# Multi-Tier Caching for Statistical-QoS Driven Digital Twins Over mURLLC-Based 6G Massive-MIMO Mobile Wireless Networks Using FBC

Xi Zhang, *Fellow, IEEE*, Qixuan Zhu, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—Digital Twin (DT) has been widely envisioned as a major intelligent application of 6G wireless networks requiring stringent quality-of-service (QoS) for *massive ultra-reliable and low latency communications* (mURLLC) to support efficient interactions between physical and virtual objects. As a key multi-tier computing (MTC) technique of 6G mobile networks, multi-tier caching stores the highly-demanded data at different wireless network tiers to significantly reduce mURLLC-streaming delay and data move. However, how to efficiently cache mURLLC data at different caching tiers in wireless networks and how to support *both delay* and *error-rate* bounded QoS for DT remain challenging problems. To conquer these difficulties, in this paper we propose to integrate multi-tier caching with finite blocklength coding for supporting mURLLC-based DT by developing multi-tier 6G massive-multiple-input-multiple-output (M-MIMO) mobile networks. First, we develop the efficient inter-tier and intra-tier collaborative multi-tier caching mechanisms, where popular DT data items are selectively cached at different wireless network caching tiers including: router tier, M-MIMO base-station (BS)/WiFi-AP tier, and mobile device tier. Second, our proposed inter-tier caching mechanisms maximize the *aggregate caching gain*, in terms of DT-based $\epsilon$-*effective capacity*, across three caching tiers to support statistical delay and error-rate bounded QoS. Third, we develop the intra-tier caching algorithm to optimize each caching-tier's QoS. Finally, our extensive numerical analyses show our developed schemes' performances-superiorities over existing schemes.**

***Index Terms*—6G, DT, mURLLC, statistical delay and error-rate bounded QoS, FBC, DT-based $\epsilon$-effective capacity, M-MIMO, multi-tier caching.**

## I. Introduction

In RECENT years, the transformation of the physical domain into the virtual domain has been accelerated by the information revolution. In order to implement this fast-paced and continuously-evolving digital transformation, *digital twin (DT)* has been widely recognized as an emerging tool to represent a physical entity by its virtual counterpart, enabling the simulation, analysis, and monitoring, while producing real-time interactions between the physical and virtual twins. By transforming elements, functions, operations, and dynamics of a physical system into its digital form, we are able to control, test, analyze, predict, and simulate the physical object. Therefore, DT transmission has been widely recognized as a promising 6G traffic type which signifies a paradigm shift towards more intelligent, interconnected, and data-centric communication systems. As a digital representation of an intended real-world physical object, a digital twin need to be timely and reliably updated by and synchronized with the corresponding physical system. However, due to the massive geographical coverage scale and complex environments of mobile devices which are also called mobile users (MUs) throughout this paper, the new challenges on how to disseminate DT data content items with stringent low-delay and high-reliability QoS requirements need to be overcome.

Towards this end, *massive ultra-reliable and low-latency communications (mURLLC)* [1] has been proposed as a key technique to create ultra real-time, reliable, and high data-rate wireless networking environments for supporting *immersive and inter-operable* DT data streaming. This is because the 6G-based mURLLC guarantees the highly-stringent quality-of-service (QoS) standards including: extra-low end-to-end delay ($<$1 ms), super-reliability ($>$99.99999%), and extra-high data rate ($>$1 Tb/s), which can best meet DTs' QoS requirements. On the other hand, the physical-virtual synchronization of DT for composite heterogeneous services, such as *metaverse* and *virtual reality (VR)/augmented reality (AR)* in 6G, demands heavily loaded computational operations and networking resources consumptions. To address these issues, *multiple-tier computing* (MTC) techniques [2], [3], [4] have also been developed to support DT data-content items by providing distributed computation, processing, and storage capabilities at different tiers of wireless networks. Under MTC-based *edge computing architectures*, in-network caching stores highly-demanded data at different wireless networking tiers along network-edge devices to efficiently reduce DT streaming delay and data move [5]. Leveraging advanced caching techniques to store the frequently accessed DT data content items and even the mobile-applications software at edge nodes, the corresponding computation tasks can be executed at the network edge to reduce the latency, thus improving QoS performances. Furthermore, the recently proposed *finite blocklength coding* (FBC) has been also shown to be

Manuscript received 6 September 2023; revised 10 January 2024; accepted 7 February 2024. Date of publication 18 March 2024; date of current version 29 March 2024. This work of Xi Zhang and Qixuan Zhu was supported in part by the U.S. National Science Foundation under Grant CCF-2142890, Grant CCF-2008975, Grant ECCS-1408601, and Grant CNS-1205726, and in part by the U.S. Air Force under Grant FA9453-15-C-0423. The work of H. Vincent Poor was supported by U.S. National Science Foundation under Grant CNS-2128448 and Grant ECCS-2335876. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Kunlun Wang. *(Corresponding author: Xi Zhang.)*

Xi Zhang and Qixuan Zhu 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; qixuan@tamu.edu).

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

Digital Object Identifier 10.1109/JSTSP.2024.3377007

a powerful technique to further enhance mURLLC by enabling *small-packet communications* to reduce processing delay while constraining decoding error-probability without requiring the arbitrarily long and even infinite large packet-size or codeword length for achieving Shannon capacity. Thus, integrating FBC with mURLLC creates the alternative promising solution for efficiently implementing statistical-QoS driven DT over the 6G mobile networks.

There are some existing works addressing the above-discussed DT-enabling techniques. The authors of [6] designed an on-demand DT content item delivery architecture through a caching based communication mode to reduce the responding time for the frequent interactions between vehicular users and roadside units. The social-aware vehicular edge caching networks were proposed by [7], where the DT enables cache controllers to grasp the social relations between vehicles, understand the vehicle flow distribution, and effectively allocate communication and storage resources for content delivery. The authors of [8] proposed a novel scheme to support DT/metaverse communications by jointly integrating communication, computing, and storage techniques through the applications of mobile edge computing integrated with mURLLC.

However, how to efficiently cache mURLLC data at different caching tiers in mobile wireless networks and how to integrate multi-tier caching with FBC to support both statistical delay and error-rate bounded QoS for DT applications still remain challenging open problems. To overcome these difficulties, in this paper we propose to develop the DT-enabled multi-tier caching 6G mURLLC mobile networks by applying statistical-QoS/FBC. First, we develop the collaborative multi-tier hierarchical caching mechanisms in both inter-tier and intra-tier scenarios to support the DT transmissions, where the popular DT data content items are selectively cached at different network-edge caching tiers: including router-tier, massive multiple-input-multiple-output (massive MIMO)-base-station (BS)/WiFi-access-point (AP)-tier, and mobile-device-tier, respectively. To support the statistical delay and error-rate bounded QoS provisioning for the cached DT-data transmission, we develop the *DT-based $\epsilon$-effective capacity* to measure the performance metrics for mobile wireless networks. Second, we propose the inter-tier collaborative caching mechanisms to maximize the aggregate $\epsilon$-effective capacity as the DT-data *caching gain* across the above described three caching tiers. Third, we propose the intra-tier optimal caching algorithms at all three caching tiers, respectively, to maximize the $\epsilon$-effective capacity within each individual caching tier.

The rest of this paper is organized as follows. Section II establishes the system architectures models for our proposed DT-enabled multi-tier caching 6G mURLLC mobile networks using statistical-QoS/FBC. Section III derives closed-form expressions for the $\epsilon$-effective capacity if downloading a cached DT data item from BS and AP, respectively. Section IV develops the inter-tier collaborative hierarchical caching mechanisms across three caching tiers. Section V develops the intra-tier collaborative caching mechanisms when downloading the DT data cached at mobile devices, BS/AP, and router tiers, respectively. Section VI develops optimal hierarchical caching mechanisms for supporting adaptive data-blocklength to minimize total transmission delay. Section VII validates and evaluates our developed schemes through numerical analyses. The paper concludes with Section VIII.

## II. The System Models

As shown in Fig. 1, the system architecture model of our proposed digital-twin-enabled multi-tier caching 6G mURLLC mobile networks supported by Statistical-QoS/FBC consists of the following three major components: 1) Physical Twin, 2) Virtual Twin, and 3) Digital-Twin-Enabled Multi-Tier-Caching 6G mURLLC Mobile Networks based on Statistical-QoS/FBC. Note that while mURLLC can help DT applications, there are many DT-specific properties, characteristics, data-requirements, etc., which cannot be completely realized by mURLLC. Thus, these DT-specific properties/characteristics/data-requirements, such as DT real-time *data collection* and dynamic *data adaptation* between physical twin and virtual twin, physical object's digital twinning, synchronized DT signaling, etc. [9], need to be also taken into account when formulating our system models and control schemes. Along with real-time data collection, data adaptation captures the actual mobile network status and accurately models the physical twin's behaviors, enabling DT to re-configure mobile network resources at physical twin (see Fig. 1) for improving their operational efficiencies. However, in these cases DTs typically impose the highly heterogeneous and multi-dimensional QoS requirements for real-time data collection and data adaptation between physical twin and virtual twin.

For achieving timely dynamic state synchronizations between physical twin and virtual twin, the physical twin's data (system states and control information) is transformed/transmitted into/to its virtual twin [10] through data collection and data adaptation by encoding the physical twin's signal into a data block with the finite blocklength. To support the heterogeneous QoS requirements for DT-specific properties in our proposed multi-tier hierarchical caching schemes, this encoded data block is considered as an original DT data item to be cached at different caching tiers of the hybrid wireline and wireless networks. Then, based on the real-time *wireless network status* (e.g., the wireless channel states, total number of MUs, and MUs' statistical QoS requirements on the DT data item) and physical twin's behavior updates, the virtual twin interacts with the physical twin accordingly by adjusting its digital transformation updated models [11]. All MUs' request frequencies for all DT data-content items are then used to derive the DT data popularity distribution function, which dictates caching strategies for multi-tier networks' *design and performance evaluations/improvements*. Finally, MUs send real-time *synchronized feedback information*, such as dictation/steering information, sensing and signaling, and MUs' service requirements, back to the physical twin to dynamically control and adjust the physical twin.

