehat{\mathbf{H}}_{\mathrm{r}}^{\mathrm{B}}$, for removing the inter-user interference as in the following equation [8]:

$$\widehat{\mathbf{H}}_{\mathrm{r}}^{\mathrm{B}} = \begin{bmatrix} \widehat{\mathbf{H}}_{\mathrm{r},1} \overline{\mathbf{V}}_{\mathrm{r},1}^{(L_\mathrm{R})} & 0 & \cdots & 0 \\ 0 & \widehat{\mathbf{H}}_{\mathrm{r},2} \overline{\mathbf{V}}_{\mathrm{r},2}^{(L_\mathrm{R})} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\mathbf{H}}_{\mathrm{r},K} \overline{\mathbf{V}}_{\mathrm{r},K}^{(L_\mathrm{R})} \end{bmatrix}. \tag{27}$$

We have $\widehat{\mathbf{H}}_{\mathrm{r}}^{\mathrm{B}} = \left[\widehat{\mathbf{H}}_{\mathrm{r},1}^{\mathrm{B}}, \ldots, \widehat{\mathbf{H}}_{\mathrm{r},K}^{\mathrm{B}} \right]$. Similar to **Case 1** in Section III-C, the rows of the estimation error $\widehat{\boldsymbol{\epsilon}}_{\mathrm{r}}$ for channel $\widehat{\mathbf{H}}_{\mathrm{r}}^{\mathrm{B}}$ are also mutually independent and distributed as $\mathcal{CN}(\mathbf{0}, \mathbf{D}_{\mathrm{r}}^{\mathrm{B}} - \widehat{\boldsymbol{D}}_{\mathrm{r}}^{\mathrm{B}})$, where $\widehat{\boldsymbol{D}}_{\mathrm{r}}^{\mathrm{B}}$ is a diagonal matrices whose kth diagonal element is represented by $\lambda_{\mathrm{r},k}^{\mathrm{B}}$. Accordingly, the overall baseband precoder at the mmWave m-MIMO BS can be expressed as follows [8]:

$$\mathbf{F}_{\mathrm{r}}^{\mathrm{B}} = \left[\overline{\mathbf{V}}_{\mathrm{r},1}^{(L_\mathrm{R})} \mathbf{V}_{\mathrm{r},1}^{(N_s)}, \ldots, \overline{\mathbf{V}}_{\mathrm{r},K}^{(L_\mathrm{R})} \mathbf{V}_{\mathrm{r},K}^{(N_s)} \right] \tag{28}$$

where $\mathbf{V}_{\mathrm{r},k}^{(N_s)}$ consists of the first N_s columns of matrix $\mathbf{V}_{\mathrm{r},k}$, which can be derived by the right singular vector of the SVD for matrix $\widehat{\mathbf{H}}_{\mathrm{r},k}^{\mathrm{B}}$ as in the following equation:

$$\widehat{\mathbf{H}}_{\mathrm{r},k}^{\mathrm{B}} = \mathbf{U}_{\mathrm{r},k} \mathbf{\Sigma}_{\mathrm{r},k} \left(\mathbf{V}_{\mathrm{r},k} \right)^\dagger \tag{29}$$

where the columns of $\mathbf{U}_{\mathrm{r},k}$ are the left singular vector of matrix $\widehat{\mathbf{H}}_{\mathrm{r},k}^{\mathrm{B}}$, $\mathbf{\Sigma}_{\mathrm{r},k}$ denotes the diagonal matrix containing the singular values of matrix $\widehat{\mathbf{H}}_{\mathrm{r},k}^{\mathrm{B}}$, and the rows of $(\mathbf{V}_{\mathrm{r},k})^\dagger$ represent the right singular vectors of matrix $\widehat{\mathbf{H}}_{\mathrm{r},k}^{\mathrm{B}}$. Finally, we can derive the RF combiner at user k as $\mathbf{W}_{\mathrm{r},k} = \mathbf{U}_{\mathrm{r},k}^{(N_s)}$, where $\mathbf{U}_{\mathrm{r},k}^{(N_s)}$ denotes the first N_s columns of matrix $\mathbf{U}_{\mathrm{r},k}$. Define the ith diagonal element of matrix $\mathbf{\Sigma}_{\mathrm{r},k}$ as $\mu_{\mathrm{r},k}^{(i)}$. Similar to the derivation of Eqs. (21) and (22) in **Case 1** in Section III-C, we can derive the downlink data transmission rate, denoted by $R_{\mathrm{r},k}^{\mathrm{mlt}}$, for the multi-path case with imperfect CSI as follows:

$$R_{\mathrm{r},k}^{\mathrm{mlt}} = T_f B \log_2 \left(\det \left(\mathbf{I}_{N_\mathrm{s}} + \mathbf{\Gamma}_{\mathrm{r},k}^{\mathrm{mlt}} \right) \right) \tag{30}$$

where

$$\mathbf{\Gamma}_{\mathrm{r},k}^{\mathrm{mlt}} \triangleq \frac{\mathcal{P}_{\mathrm{r},k} \left(\mathbf{\Sigma}_{\mathrm{r},k} \right)^\dagger \mathbf{\Sigma}_{\mathrm{r},k}}{\sum\limits_{j=1}^{K} (1-b_j) \left(\alpha_{\mathrm{r},k} - \left(\lambda_{\mathrm{r},k}^{\mathrm{B}} \right)^2 \right) \mathcal{P}_{\mathrm{r},k} + KN_\mathrm{s}\sigma^2}. \tag{31}$$

III. Cross-Layer Optimization Schemes for Heterogeneous Statistical Delay-Bounded QoS Provisioning Through Effective Capacity in the Asymptotic Regime

In this section, we develop the framework for cross-layer optimization schemes and derive optimal QoS driven resource allocation policies to maximize the effective capacity for our proposed mmWave m-MIMO scheme subject to heterogeneous statistical delay-bound QoS constraints in the asymptotic regime.

A. The Effective Capacity Theory

Based on large deviation principle (LDP), under sufficient conditions, the queue length process $\widetilde{Q}_k(t)$ converges in distribution to a random variable $\widetilde{Q}_k(\infty)$ such that

$$-\lim_{\widetilde{Q}_{\text{th},k}\to\infty}\frac{\log\left(\Pr\left\{\widetilde{Q}_k(\infty)>\widetilde{Q}_{\text{th},k}\right\}\right)}{\widetilde{Q}_{\text{th},k}}=\theta_k. \tag{32}$$

Eq. (32) states that the probability of the queue length exceeding a certain threshold $\widetilde{Q}_{\text{th},k}$ decays exponentially fast as the threshold $\widetilde{Q}_{\text{th},k}$ increases. The parameter θ_k is called QoS exponent for user k and plays a critically important role for statistical delay-bounded QoS guarantees. The larger θ_k corresponds to the more stringent QoS requirement, while the smaller θ_k leads to the looser delay constraint, which implies the system can only provide a looser QoS guarantee. For a certain $\theta_k > 0$, $\widetilde{Q}_{\text{th},k}$ denotes the queue-length bound. The effective capacity [24] is defined as the maximum constant arrival rate that a given service process can support in order to guarantee a QoS requirement specified by θ_k. Given a service process $R_k(t)$ $(t = 1, 2, \ldots)$, the effective capacity of the service process for user k, denoted by $EC(\theta_k)$, where $\theta_k > 0$, is defined as follows [24]:

$$EC(\theta_k)=-\frac{1}{\theta_k}\log\left(\mathbb{E}\left[e^{-\theta_k R_k(t)}\right]\right)=-\frac{1}{\theta_k}\log\left(\mathbb{E}\left[e^{-\theta_k R_k(1)}\right]\right) \tag{33}$$

where $\mathbb{E}(\cdot)$ is the expectation operation.

Furthermore, we apply the resource allocation policy, denoted by $\boldsymbol{\nu}_k \triangleq \boldsymbol{\nu}_k(\theta_k, \gamma_k)$, for user k which is a function of not only the SNR γ_k, but also the QoS exponent θ_k [25]. Applying the resource allocation policy, the instantaneous transmit power for user k becomes $\mathcal{P}_k(\boldsymbol{\nu}_k) = \boldsymbol{\nu}_k\overline{\mathcal{P}}_k$. We first assume that the channel is block fading. In addition, we further assume that all users are heterogeneous, that is, they have different resource allocation policies and also subject to the different unit mean power constraint. We define $\overline{\mathcal{P}}_k$ as the mean transmit power constraint for user k. Accordingly, the power-control law need to satisfy the mean power constraint as in the following equation:

$$\int_0^\infty \mathcal{P}_k(\boldsymbol{\nu}_k)p_{\alpha_k}(\alpha_k)d\alpha_k \le \overline{\mathcal{P}}_k \tag{34}$$

where $p_{\alpha_k}(\alpha_k)$ denotes the pdf of channel path gain over mmWave m-MIMO wireless fading channels.