As shown in Fig. 1, we propose a *collaborative multi-tier hierarchical caching network architecture* to support DT data transmissions over 6G mobile wireless networks, which caches the frequently requested DT data-content items along the edge of hybrid wireline and wireless networks to significantly reduce DT data transmission delay, interference, and decoding-error probability by minimizing the redundant data-move load in the Internet cloud and core networks.

Fig. 1 also shows that the edge of 6G wireless networks consists of the following three hierarchical and non-overlapped caching tiers: (1) Tier 1: routers, (2) Tier 2: massive-MIMO BS and WiFi AP, and (3) Tier 3: mobile devices.

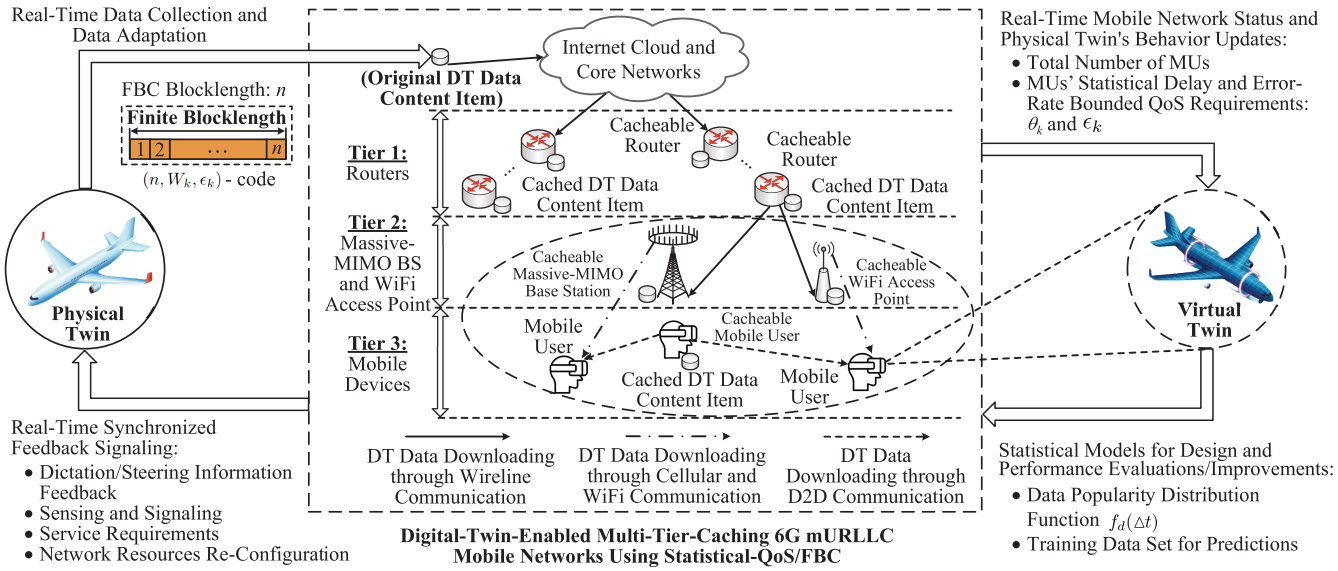



Fig. 1. System architecture model for our proposed **digital-twin-enabled multi-tier caching 6G mURLLC mobile wireless networks using statistical-QoS/FBC** between the **physical twin** and the **virtual twin**, which are able to cache the frequently requested DT data-content items in routers (**Tier 1**), massive-MIMO-BS/WiFi-AP (**Tier 2**), and/or mobile devices (**Tier 3**) along the edge of our DT-enabled multi-tier caching networks according to DT data and signaling's statistical-QoS requirements.

### *A. Caching Expenses Analysis and Statistical Delay and Error-Rate Bounded QoS Provisioning*

Assume that DT data content items can be cached through writing the file to a flash memory of a cacheable device. To store a data content item in a memory within the retention time (i.e., the duration time that a data item is saved in a cache memory), this data is programmed by a high threshold voltage, causing the damage of the hardware lifespan. We define this damage as the *caching expense*, denoted by $\xi(x)$, which is a function of the data retention time $x$, and $\xi(x)$ can be written as follows [12, Definition 1]:

$$\xi(x) = a_\sigma x^\sigma + a_{\sigma-1}x^{\sigma-1} + \cdots + a_1 x + a_0, \qquad (1)$$

where $\sigma$ is the damage degree and $a_j \geq 0, \forall j \in \{0, 1, 2, \ldots, \sigma\}$ are the coefficients for the damage function.

Let $K$ be the total number of MUs in a wireless cell, and let $k$, $\forall k \in \{1, 2, \ldots, K\}$, be the index for an MU. The streaming over mURLLC-based 6G wireless networks requires the stringent QoS on *both* statistical bounded transmission *delay* and *decoding error probability*. The FBC technique has been developed to enable *small packet communications* for adaptive error-control and real-time transmissions, where senders encode the message into short packets (i.e., packets with small numbers of bits) to reduce the packet processing and transmission delay while constraining and controlling the decoding error probability. We define an FBC scheme in the following definition.

*Definition 1:* Consider a fading channel which uses input blockcode set $\mathcal{A}$ and output blockcode set $\mathcal{B}$. We define that an $(n, W_k, \epsilon_k)$-code, $\forall k \in \{1, 2, \ldots, K\}$, for a state-dependent memoryless channel consists of [13]:

- A message set $\mathcal{W}_k = \{c_{1,k}, \ldots, c_{W_k,k}\}, \forall k$, with the cardinality $W_k$ and the message length equal to $\log_2(W_k)$.
- An encoder, which is a mapping: $\mathcal{W}_k \mapsto \mathcal{A}^n$, where $\mathcal{A}^n$ is the set of codewords with length $n$. At the receiver end, a decoder produces an estimate of the original message by observing the channel output, according to a function: $\mathcal{B}^n \mapsto \widehat{\mathcal{W}}_k$, where $\mathcal{B}^n$ is the set of received codewords with length $n$ and $\widehat{\mathcal{W}}_k$ is the estimation of $\mathcal{W}_k$.
- The decoding error probability for the $k$th MU, denoted by $\epsilon_k$, is defined as $\epsilon_k \triangleq (1/W_k) \sum_{w=1}^{W_k} \Pr\{c_{w,k} \neq \widehat{c}_{w,k}\}$, where $c_{w,k} \in \mathcal{W}_k, \widehat{c}_{w,k} \in \widehat{\mathcal{W}}_k$, and $\Pr\{\cdot\}$ is the probability of an event.

where usually $\epsilon_k > 0$ if $n < \infty$. ■

Thus, the triple-tuple $(n, W_k, \epsilon_k)$ represents that a source with the cardinality $W_k$ can successfully transmit messages with a probability of success $(1 - \epsilon_k)$ over $n$ channel uses.

The statistical delay-bounded QoS guarantees [14], [15], [16], [17], [18], [19] have been shown to be powerful in analyzing queuing behavior for the stochastic arrival and service processes over the time-varying wireless fading channels. The key statistical-QoS performance metric is the *effective capacity* which measures the maximum packet's constant arrival rate such that the given statistical *delay-bounded* QoS can be guaranteed. Based on the *large deviation principle* (LDP) [15], the queue-length process $Q_k(t)$ for the $k$th MU converges in distribution to a random variable $Q_k(\infty)$ such that

$$-\lim_{Q_{\text{th},k}\to\infty} \frac{\log(\Pr\{Q_k(\infty) > Q_{\text{th},k}\})}{Q_{\text{th},k}} = \theta_k, \ \forall k \qquad (2)$$

where $Q_{\text{th},k}$ is the queue length threshold (bound) and $\theta_k > 0$ is defined as the *QoS exponent* for MU $k$. The insights of Eq. (2) reveal that the probability of the queueing process exceeding a certain threshold $Q_{\text{th},k}$ decays exponentially fast at the rate of $\theta_k$ as the threshold $Q_{\text{th},k}$ increases and tends to infinity. As shown in [15], a smaller $\theta_k$ corresponds to a slower decay rate, which implies that the system can only provide a looser QoS guarantee, while a larger $\theta_k$ leads to a faster decay rate, which means that a more stringent QoS can be supported. When $\theta_k \to 0$, the system can tolerate long delay; when $\theta_k \to \infty$, the system cannot tolerate any delay.

However, the conventional statistical-QoS theory modeled by Eq. (2) focuses only on the statistical delay-bounded QoS

without considering the transmission reliability. To remedy these deficiencies, in this paper we propose to integrate the FBC technique with the effective-capacity theory to support *both* the statistical *delay* and *error-rate* bounded QoS provisioning. We develop and derive the statistical QoS performance metric called: the *DT-based $\epsilon$-effective capacity* to guarantee *both* the statistical *delay* and *error-rate* bounded QoS provisioning for our proposed DT wireless networks through the following definition.