B. Learning Based D2D Matching Algorithm Under Heterogeneous Statistical Delay-Bounded QoS Constraints

Consider the diverse delay-bounded QoS provisioning for different links at the same time, which represents the new heterogeneous statistical delay-bounded QoS provisioning framework and imposes many new challenges. The maximization problem $\mathbf{P_1}$ under heterogeneous statistical delay-bounded QoS requirements to maximize the effective capacity, denoted by $EC^{\max}(\theta_{\text{r},k}, \theta_{\text{d},k})$, over the kth relay-D2D link can be formulated as in the following equation:

$$\mathbf{P_1}: EC^{\max}(\theta_{\text{r},k},\theta_{\text{d},k})=\arg\max_{\{\mathcal{P}_{\text{r},k},\mathcal{P}_{\text{d},k}\}}\left\{-\frac{1}{\theta_{\iota,k}}\times\log\left\{\mathbb{E}_\gamma\left[e^{-\theta_{\iota,k}T_f B\min\{\log_2(1+\gamma_{\text{r},k}),\log_2(1+\gamma_{\text{d},k})\}}\right]\right\}\right\} \tag{35}$$

$$\begin{aligned}\text{s.t. } C1:&\ \mathcal{P}_{\text{r},k},\mathcal{P}_{\text{d},k}>0\quad\forall k;\\ C2:&\ \mathbb{E}\left[\mathcal{P}_{\text{r},k}+\mathcal{P}_{\text{d},k}\right]\le\overline{\mathcal{P}}_k;\\ C3:&\ b_k\in\{0,1\}.\end{aligned} \tag{36}$$

where $\gamma_{\text{r},k}$ and $\gamma_{\text{r},k}$ denote the SNR for the kth relay link and D2D link, respectively; $\theta_{\iota,k} = \{\theta_{\text{r},k}, \theta_{\text{d},k}\}$ $(\iota = \{\text{r}, \text{d}\})$ is the QoS exponent for the kth relay link and D2D link, respectively, with $\iota = \text{r}$ if $R_{\text{r},k} < R_{\text{d},k}$ and $\iota = \text{d}$ if $R_{\text{r},k} > R_{\text{d},k}$; and $\overline{\mathcal{P}}_k$ represents the average transmit power constraint for the relay-D2D link. The above optimization problem $\mathbf{P_1}$ is a mixed-integer optimization problem.

Due to the statistic nature of mobile users' mobility and channel state information, we propose a learning based D2D matching algorithm to solve the cross-layer design based effective capacity maximization problem over mmWave m-MIMO based 5G mobile wireless networks. To solve the collaborative-learning based D2D matching problem, we need to define the following four elements:

- Agents: K users.
- States: Define the multimedia D2D caching based state, denoted by $\boldsymbol{s}(t) \in \mathcal{S}$, where $\mathcal{S}$ is the state space, for all D2D users at time slot t, which is given as follows:

$$\boldsymbol{s}(t) = [\boldsymbol{s}_1(t), \boldsymbol{s}_2(t), \ldots, \boldsymbol{s}_K(t)] \tag{37}$$

  where $\boldsymbol{s}_k(t)$ represents the configuration of D2D user D_k given as follows:

$$\boldsymbol{s}_k(t) = (\boldsymbol{c}_k(t), \boldsymbol{r}_k(t)) \tag{38}$$

  where $\boldsymbol{c}_k(t)$ represents the caching state and $\boldsymbol{r}_k(t)$ is the requesting state at D2D user D_k, which can be given in the following equations:

$$\begin{cases}\boldsymbol{c}_k=\left[c_k^{(1)},c_k^{(2)},\ldots,c_k^{(F)}\right];\\ \boldsymbol{r}_k=\left[r_k^{(1)},r_k^{(2)},\ldots,r_k^{(F)}\right],\end{cases} \tag{39}$$

  where $\mathcal{F} = \{1, 2, \ldots, F\}$ is the multimedia file library, $c_k^{(f)}$ is the caching indicator variable which determines whether D2D user D_k caches multimedia file indexed by f $(f \in \mathcal{F})$, and $r_k^{(f)}$ is the requesting indicator variable

which represents whether multimedia file indexed by f is requested at D2D user D_k.
- Actions: Define $b_k \in \mathcal{A}$ as the action for D2D user k where $\mathcal{A}$ represents the action space.
- Rewards: Once the requested packet is delivered, a reward is collected at the D2D user immediately. Define the reward for kth D2D link as $EC(\theta_{\mathrm{r},k}, \theta_{\mathrm{d},k})$ given in Eq. (35).

The Q-value update depends on the states $\{\widetilde{\boldsymbol{s}}_k, \boldsymbol{s}_k\}$, actions b_k, and the corresponding reward, thus, can be derived as follows:

$$Q_{t+1}(\boldsymbol{s}_k, b_k) \leftarrow (1-\widehat{\alpha}) Q_t(\boldsymbol{s}_k, b_k) + \widehat{\alpha} \times \left[EC(\theta_{\mathrm{r},k}, \theta_{\mathrm{d},k}) + \delta \max_{b_k \in \mathcal{A}} \sum_{\widetilde{\boldsymbol{s}}_k \in \mathcal{S}} Q_t(\widetilde{\boldsymbol{s}}_k, b_k) \right] \tag{40}$$

where $0 < \widehat{\alpha} \leq 1$ denotes the learning rate, δ is the discount factor, and $\max_{b_k \in \mathcal{A}} Q_t(\widetilde{\boldsymbol{s}}, b_k)$ represents the estimate of the optimal future Q value based on the best selected actions at time slot t. In addition, we assume that D2D user D_k is able to use the history of neighboring devices to estimate the user preference for the multimedia files. Define the historical knowledge, denoted by $\boldsymbol{\mathcal{H}}_k(t)$, for D2D user D_k up to time slot t as follows:

$$\boldsymbol{\mathcal{H}}_k(t) = \left[(\boldsymbol{c}_k(1), \boldsymbol{r}_k(1)), \ldots, (\boldsymbol{c}_k(t), \boldsymbol{r}_k(t)) \right], \quad \forall k \tag{41}$$

At each time slot t, each mobile user obtains a sample set which includes H different most recent observations from $\mathcal{H}_k(t)$. Define the sample set, denoted by $\widehat{\mathcal{S}}(\boldsymbol{s}_k(t), \boldsymbol{\mathcal{H}}_k(t))$ as a function of H observations in $\mathcal{H}_k(t)$. In each iteration, the learning algorithm selects a cluster of non-interacting D2D user, calculates the current reward functions, and finds the best D2D matching strategy sets independently. Finally, the D2D users update their new reward functions and the D2D matching strategy. The pseudo code of the learning based D2D matching algorithm is proposed in **Algorithm 1**.

Algorithm 1 Learning Based D2D Matching Algorithm

Input: $T_f, T_{\max}, B, K$, and $\mathcal{F}$
Initialization: $[\mathcal{P}_{\mathrm{r},1}, \ldots, \mathcal{P}_{\mathrm{r},K}] = [\overline{\mathcal{P}}_{\mathrm{r},1}, \ldots, \overline{\mathcal{P}}_{\mathrm{r},K}]$ and $[\mathcal{P}_{\mathrm{d},1}, \ldots, \mathcal{P}_{\mathrm{d},K}] = [\overline{\mathcal{P}}_{\mathrm{d},1}, \ldots, \overline{\mathcal{P}}_{\mathrm{d},K}]$
for each state $\boldsymbol{s}_k$ and action b_k **do**
  $Q(\boldsymbol{s}_k, b_k) = 0$
end for
Step 1:
Mobile users broadcasts their locations and configuration information and form the bipartite graph $\mathcal{G}$ with all neighboring users
Step 2:
while $t \leq T_{\max}$ **do**
  With exploitation probability $(1-\epsilon)$ do
    Randomly select a sample set of H recent observations of the mobile users' joint actions played at network state $\boldsymbol{s}_k(t)$
    Observe current state and construct the best action for each network state $\boldsymbol{s}_k(t)$
  With probability ϵ do
    Randomly select an action
  Calculate the immediate reward $EC(\theta_{\mathrm{r},k}, \theta_{\mathrm{d},k})$
  Update the Q-value function using Eq. (40)
  Set $t \leftarrow (t+1)$
end while

IV. Heterogeneous Statistical Delay QoS Guarantees Through Effective Capacity in the Non-Asymptotic Regime

In the previous section, we have derived the D2D matching algorithm by applying the Shannon's capacity to characterize the maximum effective capacity subject to the heterogeneous statistical delay-bounded QoS constraints in the asymptotic regime. However, under the low latency requirements for the 5G multimedia data transmissions, the traditional Shannon's capacity is no longer appropriate to characterize the maximum achievable data transmission rate because of the block error probability introduced by the finite blocklength coding in the non-asymptotic regime. Towards this end, in this section we investigate the non-asymptotic case by utilizing the FBC technique for supporting both the statistical delay-bounded and error-rate bounded requirements over 5G mobile wireless networks.

A. FBC Based Block Error Probability Over mmWave M-MIMO Based Wireless Fading Channels in the Non-Asymptotic Regime

Consider a message set $\mathcal{M} = \{1, \ldots, M\}$ and a message i which is uniformly distributed on $\mathcal{M}$. We define an $(n, M, N_{\mathrm{T}}, \epsilon)$-code ($\epsilon \in [0, 1)$) over mmWave m-MIMO based 5G mobile wireless networks as follows:

- An encoder Υ: $\{1, \ldots, M\} \mapsto \mathbb{C}^{N_{\mathrm{T}} \times n}$ that maps the message $i \in \{1, \ldots, M\}$ to a codeword with length n, i.e., $\mathbf{X}^n = \Upsilon(i)$, where $\mathbf{X}^n$ represents the encoded codeword with length n.
- A decoder $\{\mathcal{D}_{\mathbf{H}}\}_{\mathbf{H} \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{T}}}}$: $\mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{T}}} \times \mathbb{C}^{N_{\mathrm{T}} \times n} \mapsto \{1, \ldots, M\} \bigcup \{e\}$ such that $\mathcal{D}_{\mathbf{H}}(\mathbf{Y}^n) = \widehat{i}$, where $\widehat{i}$ denotes the estimated received signal at the receiver, $\mathbf{Y}^n$ represents the received codeword with length n, $\mathbf{H}$ is the channel's impulse response matrix, and e is the error event. The average error probability, denoted by P_e^n, need to satisfy the following maximum error probability constraint:

$$P_e^n \triangleq \frac{1}{M} \sum_{w=1}^{M} \mathbb{E}_{\mathbf{H}} \left[\Pr \left\{ \widehat{i} \neq i | \mathbf{X}^n, \mathbf{H} \right\} \right] \leq \epsilon. \tag{42}$$