*Definition 2:* For an $(n, W_k, \epsilon_k)$-code, the *DT-based $\epsilon$-effective capacity*, denoted by $EC_k(\theta_k, \epsilon_k, \mathcal{P}_k)$, for the $k$th MU is defined as the maximum DT-data's constant arrival rate for a given service process subject to *both* statistical *delay* and *error-rate* bounded QoS requirements measured by the exponentially decaying rate $\theta_k$ of the delay-bound violation probability and the non-vanishing decoding-error probability $\epsilon_k$, respectively, under the transmit power allocation $\mathcal{P}_k$, which is specified as follows:

$$EC_k(\theta_k, \epsilon_k, \mathcal{P}_k) = -\frac{1}{n\theta_k}\log\left\{\mathbb{E}_{\gamma_k}\left[\epsilon_k\left(\gamma_k(\mathcal{P}_k)\right)\right] + \mathbb{E}_{\gamma_k}\left[1-\epsilon_k\left(\gamma_k(\mathcal{P}_k)\right)\right] e^{-\theta_k \log_2 W_k}\right\} \quad (3)$$

where $\mathbb{E}_{\gamma_k}\{\cdot\}$ denotes the expectation with respect to the random variable $\gamma_k(\mathcal{P}_k)$, and $\gamma_k(\mathcal{P}_k)$ is the signal-to-noise ratio (SNR) of the $k$th MU, which is a function of the constant transmit power $\mathcal{P}_k$. The random SNR $\gamma_k(\mathcal{P}_k)$ is due to the random fading in the wireless channel. In Eq. (3), $\epsilon_k(\gamma_k(\mathcal{P}_k))$ is given by:

$$\epsilon_k(\gamma_k(\mathcal{P}_k)) \approx Q\left(\frac{C(\gamma_k(\mathcal{P}_k)) - \frac{\log_2 W_k}{n}}{\sqrt{V(\gamma_k(\mathcal{P}_k))/n}}\right) \quad (4)$$

where $C(\gamma_k(\mathcal{P}_k))$ and $V(\gamma_k(\mathcal{P}_k))$ are the channel capacity and channel dispersion, respectively, which are given by [13]:

$$\begin{cases} C(\gamma_k(\mathcal{P}_k)) = \log_2(1+\gamma_k(\mathcal{P}_k)), \\ V(\gamma_k(\mathcal{P}_k)) \approx 1 - \dfrac{1}{(1+\gamma_k(\mathcal{P}_k))^2}, \end{cases} \quad (5)$$

and $Q(\cdot)$ is the $Q$-function. ■

### B. Inter-Tier and Intra-Tier Collaborative Hierarchical Caching Mechanisms

We develop both *inter-tier* and *intra-tier* collaborative hierarchical caching mechanisms. Our proposed inter-tier collaborative hierarchical caching mechanism aims at collaboratively caching data content items at cacheable components/devices *across* three different caching tiers, and our proposed intra-tier collaborative hierarchical caching mechanism focuses on collaboratively caching data content items at cacheable components/devices *within* one caching tier. In this paper, we optimize the inter-tier collaborative hierarchical caching to maximize the aggregate DT-based $\epsilon$-effective capacity, which is modeled as the *caching gain*, over all three caching tiers and we optimize the intra-tier collaborative hierarchical caching to maximize the DT-based $\epsilon$-effective capacities, as the caching gains, within Tier 1, Tier 2, and Tier 3, respectively. We define $EC_k^{(\mathrm{r})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, $EC_k^{(\mathrm{b})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, and $EC_k^{(\mathrm{m})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ as the $k$th MU's DT-based $\epsilon$-effective capacity when receiving a DT data content item cached at Tier 1 (routers), Tier 2 (cellular-BS/WiFi-AP), and Tier 3 (mobile devices), respectively.

## III. The DT-Based $\epsilon$-Effective Capacity for MURLLC-Driven 6G M-MIMO Mobile Networks

Since our DT-enabled 6G wireless networks deploy massive antennas on the cellular-BS and multiple antennas on WiFi-APs and MUs, we consider massive-MIMO/MIMO communications between cellular-BS/WiFi-AP and MUs. Assume that there are $M_T$ antennas on the cellular-BS, $M_A$ antennas on each WiFi-AP, and $M_R$ antennas on each MU, where $M_T \gg M_A \approx M_R$.

### A. The DT-Based $\epsilon$-Effective Capacity Under Nakagami-$m$ Fading Channel in Single Antenna Communications

To obtain the $\epsilon$-effective capacity for massive MIMO communications, we first derive its expression in single antenna communications. Assume that the channel fading amplitude $h_k$ of the $k$th MU follows the Nakagami-$m$ distribution, whose average is denoted by $\overline{h}$. Let $N_0$ be the power of additive white Gaussian noise (AWGN) and define $\gamma_k(\mathcal{P}_k) \triangleq h_k^2\mathcal{P}_k/N_0$. The probability density function (PDF) of the $k$th MU's SNR, denoted by $P_Z(\gamma_k)$, is given by [20, Eq. (2.21)]:

$$P_Z(\gamma_k) = \frac{\gamma_k^{m-1}}{\Gamma(m)}\left(\frac{m}{\overline{\gamma}_k(\mathcal{P}_k)}\right)^m \exp\left(-\frac{m}{\overline{\gamma}_k(\mathcal{P}_k)}\gamma_k\right) \quad (6)$$

where $m$ is the fading parameter of the Nakagami-$m$ distribution, $\overline{\gamma}_k(\mathcal{P}_k) \triangleq \overline{h}^2\mathcal{P}_k/N_0$ is the average of $\gamma_k(\mathcal{P}_k)$, $\forall k$, and $\Gamma(\cdot)$ is the gamma function. Employing the above Nakagami-$m$ fading channel model, we have the following theorem.

*Theorem 1:* **If** the fading amplitude $h_k$ of the single antenna wireless channel follows the Nakagami-$m$ distribution, where the PDF of SNR is given by Eq. (6), **then** a closed-form expression for the DT-based $\epsilon$-effective capacity for the $k$th MU under the $(n, W_k, \epsilon_k)$-code using FBC scheme is given by Eq. (7) shown at the bottom of this page.

*Proof:* The proof is provided in Appendix A. ■

### B. The Massive MIMO Channel for Transmitting DT Data

We consider massive MIMO communications if the $k$th MU downloads the DT data from the massive-MIMO BS. Denote by $\mathbf{g}_{k,\alpha} \in \mathbb{C}^{M_R\times 1}$ the downlink channel gain between the $k$th MU

$$EC_k(\theta_k, \epsilon_k, \mathcal{P}_k) = \begin{cases} -\dfrac{1}{n\theta_k}\log\left\{Q\left(\sqrt{n}\left[\log_2\{1+\overline{\gamma}_k(\mathcal{P}_k)\} - \dfrac{\log_2 W_k}{n}\right]\right) + \left[1-Q\left(\sqrt{n}\left[\log_2(1+\overline{\gamma}_k(\mathcal{P}_k)) - \dfrac{\log_2 W_k}{n}\right]\right)\right] W_k^{-\frac{\theta_k}{\log 2}}\right\}, & \text{if } \overline{\gamma}_k(\mathcal{P}_k) \gg 1 \\ -\dfrac{1}{n\theta_k}\log\left\{Q\left(\sqrt{\dfrac{n}{2}}\left[(\log_2 e)(\overline{\gamma}_k(\mathcal{P}_k))^{\frac{1}{2}} - \dfrac{\log_2 W_k}{n}(\overline{\gamma}_k(\mathcal{P}_k))^{-\frac{1}{2}}\right]\right) + \left[1-Q\left(\sqrt{\dfrac{n}{2}}\left[(\log_2 e)(\overline{\gamma}_k(\mathcal{P}_k))^{\frac{1}{2}} - \dfrac{\log_2 W_k}{n}(\overline{\gamma}_k(\mathcal{P}_k))^{-\frac{1}{2}}\right]\right)\right] W_k^{-\frac{\theta_k}{\log 2}}\right\}, & \text{if } 0 < \overline{\gamma}_k(\mathcal{P}_k) < 1 \end{cases} \quad (7)$$

and the $\alpha$th antenna on the massive MIMO BS, where $\mathbb{C}^{M_R\times 1}$ denotes a set of elements each consisting of a complex-valued matrix with $M_R$ rows and one column. Let $l_k$ be the distance between the $k$th MU and the BS assuming that the distance between two antennas on the massive-MIMO-BS is small as compared with the distance between an MU and the BS. We can get the downlink channel gain $\mathbf{g}_{k,\alpha}$ as follows [21, Eq. (2.19)]

$$\mathbf{g}_{k,\alpha} = \sqrt{\beta_k}\mathbf{h}_{k,\alpha} \tag{8}$$

where $\beta_k \approx [\lambda_c/(4\pi l_k)]^2$ is the large-scale fading coefficient, where $\lambda_c$ is the wavelength; and $\mathbf{h}_{k,\alpha} \in \mathbb{C}^{M_R\times 1}$ measures the effect of small-scale fading between all antennas on the $k$th MU and the $\alpha$th antenna on the massive MIMO BS. We consider that each coherence interval is divided into two phases: 1) ***uplink training phase*** to estimate the channel gain information and 2) ***downlink payload DT data transmission phase*** to download the DT data [22], which are detailed, respectively, as follows.

*1) **Uplink Training Phase**:* Denote by $\tau_{\text{ul,p}}$ the number of samples for the uplink pilot signal, where we assume that $\tau_{\text{ul,p}} \geq M_R$. Define $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_{\tau_{\text{ul,p}}}] \in \mathbb{C}^{1\times\tau_{\text{ul,p}}}$ as an orthogonal pilot training sequence satisfying $\|\boldsymbol{\phi}\|^2 = 1$, where $\|\cdot\|$ is the Euclidean norm. The pilot signal sent from the $k$th MU to the massive MIMO BS is denoted by $\mathbf{x}_k^{(\text{p})} = \sqrt{\tau_{\text{ul,p}}}\boldsymbol{\phi} \in \mathbb{C}^{1\times\tau_{\text{ul,p}}}$. In the training phase, we assign $M_R$ orthogonal pilot sequences to $M_R$ antennas of the MU $k$, which are known for both the MU $k$ and the massive MIMO BS. Let $\rho_{\text{ul}}$ be the transmit power over uplink and $\mathbf{W}^{(\text{p})} \in \mathbb{C}^{M_R\times\tau_{\text{ul,p}}}$ be the AWGN matrix, whose elements are independent and identically distributed (i.i.d.), following the complex Gaussian distribution $\mathcal{CN}(0,1)$. The received pilot signal, denoted by $\mathbf{Y}_{k,\alpha}^{(\text{p})} \in \mathbb{C}^{M_R\times\tau_{\text{ul,p}}}$, at the $\alpha$th antenna of the massive MIMO BS, is given by:

$$\mathbf{Y}_{k,\alpha}^{(\text{p})} = \sqrt{\rho_{\text{ul}}}\mathbf{g}_{k,\alpha}\mathbf{x}_k^{(\text{p})} + \mathbf{W}^{(\text{p})} = \sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\mathbf{g}_{k,\alpha}\boldsymbol{\phi} + \mathbf{W}^{(\text{p})}. \tag{9}$$