Given the blocklength n and block error probability ϵ, the maximum achievable code size, denoted by M^*, is defined as:

$$M^* \triangleq \max \{ M : \exists (n, M, N_{\mathrm{T}}, \epsilon)\text{-code} \}. \tag{43}$$

Accordingly, we can derive the maximal achievable finite blocklength coding rate, denoted by R^*, as follows:

$$R^* = \frac{\log M^*}{n}. \tag{44}$$

In addition, we define $\boldsymbol{i}(\mathbf{X}^n;\mathbf{Y}^n,\mathbf{H})$ as the information density, which is given as in the following equation:

$$\begin{aligned}\boldsymbol{i}(\mathbf{X}^n;\mathbf{Y}^n,\mathbf{H}) &\triangleq \log\frac{\partial P_{Y^n|H,X^n}\left(\mathbf{Y}^n|\mathbf{H},\mathbf{X}^n\right)}{\partial P_{Y^n|H}\left(\mathbf{Y}^n|\mathbf{H}\right)}\\ &= \frac{1}{nN_{\mathrm{T}}}\sum_{t=1}^{n}\log\frac{\partial P_{Y_t|H,X_t}\left(\boldsymbol{y}_t|\mathbf{H},\boldsymbol{x}_t\right)}{\partial P_{Y_t|H}\left(\boldsymbol{y}_t|\mathbf{H}\right)}\\ &= \frac{1}{nN_{\mathrm{T}}}\sum_{t=1}^{n}\boldsymbol{i}_t \end{aligned} \tag{45}$$

where $\log(\cdot)$ represents $\log_e(\cdot)$, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the transmit and codewords at time slot t $(t=1,\ldots,n)$, respectively, and $\boldsymbol{i}_t$ denotes the random variable with denotes the random variable with the same distribution of

$$\boldsymbol{i}_t \triangleq \log\frac{\partial P_{Y_t|H,X_t}\left(\boldsymbol{y}_t|\mathbf{H},\boldsymbol{x}_t\right)}{\partial P_{Y_t|H}\left(\boldsymbol{y}_t|\mathbf{H}\right)}. \tag{46}$$

B. Coding Rate in the Non-Asymptotic Regime

In the non-asymptotic regime, we first define $n_{\mathrm{r},k}$ as the number of symbols transmitted during one time frame (i.e., one blocklength of channel coding) for user k. In particular, researchers have derived the accurate approximation of the maximum achievable data transmission rate, denoted by $R^*_{\mathrm{r},k}$, for user k with error probability, denoted by $\epsilon_{\mathrm{r},k}$ with $0 \le \epsilon_{\mathrm{r},k} < 1$ and coding blocklength, denoted by $n_{\mathrm{r},k}$, over the real AWGN channel in the non-asymptotic regime as shown in the following equation [26]:

$$\begin{aligned}R^*_{\mathrm{r},k} &\triangleq \frac{\log M^*_{\mathrm{r},k}}{n_{\mathrm{r},k}}\\ &= C(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) - \sqrt{\frac{V(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}}\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2} + \frac{O(\log n_{\mathrm{r},k})}{n_{\mathrm{r},k}}\end{aligned} \tag{47}$$

where $Q^{-1}(\cdot)$ is the inverse of Q-function, $M^*_{\mathrm{r},k}$ is the maximum achievable code size for user k, $C(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$ and $V(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$ represent the channel capacity and channel dispersion, respectively, and $f(x)=O(g(x))$ if and only if there exists a positive real number M and a real number x_0 such that $|f(x)| \le Mg(x)$ for all $x \ge x_0$. Eq. (47) implies that, given a blockcode with finite length $n_{\mathrm{r},k}$, the maximum achievable data transmission rate can be accurately approximated by the right-hand side of Eq. (47) with block error probability no larger than $\epsilon_{\mathrm{r},k}$ over the real AWGN channel. Authors in [27] have shown that the above Eq. (47) holds true when the blocklength $n_{\mathrm{r},k}$ is as short as 100. In the following section, we will analyze the expressions for channel capacity and channel dispersion over mmWave m-MIMO wireless fading channels.

C. Channel Capacity and Channel Dispersion Over mmWave Based M-MIMO Fading Channels in the Non-Asymptotic Regime

Given the D2D mode selection indicator variable $\{b_k\}_{k=1}^{K}$, we define K_{r} and K_{d} as the number of users using cellular links and D2D links, respectively. Similar to the cases in the asymptotic regime, we investigate two different cases with and without the applying the hybrid processing structure in the non-asymptotic regime as follows.

Case 1: *(Single-Path Wireless Channel)* Similar to the case in the asymptotic regime, we can derive the channel capacity, denoted by $C^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$, over single-path mmWave m-MIMO wireless fading channels under perfect CSI scenario in the non-asymptotic regime as follows:

$$\begin{aligned}C^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) &= \log_2\left(\det\left(\mathbf{I}_{N_{\mathrm{s}}} + \frac{\mathcal{P}_{\mathrm{r},k}\left(\widetilde{\mathbf{H}}_{\mathrm{r},k}\right)^{\dagger}\widetilde{\mathbf{H}}_{\mathrm{r},k}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)\right)\\ &= \log_2\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right).\end{aligned} \tag{48}$$

In addition, we can define the channel dispersion, denoted by $V^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$, over mmWave m-MIMO based 5G mobile wireless networks in the non-asymptotic regime as in the following equation:

$$V^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) \triangleq \frac{1}{n_{\mathrm{r},k}}\sum_{t=1}^{n_{\mathrm{r},k}}\mathrm{Var}\left[\boldsymbol{i}_t\right] \tag{49}$$

where $\boldsymbol{i}_t$ is given in Eq. (46). Accordingly, we can derive the channel dispersion over the single-path mmWave m-MIMO based wireless fading channels in the non-asymptotic regime as follows [28]:

$$V^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) = 1-\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{-2}. \tag{50}$$

Case 2: *(Multi-Path Wireless Channel)* We can derive the channel capacity, denoted by $C^{\mathrm{mlt}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$, over multi-path mmWave m-MIMO wireless fading channels under perfect CSI scenario in the non-asymptotic regime as follows:

$$C^{\mathrm{mlt}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) = \log_2\left(\det\left(\mathbf{I}_{N_{\mathrm{s}}} + \frac{\mathcal{P}_{\mathrm{r},k}\left(\boldsymbol{\Sigma}_{\mathrm{r},k}\right)^{\dagger}\boldsymbol{\Sigma}_{\mathrm{r},k}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)\right). \tag{51}$$

In addition, the channel dispersion, denoted by $V^{\mathrm{mlt}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$, over the multi-hop mmWave m-MIMO based wireless fading channels in the non-asymptotic regime can be expressed as in the following equation:

$$V^{\mathrm{mlt}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) = N_{\mathrm{T}} + \mathrm{Var}\left[\log\left(\det\left(\frac{\mathcal{P}_{\mathrm{r},k}\left(\boldsymbol{\Sigma}_{\mathrm{r},k}\right)^{\dagger}\boldsymbol{\Sigma}_{\mathrm{r},k}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)\right)\right]. \tag{52}$$

D. Optimal Power Allocation Policies Subject to Statistical Delay-Bounded QoS Constraint in the Non-Asymptotic Regime

Define $\boldsymbol{\mathcal{P}}_{\mathrm{r}} \triangleq [\mathcal{P}_{\mathrm{r},1},\ldots,\mathcal{P}_{\mathrm{r},K_{\mathrm{r}}}]$ and $\overline{\boldsymbol{\mathcal{P}}}_{\mathrm{r}} \triangleq [\overline{\mathcal{P}}_{\mathrm{r},1},\ldots,\overline{\mathcal{P}}_{\mathrm{r},K_{\mathrm{r}}}]$ as the resource allocation vector and the average resource allocation vectors, respectively. Using Eq. (33), we can define the new concept of ϵ-effective capacity for the $(n_{\mathrm{r},k},M,N_{\mathrm{T}},\epsilon_{\mathrm{r},k})$-code as follows:

Definition 1: For an $(n_{\mathrm{r},k},M,N_{\mathrm{T}},\epsilon_{\mathrm{r},k})$-code $(\epsilon_{\mathrm{r},k}\in[0,1))$, given the error probability $\epsilon_{\mathrm{r},k}$, we can derive the ϵ-effective capacity, denoted by $EC_{\epsilon}(\theta_{\mathrm{r},k})$, between edge user k and

the mmWave m-MIMO BS in the non-asymptotic regime as follows:

$$EC_{\epsilon}(\theta_{\mathrm{r},k}) = -\frac{1}{\theta_{\mathrm{r},k}} \log\left(\mathbb{E}_{\alpha_{\mathrm{r},k}}\left[(1-\epsilon_{\mathrm{r},k})\, e^{-\frac{\theta_{\mathrm{r},k}\log M^{*}_{\mathrm{r},k}}{n_{\mathrm{r},k}}}\right]\right). \tag{53}$$