Applying the de-spreading scheme [21, Section 3.1.2] to the received pilot signal, the massive MIMO BS performs a de-spreading operation by correlating its received signals with the pilot signal. Denote by $\overline{\mathbf{y}}_{k,\alpha}^{(\text{p})} \in \mathbb{C}^{M_R\times 1}$ the received pilot signal at the $\alpha$th antenna of massive MIMO BS after the de-spreading operation, which is given by

$$\overline{\mathbf{y}}_{k,\alpha}^{(\text{p})} = \mathbf{Y}_{k,\alpha}^{(\text{p})}\boldsymbol{\phi}^{\text{H}} = \sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\mathbf{g}_{k,\alpha} + \overline{\mathbf{w}}^{(\text{p})} \tag{10}$$

where $(\cdot)^{\text{H}}$ denotes the Hermitian transpose, $\overline{\mathbf{w}}^{(\text{p})} \triangleq \mathbf{W}^{(\text{p})}\boldsymbol{\phi}^{\text{H}} \in \mathbb{C}^{M_R\times 1}$ is the AWGN after de-spreading, and each element of $\overline{\mathbf{w}}^{(\text{p})}$ follows $\mathcal{CN}(0,1)$. Let $\mathbf{G}_k = [\mathbf{g}_{k,1}, \mathbf{g}_{k,2}, \ldots, \mathbf{g}_{k,M_T}] \in \mathbb{C}^{M_R\times M_T}$ be the channel gain matrix between all antennas on the $k$th MU and all antennas on the massive MIMO BS. Let $\widehat{\mathbf{G}}_k = [\widehat{\mathbf{g}}_{k,1}, \widehat{\mathbf{g}}_{k,2}, \ldots, \widehat{\mathbf{g}}_{k,M_T}] \in \mathbb{C}^{M_R\times M_T}$ be the estimated channel gain matrix, indicating the estimation of $\mathbf{G}_k$. Using the minimum mean-square error (MMSE) estimation, we obtain the estimated channel gain $\widehat{g}_{k,\alpha}^{(q)}$ between the $q$th antenna (with $\forall q \in \{1, \ldots, M_R\}$) on the $k$th MU and the $\alpha$th antenna (with $\forall \alpha \in \{1, \ldots, M_T\}$) on the BS as follows [21, Eq. (3.7)] [23, Eq. (4)]:

$$\widehat{g}_{k,\alpha}^{(q)} = \mathbb{E}\left[g_{k,\alpha}^{(q)} \middle| \overline{y}_{k,\alpha}^{(\text{p},q)}\right] = \frac{\sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\beta_k}{1+\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}\overline{y}_{k,\alpha}^{(\text{p},q)} \tag{11}$$

where $\mathbb{E}[\cdot|\cdot]$ is the conditional expectation, $\widehat{g}_{k,\alpha}^{(q)}$ and $\overline{y}_{k,\alpha}^{(\text{p},q)}$ are the $q$th element of $\widehat{\mathbf{g}}_{k,\alpha} \in \mathbb{C}^{M_T\times 1}$, which is the estimation of $\mathbf{g}_{k,\alpha}$, and $\overline{\mathbf{y}}_{k,\alpha}^{(\text{p})}$, respectively. Plugging each element $\overline{y}_{k,\alpha}^{(\text{p},q)}$ of $\overline{\mathbf{y}}_{k,\alpha}^{(\text{p})}$ given by Eq. (10) into Eq. (11), the channel estimation $\widehat{\mathbf{g}}_{k,\alpha}$ is given by:

$$\widehat{\mathbf{g}}_{k,\alpha} = \frac{\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}{1+\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}\mathbf{g}_{k,\alpha} + \frac{\sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\beta_k}{1+\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}\overline{\mathbf{w}}^{(\text{p})}. \tag{12}$$

*2) **Downlink Payload DT Data Transmission Phase**:* During the downlink payload data transmission phase, the massive MIMO BS treats the channel estimation $\widehat{\mathbf{g}}_{k,\alpha}$ as the true channel to transmit the data packet to the MU $k$. Let $b_k$ be symbol intended to the MU $k$, satisfying $\mathbb{E}[|b_k|^2] = 1$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_{M_T}]^{\text{T}} \in \mathbb{C}^{M_T\times 1}$ be the weighted symbol transmitted from all antennas of the massive MIMO BS, where $(\cdot)^{\text{T}}$ is the transpose. Using the maximum ratio transmission (MRT) precoding [24] as the beamforming scheme to focus the signal of the payload DT data towards the $k$th MU, we can derive each element $x_\alpha, \forall\alpha$, of $\mathbf{x}$, which is the transmit signal on the $\alpha$th antenna of BS, as follows:

$$x_\alpha = \sum_{k=1}^{K}\sqrt{\mathcal{P}_k}\left(\boldsymbol{\eta}_{k,\alpha}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{k,\alpha}^{*}b_k, \tag{13}$$

where $(\cdot)^*$ denotes the conjugate, $\mathcal{P}_k$ is the transmit power allocation for the $k$th MU defined in the text above Eq. (3), $\boldsymbol{\eta}_{k,\alpha} \in \mathbb{R}^{1\times M_R}$ is the power control coefficient vector for the signal from the $\alpha$th antenna of the massive MIMO BS to MU $k$, each element of $\boldsymbol{\eta}_{k,\alpha}$, denoted by $\eta_{k,\alpha}^{(q)}$, satisfies $\eta_{k,\alpha}^{(q)} \in [0,1], \forall q$, and $(\cdot)^{\frac{1}{2}}$ is taking square root for each element of the vector. Considering the interference from the WiFi AP, the received signal at MU $k$, denoted by $\mathbf{y}_k \in \mathbb{C}^{M_R\times 1}$, is given by $\mathbf{y}_k = \mathbf{G}_k\mathbf{x} + \mathbf{G}_k^{(\text{AP})}\mathbf{x}_{\widetilde{k}} + \mathbf{w}_k$, where $\mathbf{G}_k^{(\text{AP})} \in \mathbb{C}^{M_R\times M_A}$ is the channel gain between the WiFi AP and the $k$th MU, $\mathbf{w}_k \in \mathbb{R}^{M_R\times 1}$ is AWGN following $\mathcal{CN}(0,1)$, and $\mathbf{x}_{\widetilde{k}} = [x_{\widetilde{k},1}, \ldots, x_{\widetilde{k},M_A}] \in \mathbb{C}^{M_A\times 1}$ is the signal vector transmitted from each antenna of the WiFi AP to its intended $\widetilde{k}$th MU, $\widetilde{k} \neq k$, where $x_{\widetilde{k},\imath}, \forall\imath \in \{1, \ldots, M_A\}$, is the signal transmitted from the $\imath$th antenna of the WiFi AP. Then, we can derive each element $y_k^{(q)}, \forall q \in \{1, 2, \ldots, M_R\}$, of $\mathbf{y}_k$ as follows:

$$\begin{aligned} y_k^{(q)} &= \mathbf{g}_k^{(q)}\mathbf{x} + \mathbf{g}_k^{(q,\text{AP})}\mathbf{x}_{\widetilde{k}} + w_k^{(q)} \\ &= \underbrace{\sqrt{\mathcal{P}_k}\sum_{\alpha=1}^{M_T} g_{k,\alpha}^{(q)}\left(\boldsymbol{\eta}_{k,\alpha}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{k,\alpha}^{*}b_k}_{\text{desired signal}} \\ &\quad + \underbrace{\sum_{\alpha=1}^{M_T}\sum_{u=1,u\neq k}^{K}\sqrt{\mathcal{P}_u}g_{k,\alpha}^{(q)}\left(\boldsymbol{\eta}_{u,\alpha}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,\alpha}^{*}b_u + w_k^{(q)}}_{\text{effective additive noise } N_k^{(q)}} \\ &\quad + \underbrace{\sum_{\imath=1}^{M_A}\sum_{\widetilde{k}=1,\widetilde{k}\neq k}^{K}\sqrt{\mathcal{P}_{\widetilde{k}}}g_{k,\imath}^{(q,\text{AP})}\left(\boldsymbol{\eta}_{\widetilde{k},\imath}\right)^{\frac{1}{2}}\left[\widehat{\mathbf{g}}_{\widetilde{k},\imath}^{(\text{AP})}\right]^{*}b_{\widetilde{k}}}_{\text{interference from the WiFi AP } I_k^{(q)}} \end{aligned} \tag{14}$$

where $\mathbf{g}_k^{(q)} \in \mathbb{C}^{1\times M_T}$ and $\mathbf{g}_k^{(q,\text{AP})} \in \mathbb{C}^{1\times M_A}$ are the $q$th row of $\mathbf{G}_k$ and $\mathbf{G}_k^{(\text{AP})}$, respectively; $w_k^{(q)}$ is the $q$th element of $\mathbf{w}_k$,

representing AWGN on the $q$th antenna of the $k$th MU; $b_u$ and $b_{\widetilde{k}}$ are symbols transmitted from the BS intended to the $u$th MU, $u \neq k$, and from the WiFi AP intended to the $\widetilde{k}$th MU, respectively; $N_k^{(q)}$ is the *effective additive noise* of the $k$th MU on its $q$th antenna; $g_{k,\imath}^{(q,\mathrm{AP})}$ is the $\imath$th element of $\mathbf{g}_k^{(q,\mathrm{AP})}$, representing the channel gain between the $q$th antenna of the $k$th MU and the $\imath$th antenna of the WiFi AP; $\boldsymbol{\eta}_{\widetilde{k},\imath} \in \mathbb{R}^{1\times M_R}$ is the power control coefficient vector for the signal from the $\imath$th antenna of the WiFi AP to MU $\widetilde{k}$; $[\widehat{\mathbf{g}}_{\widetilde{k},\imath}^{(\mathrm{AP})}]^* \in \mathbb{C}^{M_R\times 1}$ is the conjugate of the WiFi AP's channel estimation vector between the $\imath$th antenna on WiFi AP and all antennas on the $\widetilde{k}$th MU; and $I_k^{(q)}$ is the interference from the WiFi AP to the $q$th antenna on the $k$th MU. Note that since we use the MRT precoding as the beamforming scheme to transmit the signal, the inter-user interference $\sum_{\alpha=1}^{M_T}\sum_{u=1,u\neq k}^{K}\sqrt{\mathcal{P}_u}g_{k,\alpha}^{(q)}(\boldsymbol{\eta}_{u,\alpha})^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,\alpha}^{*}b_u$ in Eq. (14) is negligible and is treated as the noise. Moreover, the directions of channel gain gradually become orthogonal as $M_A$ increases under the channel property *approximately favorable propagation* [21, Eq. (7.2)] such that

$$\sum\nolimits_{\imath=1}^{M_A} g_{k,\imath}^{(q,\mathrm{AP})}\left(\boldsymbol{\eta}_{\widetilde{k},\imath}\right)^{\frac{1}{2}}\left[\widehat{\mathbf{g}}_{\widetilde{k},\imath}^{(\mathrm{AP})}\right]^{*} \to 0, \text{ if } \widetilde{k}\neq k. \tag{15}$$

In this paper, we assume that the number $M_A$ of antennas on the WiFi AP is large enough to make Eq. (15) hold true. Thus, we assume that the interference $I_k^{(q)}$ from WiFi AP to the $q$th antenna on the $k$th MU is practically negligible.