Then, using Eqs. (47) and (53), we can formulate the optimization problem $\mathbf{P_2}$ for maximizing the aggregate ϵ-effective capacity, denoted by $EC^{\max}_{\epsilon}(\theta_{\mathrm{r}})$, for all K cellular users under heterogeneous statistical delay-bounded QoS constraints in the non-asymptotic regime as follows:

$$\begin{aligned}\mathbf{P_2}: & \; EC^{\max}_{\epsilon}(\theta_{\mathrm{r}}) \\ & = \arg\max_{\boldsymbol{\mathcal{P}}_{\mathrm{r}}}\left\{\sum_{k=1}^{K_{\mathrm{r}}} -\frac{1}{\theta_{\mathrm{r},k}}\log\left(\mathbb{E}_{\alpha_{\mathrm{r},k}}\left[(1-\epsilon_{\mathrm{r},k})\, e^{-\frac{\theta_{\mathrm{r},k}\log M^{*}_{\mathrm{r},k}}{n_{\mathrm{r},k}}}\right]\right)\right\}\end{aligned} \tag{54}$$

s.t. $C1$;

$$\begin{aligned} & C4: \mathbb{E}\left[\sum_{k=1}^{K_{\mathrm{r}}}\mathcal{P}_{\mathrm{r},k}\right] \le \overline{\mathcal{P}}_{\mathrm{r}}; \\ & C5: \frac{\log M^{*}_{\mathrm{r},k}}{n_{\mathrm{r},k}} \approx C(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) - \sqrt{\frac{V(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}}\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2}; \\ & C6: n_{\mathrm{r},k} \ge n^{\mathrm{th}}_{\mathrm{r},k}, \quad \forall k, \end{aligned} \tag{55}$$

where $\overline{\mathcal{P}}_{\mathrm{r}}$ is the average transmit power for all K links at the mmWave m-MIMO BS and $n^{\mathrm{th}}_{\mathrm{r},k}$ represents the minimum blocklength for constraint $C5$ given in Eq. (55) to hold true. Then, we can convert the non-convex maximization problem $\mathbf{P_2}$ into a minimization problem $\mathbf{P_3}$ as in the following equation:

$$\begin{aligned}\mathbf{P_3}: \; \arg\min_{\boldsymbol{\mathcal{P}}_{\mathrm{r}}} & \left\{\frac{1}{\theta_{\mathrm{r,o}}}\log\left(\mathbb{E}\left[\sum_{k=1}^{K_{\mathrm{r}}}(1-\epsilon_{\mathrm{r},k})\exp\left\{-\theta_{\mathrm{r},k}\right.\right.\right. \\ & \left.\left.\left.\times\left(C(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k} - \sqrt{\frac{V(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}}\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2}\right)\right\}\right]\right)\right\}\end{aligned} \tag{56}$$

subject to the constraints $C1, C4, C5$, and $C6$ specified by Eqs. (36) and (55) where $\theta_{\mathrm{r,o}}$ ($\theta_{\min} \le \theta_{\mathrm{r,o}} \le \theta_{\max}$ and $\theta_{\min} = \min\{\theta_{\mathrm{r},1}, \theta_{\mathrm{r},2}, \ldots, \theta_{\mathrm{r},K_{\mathrm{r}}}\}, \theta_{\max} = \max\{\theta_{\mathrm{r},1}, \theta_{\mathrm{r},2}, \ldots, \theta_{\mathrm{r},K_{\mathrm{r}}}\}$) denotes the unique optimal QoS exponent and the approximation for its optimal value of $\theta_{\mathrm{r,o}}$ in the high-end SINR region is derived in [29]. Accordingly, to solve the minimization problem $\mathbf{P_3}$, we investigate the following two different cases over single-path and multi-path wireless channels for our proposed multiuser mmWave m-MIMO schemes over 5G mobile wireless networks, respectively.

Case 1: *(Single-Path Wireless Channel)* For the single-path case, according to Eq. (56), we can rewrite problem $\mathbf{P_3}$ as in the following minimization problem $\mathbf{P_4}$:

$$\begin{aligned}\mathbf{P_4}: \; \min_{\boldsymbol{\mathcal{P}}_{\mathrm{r}}} & \left\{\mathbb{E}_{\alpha_{\mathrm{r},k}}\left[\sum_{k=1}^{K_{\mathrm{r}}}(1-\epsilon_{\mathrm{r},k})\exp\left\{-\theta_{\mathrm{r},k}\left(C^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})\right.\right.\right.\right. \\ & \left.\left.\left.\left. - \sqrt{\frac{V^{\mathrm{sgl}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}}\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2}\right)\right\}\right]\right\} \\ = \min_{\boldsymbol{\mathcal{P}}_{\mathrm{r}}} & \left\{\mathbb{E}_{\alpha_{\mathrm{r},k}}\left[\prod_{k=1}^{K_{\mathrm{r}}}(1-\epsilon_{\mathrm{r},k})\exp\left\{\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2}\right.\right.\right. \\ & \left.\left.\times\sqrt{\frac{1}{n_{\mathrm{r},k}}\left(1-\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right)^{-2}\right)}\right\} \\ & \left.\times\log_2\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right)^{-\widetilde{\beta}_{\mathrm{r},k}}\right]\right\}\end{aligned} \tag{57}$$

subject to the constraints $C1, C4, C5$, and $C6$ specified by Eqs. (36) and (55) where $\widetilde{\beta}_{\mathrm{r},k} \triangleq \theta_{\mathrm{r},k}/\log 2$.

Lemma 1: Given the error probability $\epsilon_{\mathrm{r},k} \in (0, 0.5)$ and the transmit power $\mathcal{P}_{\mathrm{r},k}$, the ϵ-effective capacity $EC_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r}})$ specified by Eq. (53) over the single-path mmWave m-MIMO based wireless channel is a continuous and monotonically increasing function of the blocklength $n_{\mathrm{r},k} > 0$.

Proof: In order to show the continuity and monotonicity of the ϵ-effective capacity in terms of the blocklength $n_{\mathrm{r},k} > 0$, we need to proceed with the following steps. First, we define the following function:

$$\begin{aligned}F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) \triangleq & \; n_{\mathrm{r},k}\log_2\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right) - \frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2} \\ & \times\sqrt{n_{\mathrm{r},k}\left(1-\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right)^{-2}\right)}.\end{aligned} \tag{58}$$

Using the definition of continuity, we can prove that for every $\Delta > 0$, there always exists a $\delta > 0$ such that for all $\widetilde{n}_{\mathrm{r},k}$, we have

$$|n_{\mathrm{r},k} - \widetilde{n}_{\mathrm{r},k}| < \delta, \tag{59}$$

such that the following equation holds:

$$|F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) - F(\widetilde{n}_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})| < \Delta. \tag{60}$$

Second, we can derive the relationship between the ϵ-effective capacity and the function $F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$ as in the following equation:

$$EC_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r}}) = -\sum_{k=1}^{K_{\mathrm{r}}}\frac{1}{\theta_{\mathrm{r},k}}\log\mathbb{E}_{\alpha_{\mathrm{r},k}}\left[(1-\epsilon_{\mathrm{r},k})\, e^{-\theta_{\mathrm{r},k}\frac{F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}}\right]. \tag{61}$$

Accordingly, we can show that the ϵ-effective capacity $EC_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r}})$ is a continuous function of the blocklength $n_{\mathrm{r},k}$. In addition, the first-order derivative of the function $F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$ can be determined as follows:

$$\begin{aligned}\frac{\partial F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{\partial n_{\mathrm{r},k}} & = \log_2\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right) - \frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{2\log 2\sqrt{n_{\mathrm{r},k}}} \\ & \quad \times\sqrt{1-\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right)^{-2}} \\ & = \frac{1}{2}\left(\log_2\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^{2}}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^{2}}\right) - \frac{\log M^{*}_{\mathrm{r},k}}{n_{\mathrm{r},k}}\right)\end{aligned} \tag{62}$$

which is due to the followings:

$$\sqrt{\frac{1}{n_{\mathrm{r},k}}\left(1-\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{-2}\right)}\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2} = \log_2\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)-\frac{\log M^*_{\mathrm{r},k}}{n_{\mathrm{r},k}}. \quad (63)$$

Correspondingly, we can prove that $\frac{\partial F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{\partial n_{\mathrm{r},k}} > 0$. Therefore, $F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$ is an increasing function of the blocklength $n_{\mathrm{r},k} > 0$. Accordingly, when $Q^{-1}(\epsilon_{\mathrm{r},k}) > 0$, i.e., $\epsilon_{\mathrm{r},k} \in (0, 0.5)$, we need to analyze the monotonicity of function $\frac{F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}$ as follows:

$$\begin{aligned}\frac{\partial\left[\frac{F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}\right]}{\partial n_{\mathrm{r},k}} &= \frac{n_{\mathrm{r},k}F'_{n_{\mathrm{r},k}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})-F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{(n_{\mathrm{r},k})^2}\\ &= \frac{\sqrt{n_{\mathrm{r},k}}Q^{-1}(\epsilon_{\mathrm{r},k})\sqrt{1-\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}(\alpha_{\mathrm{r},k})^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{-2}}}{2\log 2(n_{\mathrm{r},k})^2}\\ &> 0. \end{aligned} \quad (64)$$

where $F'_{n_{\mathrm{r},k}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k}) \triangleq \frac{\partial F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{\partial n_{\mathrm{r},k}}$. Since e^{-x} is a monotonically decreasing function and $\log(x)$ is a monotonically increasing function, we can observe that the ϵ-effective capacity $EC_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r},k})$ for user k is a monotonically increasing function of $F(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})$. Thus, we can show that the aggregate ϵ-effective capacity is a monotonically increasing function of the blocklength $n_{\mathrm{r},k} > 0$, given the error probability $\epsilon_{\mathrm{r},k}$, and transmit power $\mathcal{P}_{\mathrm{r},k}$, completing the proof of Lemma 1. ∎

Theorem 1: Given the error probability $\epsilon_{\mathrm{r},k} \in (0, 0.5)$, an optimal resource allocation policy, denoted by $\mathcal{P}^{\mathrm{r,opt}}_{k,\mathrm{sgl}}$, that maximizes the aggregate ϵ-effective capacity, denoted by $EC^{\max}_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r}})$, for the single-path wireless channel data transmissions between the mmWave m-MIMO BS and user k under statistical delay-bounded and error-rate bounded QoS constraints in the non-asymptotic regime is specified as follows:

$$\mathcal{P}^{\mathrm{r,opt}}_{k,\mathrm{sgl}} = \frac{\widetilde{\beta}_{\mathrm{r},k}\prod_{k=1}^{K_{\mathrm{r}}}\left((1-\epsilon_{\mathrm{r},k})e^{\frac{\theta_{\mathrm{r},k}Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n_{\mathrm{r},k}}\log 2}}\right)^{\frac{1}{K_{\mathrm{r}}\widetilde{\beta}_{\mathrm{r},k}+1}}}{\left(\widetilde{\lambda}^{\mathrm{sgl}}_{\mathrm{o}}\right)^{\frac{1}{K_{\mathrm{r}}\widetilde{\beta}_{\mathrm{r},k}+1}}\prod_{k=1}^{K_{\mathrm{r}}}\left(\frac{\widetilde{\beta}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}(\alpha_{\mathrm{r},k})^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}}\widetilde{\beta}_{\mathrm{r},k}+1}}} - \frac{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}{N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2} \quad (65)$$

where $\widetilde{\lambda}^{\mathrm{sgl}}_{\mathrm{o}}$ is the optimal Lagrange multiplier which can be numerically obtained by substituting Eq. (65) back into constraint $C4$ specified in Eq. (55) for single-path wireless channel.

Proof: In order to derive the optimal resource allocation policy over the single-path mmWave m-MIMO wireless fading channels, we need to proceed with the following steps. First, we can form the Lagrange function, denoted by J, as in the following equations:

$$\begin{aligned}J = \mathbb{E}_{\alpha_{\mathrm{r},k}}\Bigg[\prod_{k=1}^{K_{\mathrm{r}}}(1-\epsilon_{\mathrm{r},k})\exp\Bigg\{\frac{\theta_{\mathrm{r},k}Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2\sqrt{n_{\mathrm{r},k}}}\\ \times\sqrt{\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)}\Bigg\}\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{-\widetilde{\beta}_{\mathrm{r},k}}\Bigg]\\ +\widetilde{\lambda}^{\mathrm{sgl}}\left(\mathbb{E}_{\alpha_{\mathrm{r},k}}\left[\sum_{k=1}^{K_{\mathrm{r}}}\mathcal{P}_{\mathrm{r},k}\right]-\overline{\mathcal{P}}_{\mathrm{r}}\right)\end{aligned} \quad (66)$$

where $\widetilde{\lambda}^{\mathrm{sgl}}$ represents the Lagrange multiplier associated with the constraint $C4$ given in Eq. (55) for single-path wireless channel. Second, applying the Karush-Kuhn-Tucker (KKT) condition, we can take the first-order derivative of J with respect to $\mathcal{P}_{\mathrm{r},k}$ and set the results to zero as follows:

$$\begin{aligned}\frac{\partial J}{\partial\mathcal{P}_{\mathrm{r},k}} &= -(1-\epsilon_{\mathrm{r},k})e^{\frac{\theta_{\mathrm{r},k}Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n}\log 2}}\frac{\widetilde{\beta}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\\ &\quad\times\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{-1}\\ &\quad\times\prod_{k=1}^{K_{\mathrm{r}}}\left(1+\frac{\mathcal{P}_{\mathrm{r},k}N_{\mathrm{T}}N_{\mathrm{R}}\left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)^{-\widetilde{\beta}_{\mathrm{r},k}}\\ &\quad\times p_{\alpha_{\mathrm{r},k}}(\alpha_{\mathrm{r},k})d\alpha_{\mathrm{r},k}+\widetilde{\lambda}^{\mathrm{sgl}}p_{\alpha_{\mathrm{r},k}}(\alpha_{\mathrm{r},k})d\alpha_{\mathrm{r},k}=0.\end{aligned} \quad (67)$$

Third, using Eq. (67), under statistical delay-bounded and error-rate bounded QoS constraints, an optimal resource allocation policy that maximizes the ϵ-effective capacity under FBC in the high SNR region can be derived as in Eq. (65), completing the proof of Theorem 1. ∎

Case 2: *(Multi-Path Wireless Channel)* Similarly, we can derive the aggregate ϵ-effective capacity over the multi-path mmWave m-MIMO based wireless fading channels in the non-asymptotic regime as in the following equations:

$$\begin{aligned}EC_{\epsilon,\mathrm{mlt}}(\theta_{\mathrm{r}}) &= -\frac{1}{\theta_{\mathrm{r,o}}}\log\Bigg\{\mathbb{E}_{\alpha_{\mathrm{r},k}}\Bigg[\sum_{k=1}^{K_{\mathrm{r}}}(1-\epsilon_{\mathrm{r},k})\exp\Bigg\{-\theta_{\mathrm{r},k}\\ &\quad\times\left(C^{\mathrm{mlt}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})-\sqrt{\frac{V^{\mathrm{mlt}}(n_{\mathrm{r},k},\mathcal{P}_{\mathrm{r},k})}{n_{\mathrm{r},k}}}\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2}\right)\Bigg\}\Bigg]\Bigg\}\\ &= -\frac{1}{\theta_{\mathrm{r,o}}}\log\Bigg\{\mathbb{E}_{\Lambda}\Bigg[\prod_{k=1}^{K_{\mathrm{r}}}(1-\epsilon_{\mathrm{r},k})\exp\Bigg\{\frac{Q^{-1}(\epsilon_{\mathrm{r},k})}{\log 2}\\ &\quad\times\sqrt{\frac{1}{n_{\mathrm{r},k}}\left(N_{\mathrm{T}}+\mathrm{Var}\left[\log\left(\prod_{i=1}^{\widetilde{N}}\mu^{(i)}_{\mathrm{r},k}\right)\right]\right)}\Bigg\}\\ &\quad\times\left(\sum_{i=1}^{\widetilde{N}}\left(1+\frac{\mathcal{P}_{\mathrm{r},k}\left(\Lambda_{\mathrm{r},i}\right)^2}{K_{\mathrm{r}}N_{\mathrm{s}}\sigma^2}\right)\right)^{-\widetilde{\beta}_{\mathrm{r},k}}\Bigg]\Bigg\}\end{aligned} \quad (68)$$

where $\widetilde{N} = \min\{N_{\mathrm{T}}, N_{\mathrm{R}}\}$ and $\{\Lambda_{\mathrm{r},i}, i = 1,\ldots,\widetilde{N}\}$ denotes the eigenvalue of square matrix $\left(\left(\mathbf{\Sigma}_{\mathrm{r},k}\right)^{\dagger}\mathbf{\Sigma}_{\mathrm{r},k}\right)$. Accordingly, we can obtain the following upper bound on the aggregate

ϵ-effective capacity in the non-asymptotic regime:

$$EC_{\epsilon,\text{mlt}}(\theta_\text{r}) \leq \frac{1}{\theta_{\text{r,o}}} \log \left\{ \mathbb{E}_\Lambda \left[\prod_{k=1}^{K_\text{r}} (1-\epsilon_{\text{r},k}) \exp \left\{ \sqrt{\frac{N_\text{T}}{n_{\text{r},k}}} \times \frac{Q^{-1}(\epsilon_{\text{r},k})}{\log 2} \right\} \left(\sum_{i=1}^{\widetilde{N}} \left(1 + \frac{\mathcal{P}_{\text{r},k} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right) \right)^{-\widetilde{\beta}_{\text{r},k}} \right] \right\} \quad (69)$$

Correspondingly, we can rewrite the minimization problem $\mathbf{P_3}$ in the high SNR region as in the following optimization problem $\mathbf{P_5}$:

$$\begin{aligned} \mathbf{P_5}: \min_{\boldsymbol{\mathcal{P}}_\text{r}} & \left\{ \mathbb{E}_\Lambda \left[\prod_{k=1}^{K_\text{r}} (1-\epsilon_{\text{r},k}) \exp \left\{ \sqrt{\frac{N_\text{T}}{n_{\text{r},k}}} \frac{Q^{-1}(\epsilon_{\text{r},k})}{\log 2} \right\} \times \left(\sum_{i=1}^{\widetilde{N}} \left(1 + \frac{\mathcal{P}_{\text{r},k} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right) \right)^{-\widetilde{\beta}_{\text{r},k}} \right] \right\} \\ \approx \min_{\boldsymbol{\mathcal{P}}_\text{r}} & \left\{ \mathbb{E}_\Lambda \left[\prod_{k=1}^{K_\text{r}} (1-\epsilon_{\text{r},k}) \exp \left\{ \sqrt{\frac{N_\text{T}}{n_{\text{r},k}}} \frac{Q^{-1}(\epsilon_{\text{r},k})}{\log 2} \right\} \times \left(\frac{\mathcal{P}_{\text{r},k} \sum_{i=1}^{\widetilde{N}} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right)^{-\widetilde{\beta}_{\text{r},k}} \right] \right\}. \end{aligned} \quad (70)$$

Similar to **Case 1** in Section V-D, we can derive the following lemma to analyze the convexity for the minimization problem $\mathbf{P_5}$ over the multi-path mmWave m-MIMO wireless channels.