Using Eq. (14), the SNR on the $q$th antenna of the $k$th MU, denoted by $\gamma_k^{(q)}(\mathcal{P}_k), \forall q$, is given by [21]:

$$\gamma_k^{(q)}(\mathcal{P}_k) = \frac{\mathrm{Var}\left[\sqrt{\mathcal{P}_k}b_k\sum_{\alpha=1}^{M_T}g_{k,\alpha}^{(q)}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\alpha}^{(i)}}\left(\widehat{g}_{k,\alpha}^{(i)}\right)^{*}\right]}{\mathrm{Var}\left[\sum_{\alpha=1}^{M_T}\sum_{u=1,u\neq k}^{K}\sqrt{\mathcal{P}_u}g_{k,\alpha}^{(q)}\left(\boldsymbol{\eta}_{u,\alpha}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,\alpha}^{*}b_u\right]+1} \tag{16}$$

where $\mathrm{Var}[\cdot]$ is derived with respect to both the random distance $l_k$ and the random small-scale fading $h_{k,\alpha}^{(q)}$, which is the $q$th element of $\mathbf{h}_{k,\alpha}$ defined in the text following Eq. (8), for the wireless channel between the $q$th antenna on MU $k$ and the $\alpha$th antenna of the massive MIMO BS. We assume that all $h_{k,\alpha}^{(q)}, \forall q, \alpha$, are i.i.d., following the Nakagami-$m$ distribution. Then, using Eq. (16), we obtain the following theorem.

*Theorem 2:* **If** all small-scale fading $h_{k,\alpha}^{(q)}, \forall q, \alpha$, of wireless channels for massive MIMO communications are i.i.d., following the Nakagami-$m$ distribution, and assume that all MUs are uniformly distributed within a wireless cell with the inner radius $R_{\min}$ and the outer radius $R_{\max}$, **then** the closed-form expression for the SNR $\gamma_k^{(q)}(\mathcal{P}_k)$ on the $q$th antenna of the $k$th MU is specified as follows:

$$\gamma_k^{(q)}(\mathcal{P}_k) = \frac{\mathcal{P}_k N_{k,1}(M_T)}{\left(\sum_{u=1,u\neq k}\mathcal{P}_u\right)N_{k,2}(M_T)+1}, \ \forall q \tag{17}$$

where $\forall q \in \{1,2,\ldots,M_R\}$ and

$$N_{k,1}(M_T) = \frac{\beta_k^3\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\overline{\eta}M_TM_R\overline{h}^2}{(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_k)^2}\left(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_kM_TM_R\overline{h}^2\right) \tag{18}$$

and

$$\begin{aligned}N_{k,2}(M_T) =& \frac{\beta_k\overline{\eta}M_TM_R\overline{h}^2\lambda_c^2}{(4\pi)^2(R_{\max}^2-R_{\min}^2)}\left\{\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\left(\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\overline{h}^2+1\right)(X_{\max}-X_{\min})\right.\\ &\left.+\frac{M_T(M_R-1)\lambda_c^2}{(4\pi)^2(R_{\max}^2-R_{\min}^2)}\left[\log\left(\frac{X_{\max}}{X_{\min}}\right)\right]^2\right\}\end{aligned} \tag{19}$$

where $\overline{\eta} \triangleq \mathbb{E}[\eta_{k,\alpha}^{(i)}]$, $\overline{h}$ is defined in the text above Eq. (6), $\lambda_c$ is defined in the text following Eq. (8), and

$$\begin{cases}X_{\max} = \dfrac{\lambda_c^2}{16\pi^2(R_{\min}^2+\imath^2)+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\lambda_c^2},\\ X_{\min} = \dfrac{\lambda_c^2}{16\pi^2(R_{\max}^2+\imath^2)+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\lambda_c^2},\end{cases} \tag{20}$$

where $\imath$ denotes the height of a BS/AP.

*Proof:* The proof is provided in [25, Eqs. (14)–(28)]. ■

*Remarks on Theorem 2:* Theorem 2 reveals that all $\gamma_k^{(q)}(\mathcal{P}_k), \forall q$, are the same, since random variables $h_{k,\alpha}^{(q)}, \forall q, \alpha$ are i.i.d.

### C. The $\epsilon$-Effective Capacity for Transmitting DT Data Cached At Massive MIMO BS and WiFi AP

If a DT data is cached at the massive MIMO BS in Tier 2 and the $k$th MU downloads this data using massive MIMO channels, using Theorems 1 and 2, we can derive the $\epsilon$-effective capacity for the $k$th MU in the following theorem.

*Theorem 3:* **If** the $k$th MU receives the cached DT data from the massive MIMO BS and all small-scale fading $h_{k,\alpha}^{(q)}, \forall q, \alpha$, of wireless channels for massive MIMO communications are i.i.d. following the Nakagami-$m$ distribution, **then** a closed-form expression for the $\epsilon$-effective capacity, denoted by $EC_k^{\mathrm{BS}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{BS}})$, for the $k$th MU under the $(n, W_k, \epsilon_k)$-coding scheme in the non-asymptotic regime is determined by:

$$EC_k^{\mathrm{BS}}(\theta_k,\epsilon_k,\mathcal{P}_{\mathrm{BS}}) \triangleq EC_k^{\mathrm{MIMO}}(\theta_k,\epsilon_k,\mathcal{P}_{\mathrm{BS}}) \tag{21}$$

$$= \begin{cases}-\dfrac{1}{n\theta_k}\log\left\{Q_1+(1-Q_1)W_k^{-\frac{\theta_k}{\log 2}}\right\}, & \text{if } \overline{\gamma}_k(\mathcal{P}_{\mathrm{BS}})\gg 1\\ -\dfrac{1}{n\theta_k}\log\left\{Q_2+(1-Q_2)W_k^{-\frac{\theta_k}{\log 2}}\right\}, & \text{if } 0<\overline{\gamma}_k(\mathcal{P}_{\mathrm{BS}})<1\end{cases} \tag{22}$$

where $EC_k^{\mathrm{MIMO}}(\cdot,\cdot,\cdot)$ denotes the general function for the $\epsilon$-effective capacity over the massive-MIMO channel to the $k$th MU, $\overline{\gamma}_k(\mathcal{P}_{\mathrm{BS}}) \triangleq \gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}}), \forall q$, is given by Eq. (17) after replacing $\mathcal{P}_k$ by $\mathcal{P}_{\mathrm{BS}}$, and we define $Q_1$ and $Q_2$ as follows:

$$\begin{cases}Q_1 \triangleq \left[Q\left(\sqrt{n}\left[\log_2\{1+\overline{\gamma}_k(\mathcal{P}_{\mathrm{BS}})\}-\dfrac{\log_2 W_k}{n}\right]\right)\right]^{M_R},\\ Q_2 \triangleq \left[Q\left(\sqrt{\frac{n}{2}}\left[(\log_2 e)[\overline{\gamma}_k(\mathcal{P}_{\mathrm{BS}})]^{\frac{1}{2}}-\frac{\log_2 W_k}{n}[\overline{\gamma}_k(\mathcal{P}_{\mathrm{BS}})]^{-\frac{1}{2}}\right]\right)\right]^{M_R}.\end{cases} \tag{23}$$

*Proof:* The proof is provided in Appendix B. ■

In addition to downloading the cached DT data from a massive MIMO BS using the massive MIMO communications, an MU can also download a cached DT data from a multi-antenna-equipped WiFi AP in Tier 2 using MIMO communications.

Then, we have the following theorem if the $k$th MU receives the DT data from the WiFi AP.

*Theorem 4:* **If** the $k$th MU receives the cached DT data from the WiFi AP equipped with $M_A$ antennas with $M_A \approx M_R$ and all small-scale fading $h_{k,\alpha}^{(q)}, \forall q, \alpha$, of wireless channels for MIMO communications are i.i.d. following the Nakagami-$m$ distribution, **then** a closed-form expression for the DT-based $\epsilon$-effective capacity, denoted by $EC_k^{\text{AP}}(\theta_k, \epsilon_k, \mathcal{P}_{\text{AP}})$, for the $k$th MU under the $(n, W_k, \epsilon_k)$-coding scheme in the non-asymptotic regime is determined by:

$$EC_k^{\text{AP}}(\theta_k, \epsilon_k, \mathcal{P}_{\text{AP}}) \triangleq EC_k^{\text{MIMO}}(\theta_k, \epsilon_k, \mathcal{P}_{\text{AP}}) \tag{24}$$

$$= \begin{cases} -\frac{1}{n\theta_k}\log\left\{Q_3 + (1-Q_3)W_k^{-\frac{\theta_k}{\log 2}}\right\}, & \text{if } \overline{\gamma}_k(\mathcal{P}_{\text{AP}}) \gg 1 \\ -\frac{1}{n\theta_k}\log\left\{Q_4 + (1-Q_4)W_k^{-\frac{\theta_k}{\log 2}}\right\}, & \text{if } 0 < \overline{\gamma}_k(\mathcal{P}_{\text{AP}}) < 1 \end{cases} \tag{25}$$