Lemma 2: Given the error probability $\epsilon_{\text{r},k} \in (0, 0.5)$ and the transmit power $\mathcal{P}_{\text{r},k}$ in the high SNR region, the aggregate ϵ-effective capacity $EC_{\epsilon,\text{mlt}}(\theta_\text{r})$ specified by Eq. (68) is a continuous and monotonically increasing function of the blocklength $n_{\text{r},k} > 0$.

Proof: The proof of Lemma 2 is similar to the proof for Lemma 1 and thus is omitted in this paper. ∎

Theorem 2: In the high SNR region, the optimal resource allocation policy, denoted by $\mathcal{P}^{\text{r,opt}}_{k,\text{mlt}}$, that maximizes the aggregate ϵ-effective capacity for the wireless data transmissions between the mmWave m-MIMO BS and edge user k under statistical delay-bounded and error-rate bounded QoS constraints in the non-asymptotic regime is determined by the following equation:

$$\mathcal{P}^{\text{r,opt}}_{k,\text{mlt}} = \frac{\widetilde{\beta}_{\text{r},k} \prod_{k=1}^{K_\text{r}} \left((1-\epsilon_{\text{r},k}) e^{\sqrt{\frac{N_\text{T}}{n_{\text{r},k}}} \frac{Q^{-1}(\epsilon_{\text{r},k})}{\log 2}} \right)^{\frac{1}{K_\text{r}\widetilde{\beta}_{\text{r},k}+1}}}{\left(\widetilde{\lambda}^{\text{mlt}}_\text{o}\right)^{\frac{1}{K_\text{r}\widetilde{\beta}_{\text{r},k}+1}} \prod_{k=1}^{K_\text{r}} \left(\frac{\widetilde{\beta}_{k\text{r},} \sum_{i=1}^{\widetilde{N}} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right)^{\frac{\widetilde{\beta}_{\text{r},k}}{K_\text{r}\widetilde{\beta}_{\text{r},k}+1}}} \quad (71)$$

where $\widetilde{\lambda}^{\text{mlt}}_\text{o}$ is the optimal Lagrange multiplier which can be numerically obtained by substituting Eq. (71) back into constraint $C4$ specified in Eq. (55) over multi-path mmWave m-MIMO wireless fading channels.

Proof: In order to derive the optimal resource allocation policy over the multi-path mmWave m-MIMO wireless fading channels, we need to proceed with the following steps. First, we can form the Lagrange function, denoted by J, as shown in the following equations:

$$\begin{aligned} J = \mathbb{E}_\Lambda & \left[\prod_{k=1}^{K_\text{r}} (1-\epsilon_{\text{r},k}) \exp \left\{ \sqrt{\frac{N_\text{T}}{n_{\text{r},k}}} \frac{Q^{-1}(\epsilon_{\text{r},k})}{\log 2} \right\} \times \left(\frac{\mathcal{P}_{\text{r},k} \sum_{i=1}^{\widetilde{N}} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right)^{-\widetilde{\beta}_{\text{r},k}} \right] \\ & + \widetilde{\lambda}^{\text{mlt}} \left(\mathbb{E}_{\alpha_{\text{r},k}} \left[\sum_{k=1}^{K_\text{r}} \mathcal{P}_{\text{r},k} \right] - \overline{\mathcal{P}}_\text{r} \right) \end{aligned} \quad (72)$$

where $\widetilde{\lambda}^{\text{mlt}}$ represents the Lagrange multiplier associated with the constraint $C4$ given in Eq. (55) for multi-path wireless channel. Then, applying the KKT condition, we can take the derivative of J with respect to $\mathcal{P}_{\text{r},k}$ and set the results to zero as follows:

$$\begin{aligned} \frac{\partial J}{\partial \mathcal{P}_{\text{r},k}} = & -(1-\epsilon_{\text{r},k}) e^{\sqrt{\frac{N_\text{T}}{n_{\text{r},k}}} \frac{Q^{-1}(\epsilon_{\text{r},k})}{\log 2}} \frac{\widetilde{\beta}_{\text{r},k} \sum_{i=1}^{\widetilde{N}} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \\ & \times \left(\frac{\mathcal{P}_{\text{r},k} \sum_{i=1}^{\widetilde{N}} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right)^{-1} \prod_{k=1}^{K_\text{r}} \left(\frac{\mathcal{P}_{\text{r},k} \sum_{i=1}^{\widetilde{N}} \left(\Lambda_{\text{r},i}\right)^2}{K_\text{r} N_\text{s} \sigma^2} \right)^{-\widetilde{\beta}_{\text{r},k}} \\ & \times p_{\alpha_{\text{r},k}}(\alpha_{\text{r},k}) d\alpha_{\text{r},k} + \widetilde{\lambda}^{\text{mlt}} p_{\alpha_{\text{r},k}}(\alpha_{\text{r},k}) d\alpha_{\text{r},k} = 0. \end{aligned} \quad (73)$$

Consequently, using Eq. (73), under statistical delay-bounded QoS constraints, we can derive the optimal resource allocation policy that maximizes the ϵ-effective capacity using FBC in the high SNR region as in Eq. (71). As a result, the proof of Theorem 2 is completed. ∎

E. Maximizing ϵ-Effective Capacity in the Non-Asymptotic Regime

To derive the maximum aggregate ϵ-effective capacity, we can investigate the following two different cases under signal-path wireless channel and multi-path wireless channels in the non-asymptotic regime as follows.

Case 1: *(Single-Path Wireless Channel)* Using Eq. (57) and our proposed optimal resource allocation policy specified in Eq. (65) in Theorem 1, we can derive the maximum aggregate ϵ-effective capacity, denoted by $EC^{\max}_{\epsilon,\text{sgl}}(\theta_\text{r})$, in the non-asymptotic regime as follows:

$$EC^{\max}_{\epsilon,\text{sgl}}(\theta_\text{r}) = \sum_{k=1}^{K_\text{r}} -\frac{1}{\theta_{\text{r},k}} \log \left\{ \mathbb{E}_{\alpha_{\text{r},k}} \left[(1-\epsilon_{\text{r},k}) f(\alpha_{\text{r},k}) g(\alpha_{\text{r},k}) \right] \right\} \quad (74)$$

where $f(\alpha_{\text{r},k})$ and $g(\alpha_{\text{r},k})$ are defined in Eq. (75), as shown at the bottom of the next page. Since $\mathbb{E}_{\alpha_{\text{r},k}}[f(\alpha_{\text{r},k}) g(\alpha_{\text{r},k})] > \mathbb{E}_{\alpha_{\text{r},k}}[f(\alpha_{\text{r},k})] \mathbb{E}_{\alpha_{\text{r},k}}[g(\alpha_{\text{r},k})]$, we can obtain an approximation

of the maximum aggregate ϵ-effective capacity, denoted by $EC^{\max}_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r}})$, as in Eq. (76), as shown at the bottom of this page.

Case 2: *(Multi-Path Wireless Channel)* Similarly, using Eq. (71) in Theorem 2, in the non-asymptotic regime, we can derive the maximum aggregate ϵ-effective capacity over multi-path wireless channels in the high SNR region.

V. Performance Evaluations

We conduct simulation experiments and numerical analyses to validate and evaluate our proposed mmWave m-MIMO based schemes. Throughout our simulations and numerical analyses, we set the bandwidth $B = 100$ KHz, the time frame length $T_f = 2$ ms, the average transmit power for the mmWave m-MIMO BS $\overline{\mathcal{P}}_{\mathrm{r},k}$ can be choose from $[5, 20]$ Watt, the average transmit power for the kth relay node $\overline{\mathcal{P}}_{\mathrm{d},k}$ can be choose from $[1, 5]$ Watt, and $\omega = 2K$.

Assume that there are 2 mobile users ($K = 2$) in the entire wireless networks. Let $\theta_{\mathrm{r},1} = 0.001$ and $\theta_{\mathrm{r},2} = 0.5$. Set the number of receive antennas at the mobile user $N_{\mathrm{R}} = 10$ and data streams for each mobile user $N_s = 4$. Using Eqs. (22) and (30), Fig. 2 plots the achievable data transmission rate with different values of SNR for our proposed hybrid block diagonalization based model compared with the traditional ZF beamforming model for the multiuser mmWave m-MIMO scheme. We can observe from Fig. 2 that in the asymptotic regime, as the value or SNR increases, the achievable data transmission rate also increases and finally converges to a certain value. Compared with the traditional ZF beamforming model, Fig. 2 shows that our proposed hybrid block diagonalization model outperforms the tradition ZF beamforming model in terms of the achievable data transmission rate over mmWave m-MIMO based 5G mobile wireless networks.

Using the learning based D2D matching algorithm, Fig. 3 depicts the Q-value update given in Eq. (40). Given the number of transmit antennas N_{T}, Fig. 3 shows that as the number of iteration increases, Q-value update converges to a certain value, which indicates that our learning based D2D matching algorithm achieves the optimal D2D matching policy. Fig. 4 plots upper and lower bounds on the aggregate effective capacity with varying number of antennas at the BS over mmWave m-MIMO based 5G mobile wireless networks in the asymptotic regime. As shown in Fig. 4, as the value of

$$\begin{cases} f(\alpha_{\mathrm{r},k}) \triangleq \left(\dfrac{\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} \left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left((1-\epsilon_{\mathrm{r},k}) e^{\frac{\theta_{\mathrm{r},k} Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n_{\mathrm{r},k}} \log 2}} \right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}}{\left(\widetilde{\lambda}_{\mathrm{o}}^{\mathrm{sgl}}\right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left(\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} \left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \right)^{\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}} \right)^{-\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}} ; \\ g(\alpha_{\mathrm{r},k}) \triangleq \exp\left\{ \dfrac{\theta_{\mathrm{r},k} Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n} \log 2} \sqrt{\left(1 - \left(\dfrac{\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} \left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left((1-\epsilon_{\mathrm{r},k}) e^{\frac{\theta_{\mathrm{r},k} Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n_{\mathrm{r},k}} \log 2}} \right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}}{\left(\widetilde{\lambda}_{\mathrm{o}}^{\mathrm{sgl}}\right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left(\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} \left(\alpha_{\mathrm{r},k}\right)^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \right)^{\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}} \right)^{-2} \right)} \right\}. \end{cases} \tag{75}$$

$$\begin{aligned} EC^{\max}_{\epsilon,\mathrm{sgl}}(\theta_{\mathrm{r}}) \approx \sum_{k=1}^{K_{\mathrm{r}}} & -\frac{1}{\theta_{\mathrm{r},k}} \log\left(1-\epsilon_{\mathrm{r},k}\right) - \frac{1}{\theta_{\mathrm{r},k}} \log\left\{ \mathbb{E}_{\alpha_{\mathrm{r},k}} \left[\exp\left\{ \frac{\theta_{\mathrm{r},k} Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n} \log 2} \right. \right. \right. \\ & \left. \left. \left. \times \sqrt{\left(1 - \left(\frac{\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} (\alpha_{\mathrm{r},k})^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left((1-\epsilon_{\mathrm{r},k}) e^{\frac{\theta_{\mathrm{r},k} Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n_{\mathrm{r},k}} \log 2}} \right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}}{\left(\widetilde{\lambda}_{\mathrm{o}}^{\mathrm{sgl}}\right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left(\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} (\alpha_{\mathrm{r},k})^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \right)^{\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}} \right)^{-2} \right)} \right\} \right] \right\} \\ & - \frac{1}{\theta_{\mathrm{r},k}} \log\left\{ \mathbb{E}_{\alpha_{\mathrm{r},k}} \left[\left(\frac{\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} (\alpha_{\mathrm{r},k})^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left((1-\epsilon_{\mathrm{r},k}) e^{\frac{\theta_{\mathrm{r},k} Q^{-1}(\epsilon_{\mathrm{r},k})}{\sqrt{n_{\mathrm{r},k}} \log 2}} \right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}}{\left(\widetilde{\lambda}_{\mathrm{o}}^{\mathrm{sgl}}\right)^{\frac{1}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}} \prod\limits_{k=1}^{K_{\mathrm{r}}} \left(\frac{\widetilde{\beta}_{\mathrm{r},k} N_{\mathrm{T}} N_{\mathrm{R}} (\alpha_{\mathrm{r},k})^2}{K_{\mathrm{r}} N_{\mathrm{s}} \sigma^2} \right)^{\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}}} \right)^{-\frac{\widetilde{\beta}_{\mathrm{r},k}}{K_{\mathrm{r}} \widetilde{\beta}_{\mathrm{r},k}+1}} \right] \right\} \end{aligned} \tag{76}$$

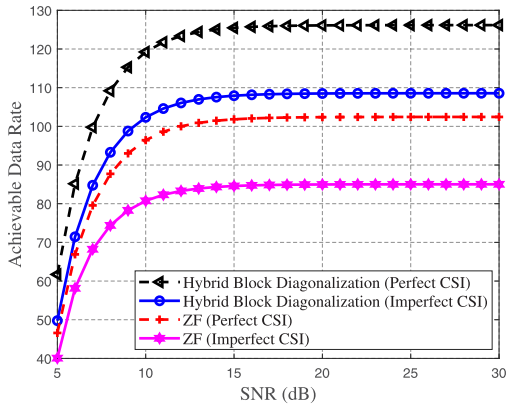


Fig. 2. Achievable data transmission rate for the proposed hybrid block diagonalization model v.s. tradition ZF beamforming model.

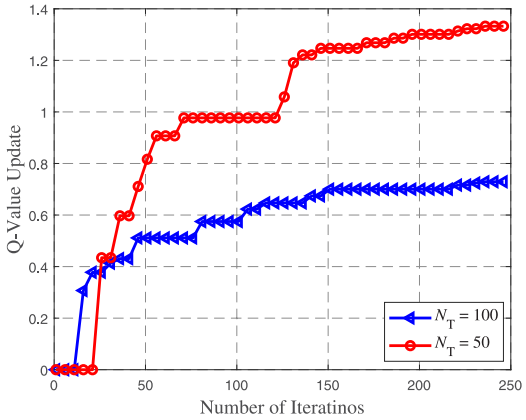


Fig. 3. The simulated Q-value update under learning based D2D matching algorithm over mmWave m-MIMO based 5G mobile wireless networks.

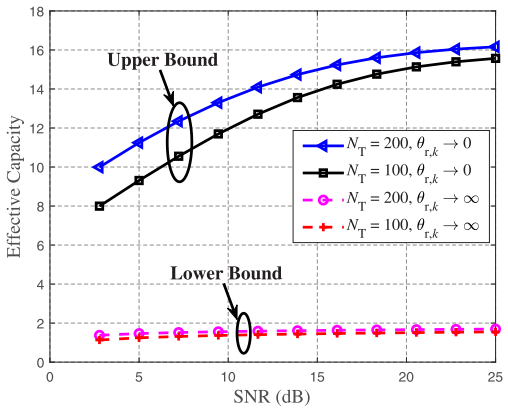


Fig. 4. Upper and lower bounds on the effective capacity with varying number of antennas at the mmWave m-MIMO BS in the asymptotic regime.

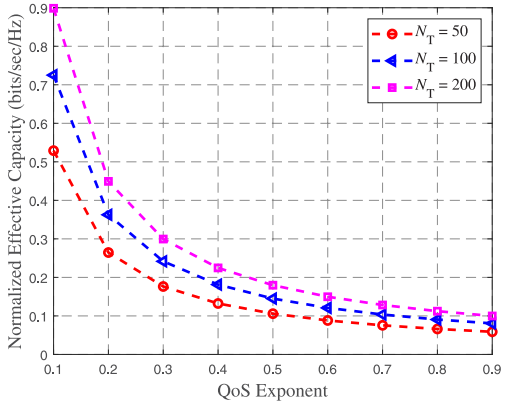


Fig. 5. The normalized effective capacity with varying QoS exponents in the asymptotic regime.

SNR increases, the aggregate effective capacity also increases. We can observe from Fig. 4 that as the number of transmit antenna increases, the system with larger antenna arrays at the mmWave m-MIMO BS achieves better effective capacity, implying that our proposed multiuser mmWave m-MIMO scheme outperforms the traditional multiuser MIMO scheme in terms of the effective capacity. Fig. 4 also shows that very loose QoS constraint ($\theta_{r,k} \to 0$) and very stringent QoS constraint ($\theta_{r,k} \to \infty$) set upper and lower bounds on the effective capacity, respectively. Also, as we increase the number of the transmit antennas, the effective capacity is also increased, which implies that larger antenna arrays at the BS achieve better aggregate effective capacity for our proposed multiuser mmWave m-MIMO scheme.

Setting $K = 5$, $N_R = 10$, and the data stream $N_s = 8$, Fig. 5 depicts the normalized aggregate effective capacity for our proposed multiuser mmWave m-MIMO scheme under different statistical delay-bounded QoS constraints in the asymptotic regime. As shown in Fig. 5, we can observe that as the QoS exponent θ becomes larger, the value of the aggregate effective capacity decreases. In addition, as the number of transmit antennas N_T gets larger, the gap between each curve for the normalized effective capacity becomes smaller. We can also observe from Fig. 5 that for all three curves, the normalized effective capacity increases as the value of N_T, which implies that as the antenna arrays becomes larger, the m-MIMO scheme achieves better performance under the statistical delay-bounded QoS constraints.

We set the block error probability $\epsilon_{r,k} \in \{10^{-6}, 10^{-3}\}$. Fig. 6 plots an upper bound on channel capacity with various SNR values for our proposed scheme over mmWave m-MIMO based 5G mobile wireless networks in the non-asymptotic regime. As shown in Fig. 6, given a desired SNR, the upper bound on channel capacity increases with the number of transmit antennas N_T. We can observe from Fig. 6 that the loose error-rate constraint $\epsilon_{r,k} = 1 \times 10^{-3}$ and stringent error-rate constraint $\epsilon_{r,k} = 1 \times 10^{-6}$ represent upper and lower bounds on the channel capacity, respectively.

Setting SNR to be 20 dB and blocklength $n_{r,k} = 800$, Fig. 7 plots the achievable data transmission rate vs. varying values of average error probability over mmWave m-MIMO based 5G mobile wireless networks in the non-asymptotic regime. Fig. 7 shows that the achievable finite blocklength coding rate decreases when the error-rate bound becomes more stringent, i.e., $\epsilon_{r,k} \to 0$. This implies that a larger value of $\epsilon_{r,k}$ corresponds to a more stringent error-rate requirement. On the other hand, a smaller value of $\epsilon_{r,k}$ implies a looser error-rate constraint.

Then, we set the SNR to be 20 dB. Fig. 8 depicts the achievable finite blocklength coding rate with different blocklengths over mmWave m-MIMO based 5G mobile wireless networks in the non-asymptotic regime. Fig. 8 shows that the achievable

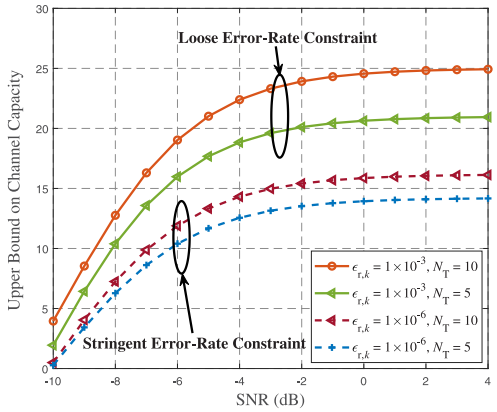


Fig. 6. An upper bound on channel capacity vs. blocklength over mmWave m-MIMO wireless channel model in the non-asymptotic regime.