where $EC_k^{\text{MIMO}}(\cdot,\cdot,\cdot)$ given in Eq. (24) is described in the text following Eq. (21) and

$$\overline{\gamma}_k(\mathcal{P}_{\text{AP}}) \triangleq \gamma_k^{(q)}(\mathcal{P}_{\text{AP}}) = \frac{\mathcal{P}_{\text{AP}} N_{k,1}(M)}{\left(\sum_{u=1,u\neq k} \mathcal{P}_u\right) N_{k,2}(M) + 1}, \tag{26}$$

where $M \triangleq \min\{M_R, M_A\}$ is taking the minimum between $M_A$ and $M_R$, after replacing $M_T$ in Eq. (18) and Eq. (19) by $M$ and replacing $\mathcal{P}_k$ by $\mathcal{P}_{\text{AP}}$ in Eq. (17), respectively; and we define two new auxiliary variables $Q_3$ and $Q_4$ as follows:

$$\begin{cases} Q_3 \triangleq \left[Q\left(\sqrt{n}\left[\log_2\{1+\overline{\gamma}_k(\mathcal{P}_{\text{AP}})\} - \frac{\log_2 W_k}{n}\right]\right)\right]^M, \\ Q_4 \triangleq \\ \left[Q\left(\sqrt{\frac{n}{2}}\left[(\log_2 e)[\overline{\gamma}_k(\mathcal{P}_{\text{AP}})]^{\frac{1}{2}} - \frac{\log_2 W_k}{n}[\overline{\gamma}_k(\mathcal{P}_{\text{AP}})]^{-\frac{1}{2}}\right]\right)\right]^M. \end{cases} \tag{27}$$

*Proof:* The proof is omitted due to lack of space, but can be obtained in the way similar to the derivations given in Appendix B. ■

## IV. Inter-Tier Collaborative Hierarchical Caching Mechanisms

Denote by $S_{\text{r}}, S_{\text{b}}$, and $S_{\text{m}}$ the total caching capability (i.e., maximum number of cached data content items assuming that the sizes of all data content items are the same) for caching Tier 1 (routers), caching Tier 2 (cellular-BS/WiFi-AP), and caching Tier 3 (mobile devices), respectively, and denote by $D$ the total number of DT data content items. Also denote by $f_d(\triangle t)$ the *instantaneous popularity* of the $d$th data content item, $\forall d \in \{1, 2, \ldots, D\}$, representing the probability that the $d$th data is requested by at least one of MUs during the last period $\triangle t$ from the current observing time. Note that the instantaneous value of $f_d(\triangle t)$ is a function of the observation time and can increase or decrease depending on whether the $d$th data content item is getting more or less visits during the last $\triangle t$ period from the current observing time. For the inter-tier collaborative hierarchical caching algorithm, we propose that a popular DT data content item is randomly cached at a caching tier if and only if its popularity $f_d(\triangle t)$ exceeds a predefined threshold for that tier. Denoted by $f^{(\text{T1})}$, $f^{(\text{T2})}$, and $f^{(\text{T3})}$ the popularity thresholds for Tier 1, Tier 2, and Tier 3, respectively, and we set $f^{(\text{T1})} < f^{(\text{T2})} < f^{(\text{T3})}$. This ensures that we cache more popular data content items at cache stations closer to MUs. If a data content item's popularity is larger than the threshold of a tier, this data content item enters one of the cache stations of routers (i.e., Tier 1), cellular-BS/WiFi-AP (i.e., Tier 2), or mobile devices (i.e., Tier 3) following a Poisson process with the arrival rate of $\lambda$.

According to each tier's caching capability and caching expense, the data content items stored at different tiers have different *caching lifespans*, and the cache stations delete their cached data content items after this lifespan period ends for replacing the old data content item by a new one. Thus, we propose that a DT data content item departs its caching tiers once its caching lifespan elapses or its popularity $f_d(\triangle t)$ drops below the threshold of its current caching tier. We define random variables $T_{\text{r}}$, $T_{\text{b}}$, and $T_{\text{m}}$ to be the popular data content items' lifespans for caching in the tiers of routers, cellular-BS/WiFi-AP, and mobile devices, respectively. We propose that data content items depart the their caches by following the exponential distribution with departure rates $\mu_{\text{r}}$, $\mu_{\text{b}}$, and $\mu_{\text{m}}$ ($\mu_{\text{r}} \neq \mu_{\text{b}} \neq \mu_{\text{m}}$), when they are cached in routers, cellular-BS/WiFi-AP, and mobile devices, respectively.

Each tier's caching expense under the lifespan of $T_i, \forall i \in \{\text{r}, \text{b}, \text{m}\}$, is denoted by $\xi(T_i)$, which is defined by Eq. (1). Let $\pi_{\text{r}}(\lambda, \mu_{\text{r}})$, $\pi_{\text{b}}(\lambda, \mu_{\text{b}})$, and $\pi_{\text{m}}(\lambda, \mu_{\text{m}})$ be the probabilities of a data content item to be cached at Tier 1, Tier 2, and Tier 3, respectively. The objective of our inter-tier collaborative hierarchical caching mechanism is to optimize the cached DT-data items' departure rates $\mu_{\text{r}}$, $\mu_{\text{b}}$, and $\mu_{\text{m}}$ and the other cache-controlling variables such that our caching mechanism can maximize the aggregate $\epsilon$-effective capacity over three caching tiers for delivering all MUs' requested DT-data items under the caching expense constraint $\xi^{\max}$ and caching capability, which can be formulated as the following optimization problem:

$$\max_{\substack{\mu_{\text{r}},\mu_{\text{b}},\mu_{\text{m}},l,\pi_{(\text{m}),d},\\ \pi_{\text{b}}^{(\text{BS})},\pi_{\text{b}}^{(\text{AP})},\widetilde{\pi}_{\text{r}}^{(\text{BS})},\widetilde{\pi}_{\text{r}}^{(\text{AP})}}} \left\{\sum_{i=\{\text{r},\text{b},\text{m}\}}\left[\sum_{k=1}^{K}\pi_i(\lambda,\mu_i)EC_k^{(i)}(\theta_k,\epsilon_k,\mathcal{P}_k)\right]\right\} \tag{28}$$

$$\text{s.t.: C1: } \pi_{\text{r}}(\lambda,\mu_{\text{r}})\mathbb{E}[\xi(T_{\text{r}})] + \pi_{\text{b}}(\lambda,\mu_{\text{b}})\mathbb{E}[\xi(T_{\text{b}})] + \pi_{\text{m}}(\lambda,\mu_{\text{m}})\mathbb{E}[\xi(T_{\text{m}})] \leq \xi^{\max}, \tag{29}$$

$$\text{C2: } \pi_{\text{r}}(\lambda,\mu_{\text{r}}) + \pi_{\text{b}}(\lambda,\mu_{\text{b}}) + \pi_{\text{m}}(\lambda,\mu_{\text{m}}) = 1, \tag{30}$$

$$\text{C3: } \pi_i(\lambda,\mu_i)D \leq S_i,\ \forall i \in \{\text{r},\text{b},\text{m}\}, \tag{31}$$

$$\text{C4: } l \leq 1,\ \pi_{(\text{m}),d} \leq 1, \tag{32}$$

$$\text{C5: } \pi_{\text{b}}^{(\text{BS})} + \pi_{\text{b}}^{(\text{AP})} = \pi_{\text{b}}(\lambda,\mu_{\text{b}}), \tag{33}$$

$$\text{C6: } \widetilde{\pi}_{\text{r}}^{(\text{BS})} + \widetilde{\pi}_{\text{r}}^{(\text{AP})} = \pi_{\text{r}}(\lambda,\mu_{\text{r}}), \tag{34}$$

where $EC_k^{(i)}(\theta_k, \epsilon_k, \mathcal{P}_k), \forall i \in \{\text{r}, \text{b}, \text{m}\}$, is defined in Section II-B and will be further detailed in the following Sections V-A, V-B and V-C, respectively; $l$ is the MUs' device-to-device (D2D) communication range for Tier 3, $\pi_{(\text{m}),d}$ is the probability that the $d$th data content item is cached at a cache unit of a mobile device at Tier 3; $\pi_{\text{b}}^{(\text{BS})}$ and $\pi_{\text{b}}^{(\text{AP})}$ are probabilities that a data content item is cached at the cellular BS and WiFi AP at Tier 2, respectively; and $\widetilde{\pi}_{\text{r}}^{(\text{BS})}$ and $\widetilde{\pi}_{\text{r}}^{(\text{AP})}$ are probabilities that a data content item is cached at a router in Tier 1 to be transmitted going through the cellular BS and WiFi AP, respectively. Using the optimal $\mu_{\text{r}}$, $\mu_{\text{b}}$, and $\mu_{\text{m}}$, we

can obtain $\mathbb{E}[T_i], \forall i$, representing the collaborative caching algorithm on how long to cache a DT data item at each caching tier, and the probability $\pi_i(\lambda, \mu_i), \forall i$, representing the caching algorithm on where to cache a DT data item.

## V. Intra-Tier Collaborative Caching Mechanisms for Each Caching Tier Over DT-Enabled 6G MURLLC Mobile Wireless Networks

We consider the intra-tier collaborative caching mechanism within each caching tier for Tier 1, Tier 2, and Tier 3, respectively, as follows.

### A. Intra-Tier Collaborative Caching Mechanism for Caching Tier 3 At Mobile Users

When the popularity $f_d(\triangle t)$ of a DT data content item is larger than the popularity threshold of Tier 3 (i.e., $f_d(\triangle t) > f^{(\mathrm{T3})}$), we propose to cache it at a mobile device at Tier 3 and MUs will download their requested data content items by using D2D communications. Define the mobile device's D2D communication range as $l$ [26], [27], implying that an MU can communicate with another mobile device within a distance $l$. Let $\mathcal{M}_k$ be the D2D communication set of the $k$th MU, such that all elements in this set are MUs within the distance $l$ to the $k$th MU. The density of MUs in this wireless cell is defined as $k_0 \triangleq K/(\pi R_{\max}^2)$, where $R_{\max}$ is the outer radius of the wireless cell. Denote by $|\mathcal{M}_k|$ the cardinality of the set $\mathcal{M}_k$. The probability that there are $|\mathcal{M}_k|$ MUs around the $k$th MU within the distance $l$ is denoted by $p(|\mathcal{M}_k|, \pi l^2, k_0)$, where $p(x, y, z) \triangleq [(yz)^x e^{-yz}]/(x!)$ is the probability mass function of a Poisson Point Process, representing the probability that there are $x$ MUs in the area $y$ with user density equal to $z$. In our proposed intra-tier collaborative caching scheme, the $k$th MU, $\forall k$, and all mobile devices in its D2D communication set $\mathcal{M}_k$ update and exchange their caching information (which data content items they are currently caching) with each other. The *cache hitting rate (probability)*, denoted by $P_c(k)$, of a DT data content item requested by MU $k$ can be defined as the probability that the requested DT data content item of MU $k$ can be found within $\mathcal{M}_k$. Denote by $\pi_{(\mathrm{m}),d}$ the probability that the $d$th data content item is cached at a cache unit of a mobile device at Tier 3, where we use the "cache unit" to represent the memory space to cache one data content item. Let $c_j \in \mathbb{N}$ be the $j$th mobile device's *caching capability* (i.e., total number of cache units), $\forall j \in \mathcal{M}_k$, representing the maximum number of data content items that can be cached in the mobile device $j$. Let $\{X_d\}$, $\forall d \in \{1, \ldots, D\}$, be a set of $D$ independent random variables, and let $X$ be the sum of $X_d, \forall d$. We set $X_d \in \{0, 1\}$, where $\{X_d = 1\}$ represents that the $d$th data content item is requested by at least one MU during the last period $\triangle t$, while $\{X_d = 0\}$ represents that the $d$th data content item is not requested by any MU during the last period $\triangle t$. Thus, we have $f_d(\triangle t) = \mathbb{E}[X_d]$.

We can derive $P_c(k)$ as a function of $l$ and $\pi_{(\mathrm{m}),d}$ as follows:

$$P_c(k) = \sum_{d=1}^{D} \Bigg\{ f_d(\triangle t)\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}}) \times \sum_{\kappa=0}^{K-1} \left[ p\big(\kappa, \pi l^2, k_0\big) \left( 1 - \prod_{j=1}^{\kappa} \big(1 - \pi_{(\mathrm{m}),d}\big)^{c_j} \right) \right] \Bigg\}. \quad (35)$$

Using Eq. (35), the objective of the intra-tier collaborative caching mechanism for Tier 3 is for maximizing the $k$th MU's $\epsilon$-effective capacity, $\forall k$, for receiving its requested DT data by using D2D communications as follows:

$$\max_{l, \pi_{(\mathrm{m}),d}} \Big\{ EC_k^{(\mathrm{m})}(\theta_k, \epsilon_k, \mathcal{P}_k) \Big\} = \max_{l, \pi_{(\mathrm{m}),d}} \big\{ P_c(k) EC_k^{\mathrm{D2D}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{D2D}}) \big\} \quad (36)$$

$$\text{s.t.: C3} : \pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}}) D \le S_{\mathrm{m}}$$
$$\text{C4} : l \le 1, \ \pi_{(\mathrm{m}),d} \le 1,$$

where $EC_k^{\mathrm{D2D}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{D2D}})$ is given by:

$$EC_k^{\mathrm{D2D}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{D2D}}) \triangleq EC_k^{\mathrm{MIMO}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{D2D}}) \quad (37)$$

$$= \begin{cases} -\dfrac{1}{n\theta_k} \log \left\{ Q_5 + (1 - Q_5) W_k^{-\frac{\theta_k}{\log 2}} \right\}, & \text{if } \overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}}) \gg 1 \\ -\dfrac{1}{n\theta_k} \log \left\{ Q_6 + (1 - Q_6) W_k^{-\frac{\theta_k}{\log 2}} \right\}, & \text{if } 0 < \overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}}) < 1 \end{cases} \quad (38)$$

where $EC_k^{\mathrm{MIMO}}(\cdot, \cdot, \cdot)$ given in Eq. (37) is described in Eq. (21), $\overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}})$ can be derived by using the way similar to the derivations for Eq. (26) as follows:

$$\overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}}) \triangleq \gamma_k^{(q)}(\mathcal{P}_{\mathrm{D2D}}) = \frac{\mathcal{P}_{\mathrm{D2D}} N_{k,1}(M_R)}{\left( \sum_{u=1, u \ne k} \mathcal{P}_u \right) N_{k,2}(M_R) + 1}, \quad (39)$$

and we define two new auxiliary variables $Q_5$ and $Q_6$ as follows:

$$\begin{cases} Q_5 \triangleq \left[ Q\left( \sqrt{n} \left[ \log_2\{1 + \overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}})\} - \dfrac{\log_2 W_k}{n} \right] \right) \right]^{M_R}, \\ Q_6 \triangleq \left[ Q\left( \sqrt{\dfrac{n}{2}} \left[ (\log_2 e)[\overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}})]^{\frac{1}{2}} - \dfrac{\log_2 W_k}{n} [\overline{\gamma}_k(\mathcal{P}_{\mathrm{D2D}})]^{-\frac{1}{2}} \right] \right) \right]^{M_R}. \end{cases} \quad (40)$$

In Eq. (36), we use the D2D communication distance $l$ as the maximizer because $l$ determines the elements in $\mathcal{M}_k$. Thus, to solve Eq. (36), we need to further derive a closed-form expression of $P_c(k)$ as a function of $l$ and $\pi_{(\mathrm{m}),d}$.

To obtain the maximum $P_c(k)$, we need to derive the algorithm that determines which data content items are cached on the mobile devices. Define $P_d^{(\mathrm{m})}(j), 1 \le j \le |\mathcal{M}_k|$, as the $d$th data content item's caching probability in prior $(j-1)$ mobile devices in the set $\mathcal{M}_k$ and define $P_d^{(\mathrm{m})}(0) = 0$ [28]. Assume that all MUs in the set $\mathcal{M}_k$ are ordered in which they entered this set, and they know their own orders. For example, if $j = 4$ and the $d$th data content item has been cached in 2 mobile devices of the prior 3 mobile devices, then we have $P_d^{(\mathrm{m})}(4) = 2/3$. To maximize the cache hitting rate of Tier 3, we apply the **Algorithm 1** given at the top of the next page. We observe from **Algorithm 1** that $\pi_{(\mathrm{m}),d}$ is proportional to the value of $s_d$, which is defined in **Algorithm 1**. We can observe that, the goal of **Algorithm 1** is caching data content items that have the largest request-cache probability difference.

To obtain a more precise expression for Eq. (36), we further study the lower-bound and upper-bound, respectively, on the cache hitting rate (probability), denoted by $P_c(k)$, as follows.

*1) **The Lower-Bound on Cache Hitting Rate**:* We define $\kappa \triangleq |\mathcal{M}_k|$ as the cardinality of the set $\mathcal{M}_k$, indicating the number of MUs that can communicate with the MU $k$ using D2D. According to Eq. (35), we can derive the lower-bound on

**Algorithm 1:** Cache Placement Algorithm to Maximize the Cache Hitting Rate on Tier 3.

1: **for** each mobile device $j$ in the set $\mathcal{M}_k$ **do**
2: Calculate the difference between data popularity $f_d(\triangle t)$ and caching probability $P_d^{(\mathrm{m})}(j)$ for each data content item, denoted by $s_d$, where $s_d \triangleq f_d(\triangle t) - P_d^{(\mathrm{m})}(j), \forall d$.
3: Cache the first $c_j$ data content items, which have the maximum values of $s_d$, into the mobile device $j$.
4: **end for**

the cache hitting rate $P_c(k)$ as follows:

$$P_c(k) \overset{(a)}{\geq} \sum_{d=1}^{D} f_d(\triangle t)\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d \tag{41}$$

where

$$\Lambda_d \triangleq \sum_{\kappa=0}^{K-1}\left[p\left(\kappa,\pi l^2,k_0\right)\prod_{j=1}^{\kappa}\pi_{(\mathrm{m}),d}^{c_j}\right], \tag{42}$$

and (a) is due to $(1-\prod_i x^i) \geq \prod_i (1-x)^i, \forall x \in (0,1)$. Define $\tau$ as the expectation of $X$, which is the popularity for all data content items during the last period $\triangle t$ and is derived as follows:

$$\tau = \mathbb{E}[X] = \mathbb{E}\left[\sum_{d=1}^{D} X_d\right] = \sum_{d=1}^{D}\mathbb{E}\left[X_d\right] = \sum_{d=1}^{D} f_d(\triangle t). \tag{43}$$

For any $0<\delta<1$, we define a range $\mathcal{R} = [(1-\delta)\tau, (1+\delta)\tau]$ around $\tau$. Applying the Chernoff bound, we can derive the probability that $X$ deviates from $\tau$ as follows:

$$\begin{cases} \Pr\{X > (1+\delta)\tau\} \leq e^{-\frac{\delta^2\tau}{3}}, \\ \Pr\{X < (1-\delta)\tau\} \leq e^{-\frac{\delta^2\tau}{2}}. \end{cases} \tag{44}$$

Using Eq. (44), we can further derive Eq. (41) as follows:

$$\begin{aligned} P_c(k) &\geq \sum_{d=1}^{D}\{f_d(\triangle t)\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d\} \geq \Pr\{X\in\mathcal{R}\}\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d \\ &= \Pr\left\{(1-\delta)\tau \leq X \leq (1+\delta)\tau\right\}\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d \\ &\overset{(b)}{\geq} \left(1-e^{-\frac{\delta^2\tau}{3}}\right)\left(1-e^{-\frac{\delta^2\tau}{2}}\right)\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d \\ &\overset{(c)}{\geq} \left(1-2e^{-\frac{\delta^2\tau}{3}}\right)\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d, \end{aligned} \tag{45}$$

where (b) follows by using Eq. (44), and (c) holds when the popularity $f_d(\triangle t)$ and $\Lambda_d$ satisfy

$$d^* = \arg\min_{d}\left\{f_d(\triangle t)\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Lambda_d\right\}. \tag{46}$$