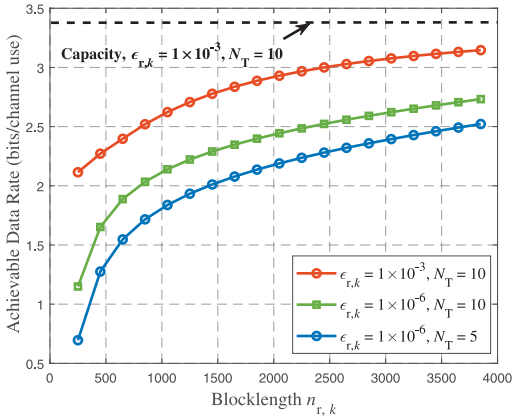


Fig. 8. Achievable data transmission rate vs. blocklength over mmWave m-MIMO wireless channel model in the non-asymptotic regime.

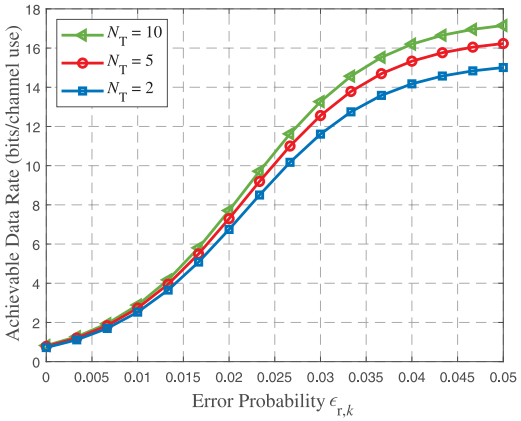


Fig. 7. Achievable data transmission rate vs. average error probability over mmWave m-MIMO wireless channel model in the non-asymptotic regime.

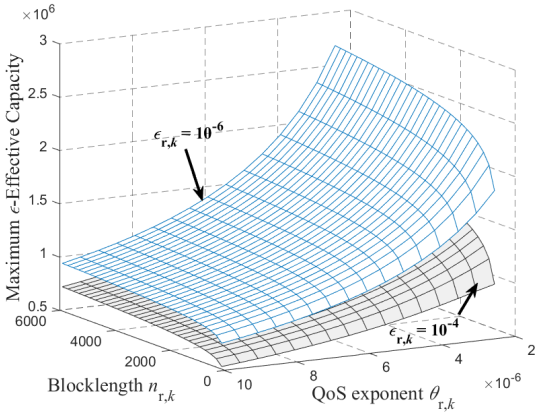


Fig. 9. Maximum ϵ-effective capacity vs. blocklength and QoS exponent over 5G mobile wireless networks in the non-asymptotic regime.

finite blocklength coding rate increases with the number of transmit antennas N_T at the BS over mmWave m-MIMO based 5G mobile wireless networks. We can observe in Fig. 8 that for a given error probability, the achievable finite blocklength coding rate is an increasing function of the corresponding blocklength. Fig. 8 also shows that we achieve a higher value of the achievable finite blocklength coding rate under a loose error-rate constraint. The dashed lines in Fig. 8 imply the channel capacity of the ideal model in which the data rate is equal to the instantaneous capacity, i.e., $n_{r,k} \to \infty$ and $\epsilon_{r,k} = 0$.

Setting the SNR to be 20 dB, Fig. 9 depicts the maximum ϵ-effective capacity with different blocklengths and QoS exponents for our proposed mmWave m-MIMO scheme in the non-asymptotic regime. We can observe in Fig. 9 that for a given error probability, the maximum ϵ-effective capacity is a decreasing function of the corresponding QoS exponent. Fig. 9 also shows that we achieve a higher value of maximum ϵ-effective capacity with a smaller error probability.

VI. Conclusions

We have proposed heterogeneous statistical QoS driven resource allocation policies by designing multiuser mmWave m-MIMO based schemes in both the asymptotic and non-asymptotic regimes. In particular, we have developed mmWave m-MIMO based 5G mobile networking system architectures and then established corresponding mmWave m-MIMO based 5G mobile wireless networking system models. Then, we have analyzed hybrid block diagonalization model for the mmWave m-MIMO scheme with imperfect knowledge of CSI. Given the heterogeneous statistical delay-bounded QoS constraints, we have formulated and solved the effective capacity by deriving optimal resource-allocation policies over relay-D2D link for our proposed mmWave m-MIMO scheme in both the asymptotic and non-asymptotic regimes. We also have conducted a set of simulations that validate our proposed schemes and show that they outperform the other existing schemes under heterogeneous statistical delay-bounded QoS constraints over multiuser mmWave m-MIMO based 5G mobile wireless networks.

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-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, Jun. 2008.

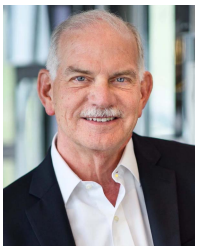
[3] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous statistical QoS provisioning over 5G mobile wireless networks," *IEEE Netw.*, vol. 28, no. 6, pp. 46–53, Nov. 2014.
[4] W. Cheng, X. Zhang, and H. Zhang, "Heterogeneous statistical QoS provisioning over 5G wireless full-duplex networks," in *Proc. IEEE INFOCOM*, Apr./May 2015, pp. 55–63.
[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
[6] H. Q. Ngo, H. A. Suraweera, M. Matthaiou, and E. G. Larsson, "Multipair full-duplex relaying with massive arrays and linear processing," *IEEE Trans. Veh. Technol.*, vol. 32, no. 9, pp. 1721–1737, May 2014.
[7] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
[8] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.
[9] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.
[10] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, Jr., "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.
[11] X. Ma, F. Yang, S. Liu, J. Song, and Z. Han, "Design and optimization on training sequence for mmWave communications: A new approach for sparse channel estimation in massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1486–1497, Jul. 2017.
[12] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[13] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2823–2835, Aug. 2015.
[14] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability," *IEEE Trans. Inf. Theory*, vol. 61, no. 8, pp. 5–21, Jan. 2014.
[15] 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.
[16] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.
[17] G. Ozcan and M. Gursoy, "Throughput of cognitive radio systems with finite blocklength codes," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541–2554, Nov. 2013.
[18] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Aug. 2016.
[19] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854–6883, Dec. 2016.
[20] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[21] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.
[22] I. Chafaa and M. Djeddou, "Improved channel estimation in mmWave communication system," in *Proc. Seminar Detection Syst. Architectures Technol. (DAT)*, Feb. 2017, pp. 1–5.
[23] B. Chen and C. Yang, "Energy costs for traffic offloading by cache-enabled D2D communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2016, pp. 1–6.
[24] 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.
[25] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.
[26] S. Xu, T.-H. Chang, S.-C. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.
[27] 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.
[28] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Dispersion of Gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2009, pp. 2204–2208.
[29] 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. 2018.

Xi Zhang (S'89–SM'98–F'16) 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 the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He is a fellow of the IEEE for contributions to quality of service (QoS) theory in mobile wireless networks. He was with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hill, NJ, USA, and AT&T Laboratories Research, Florham Park, NJ, USA, in 1997. He was a Research Fellow 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 350 research papers on wireless networks and communications systems, network protocol design and modeling, statistical communications, random signal processing, information theory, and control theory and systems. He received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He received five Best Paper Awards at IEEE ICC 2018, IEEE GLOBECOM 2014, IEEE GLOBECOM 2009, IEEE GLOBECOM 2007, and IEEE WCNC 2010, respectively. One of his IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS papers has been listed as the IEEE Best Readings Paper (receiving the highest citation rate among all IEEE Transactions/Journal papers in the area) on Wireless Cognitive Radio Networks and Statistical QoS Provisioning over Mobile Wireless Networking. He is an IEEE Distinguished Lecturer of both the IEEE Communications Society and the IEEE Vehicular Technology Society. He also received the TEES Select Yoiung Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering, Texas A&M University, College Station, TX, in 2006.

Professor Zhang is serving or has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, twice as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for two special issues on "Broadband Wireless Communications for High Speed Vehicles" and "Wireless Video Transmissions," an Associate Editor for IEEE COMMUNICATIONS LETTERS, twice as the Lead Guest Editor for *IEEE Communications Magazine* for two special issues on "Advances in Cooperative Wireless Networking" and "Underwater Wireless Communications and Networks: Theory and Applications," a Guest Editor for *IEEE Wireless Communications Magazine* for special issue on "Next Generation CDMA vs. OFDMA for 4G Wireless Applications," an Editor for 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* (Wiley), and an Area Editor for the *Journal on Computer Communications* (Elsevier), among many others. He is serving or has served as the TPC Chair for the IEEE GLOBECOM 2011, TPC Vice-Chair for IEEE INFOCOM 2010, TPC Area Chair for IEEE INFOCOM 2012, Panel/Demo/Poster Chair for ACM MobiCom 2011, General Chair for IEEE WCNC 2013, and TPC Chair for IEEE INFOCOM 2017-2019 Workshops on "Integrating Edge Computing, Caching, and Offloading in Next Generation Networks," 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 the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, under the supervision of Professor Xi Zhang. Her research interests focus on big data based 5G wireless networks technologies, statistical QoS provisioning, and cognitive radio networks. She won the Best Paper Award at the IEEE GLOBECOM 2014 and also received the Hagler Institute for Advanced Study Heep Graduate Fellowship Award from Texas A&M University in 2018.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign.

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

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