Similar to the derivations for Eq. (43), let $\{Y_\kappa\}, \forall \kappa \in \{0,1,\ldots,K-1\}$, be a set consisting of $K$ independent random variables, and let $Y$ be the sum of all $Y_\kappa$'s, $\forall\kappa$. We set $Y_\kappa \in \{0,1\}, \forall\kappa$, where $\{Y_\kappa = 1\}$ represents that there are $\kappa$ mobile devices in the D2D communication range of an MU, while $\{Y_\kappa = 0\}$ indicating otherwise. Thus, we have $\mathbb{E}[Y_\kappa] = p(\kappa,\pi l^2,k_0)$. Let $\nu$ be the expectation of $Y$ and thus we can derive $\nu$ as follows:

$$\nu = \mathbb{E}[Y] = \mathbb{E}\left[\sum_{\kappa=0}^{K-1} Y_\kappa\right] = \sum_{\kappa=0}^{K-1}\mathbb{E}[Y_\kappa] = \sum_{\kappa=0}^{K-1} p\left(\kappa,\pi l^2,k_0\right). \tag{47}$$

Defining the range $\widetilde{\mathcal{R}} = [(1-\delta)\nu, (1+\delta)\nu]$ and applying the Chernoff bound again in Eq. (45), we can derive $P_c(k)$ as follows:

$$\begin{aligned} P_c(k) &\geq \left(1-2e^{-\frac{\delta^2\tau}{3}}\right)\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\Pr\left\{Y\in\widetilde{\mathcal{R}}\right\}\prod_{j=1}^{\kappa^*}\pi_{(\mathrm{m}),d^*}^{c_j} \\ &\geq \pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\left(1-2e^{-\frac{\delta^2\tau}{3}}\right)\left(1-2e^{-\frac{\delta^2\nu}{3}}\right)\prod_{j=1}^{\kappa^*}\pi_{(\mathrm{m}),d^*}^{c_j} \end{aligned} \tag{48}$$

where $\kappa^* \in \widetilde{R}$ is the optimal number of MUs in the D2D communication range of MU $k$ such that

$$\kappa^* = \arg\min_{\kappa\in\widetilde{R}}\left\{p\left(\kappa,\pi l^2,k_0\right)\prod_{j=1}^{\kappa}\pi_{(\mathrm{m}),d^*}^{c_j}\right\}. \tag{49}$$

Since there are totally $K$ users in the entire wireless cell with area of $\pi R_{\max}^2$, each user occupies the average area $(\pi R_{\max}^2)/K$. Then, since the $k$th MU's communication area, namely $\pi l^2$, is proportional to $(\pi R_{\max}^2)/K$, we can obtain that $l$ is proportional to $\sqrt{1/K}$, which can be represented by $l = \Theta(\sqrt{1/K})$. We define that for functions $\varphi_1(x)$ and $\varphi_2(x)$, if $\varphi_1(x) = \Theta(\varphi_2(x))$, we have $\varrho_1\varphi_2(x) \leq |\varphi_1(x)| \leq \varrho_2\varphi_2(x)$, where $\varrho_1$ and $\varrho_2$ are constants. We can observe from Eq. (47) that $\nu$ is proportional to $Kl^2 = \Theta(1)$. Since $\kappa^* \in \widetilde{\mathcal{R}}$, we have

$$\kappa^* \geq (1-\delta)\nu \geq \varrho_2(1-\delta). \tag{50}$$

We denote by $j^*$ the optimal $j$ which yields the minimum value of $\pi_{(\mathrm{m}),d}^{c_j}$ and define $j^* \in [\kappa^*\pi_{(\mathrm{m}),1}(1-\delta), \kappa^*\pi_{(\mathrm{m}),1}(1+\delta)]$. Then, to satisfy $(\kappa^* - j^*) \geq 1$, we need to guarantee

$$\kappa^* - j^* \geq \kappa^* - \kappa^*\pi_{(\mathrm{m}),1}(1+\delta) \geq 1 \Rightarrow \kappa^* \geq \frac{1}{1-\pi_{(\mathrm{m}),1}(1+\delta)}, \tag{51}$$

where, to satisfy Eq. (51), we must have

$$\varrho_2(1-\delta) \geq \frac{1}{1-\pi_{(\mathrm{m}),1}(1+\delta)} \Rightarrow \varrho_2 \geq \frac{1}{1-\delta-\pi_{(\mathrm{m}),1}(1-\delta^2)}. \tag{52}$$

Substituting Eq. (51) into Eq. (48), we can obtain the more precise lower-bound of $P_c(k)$.

*2) **The Upper-Bound on the Cache Hitting Rate**:* Applying the Zipf distribution to characterize the popularity of DT data and defining $\eta$ as the exponent for the Zipf distribution, we derive the upper-bound on $P_c(k)$ as follows:

$$P_c(k) \overset{(d)}{\leq} \pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})\sum_{\kappa=0}^{K-1}\frac{2\kappa^{2-\eta}}{D^{1-\eta}}p\left(\kappa,\pi l^2,k_0\right)\Phi_d(\kappa) \tag{53}$$

where

$$\Phi_d(\kappa) \triangleq 1 - \prod_{j=1}^{\kappa}\left\{1-\pi_{(\mathrm{m}),d^*}\right\}^{c_j}, \tag{54}$$

and (d) holds due to $\sum_{d=1}^{\kappa} f_d(\triangle t) \leq (2\kappa^{1-\eta})/D^{1-\eta} \leq (2\kappa^{2-\eta})/D^{1-\eta}$ [29, Appendix D]. We can further derive Eq. (53) as follows:

$$P_c(k) \leq \frac{2\pi_{\mathrm{m}}(\lambda,\mu_{\mathrm{m}})}{D^{1-\eta}}\sum_{\kappa=0}^{K-1}\kappa^2\, p\left(\kappa,\pi l^2,k_0\right)\Phi_d(\kappa)$$

$$\overset{\text{(e)}}{\leq} \frac{2\pi_{\rm m}(\lambda,\mu_{\rm m})}{D^{1-\eta}}\sum_{\kappa=0}^{K-1}\kappa^2 p\big(\kappa,\pi l^2,k_0\big)\sum_{\kappa=0}^{K-1}\Phi_d(\kappa)$$
$$= \frac{2\pi_{\rm m}(\lambda,\mu_{\rm m})}{D^{1-\eta}}\mathbb{E}\left[\kappa^2\right]\left\{K-\sum_{\kappa=0}^{K-1}\{1-\pi_{({\rm m}),{\rm d}^*}\}^{\sum_{j=1}^{\kappa}c_j}\right\}$$
$$\leq \frac{2\pi_{\rm m}(\lambda,\mu_{\rm m})}{D^{1-\eta}}\mathbb{E}\left[\kappa^2\right]\left\{K-\{1-\pi_{({\rm m}),d^*}\}^{\sum_{j=1}^{K-1}c_j}\right\} \quad (55)$$

where (e) is due to $\sum_i x_i y_i \leq \sum_i x_i \sum_i y_i, \forall x_i, y_i > 0$. To calculate $\mathbb{E}[\kappa^2]$, define the rate for the Poisson process, denoted by $\lambda_{\rm p}$, as $\lambda_{\rm p} \triangleq \pi l^2 k_0$. Then, we can calculate $\mathbb{E}[\kappa^2]$ as follows:

$$\mathbb{E}\big[\kappa^2\big] = \mathbb{E}\left[\kappa(\kappa-1)+\kappa\right] = \mathbb{E}\left[\kappa(\kappa-1)\right]+\mathbb{E}\left[\kappa\right]$$
$$= \sum_{\kappa=0}^{K-1}\kappa(\kappa-1)\frac{e^{-\lambda_{\rm p}}\lambda_{\rm p}^{\kappa}}{\kappa!}+\lambda_{\rm p} \leq \sum_{\kappa=0}^{\infty}\kappa(\kappa-1)\frac{e^{-\lambda_{\rm p}}\lambda_{\rm p}^{\kappa}}{\kappa!}+\lambda_{\rm p}$$
$$= \sum_{\kappa=2}^{\infty}\kappa(\kappa-1)\frac{e^{-\lambda_{\rm p}}\lambda_{\rm p}^{\kappa}}{\kappa!}+\lambda_{\rm p} = \sum_{\kappa=2}^{\infty}\frac{e^{-\lambda_{\rm p}}\lambda_{\rm p}^{\kappa}}{(\kappa-2)!}+\lambda_{\rm p}$$
$$= e^{-\lambda_{\rm p}}\lambda_{\rm p}^2\sum_{\kappa=2}^{\infty}\frac{\lambda_{\rm p}^{\kappa-2}}{(\kappa-2)!}+\lambda_{\rm p} = e^{-\lambda_{\rm p}}\lambda_{\rm p}^2 e^{\lambda_{\rm p}}+\lambda_{\rm p} = \lambda_{\rm p}^2+\lambda_{\rm p}. \quad (56)$$

Thus, using Eq. (56), we can calculate the term of $\mathbb{E}[\kappa^2]$ in Eq. (55) and we can further derive Eq. (55), which yields the upper-bound of $P_c(k)$, as follows:

$$P_c(k) \leq \frac{2\pi_{\rm m}(\lambda,\mu_{\rm m})}{D^{1-\eta}}\left(\pi^2 l^4 k_0^2+\pi l^2 k_0\right) \times\left[K-\{1-\pi_{({\rm m}),d^*}\}^{\sum_{j=1}^{K-1}c_j}\right]. \quad (57)$$

Using $P_c(k)$'s lower-bound derived in Eq. (48) and upper-bound derived in Eq. (57), we obtain the range of cache hitting rate $P_c(k)$ for the $k$th MU, and thus we can also further derive the corresponding optimal $l$ and $\pi_{({\rm m}),d}$ which maximize the $k$th MU's $\epsilon$-effective capacity defined in the optimization problem formulated by Eq. (36).

### B. Intra-Tier Collaborative Caching Mechanism for Caching Tier 2 At Cellular Base Station and WiFi Access Point

When the popularity $f_d(\triangle t)$ of a DT data content item is higher than the threshold of Tier 2 but is lower than a threshold of Tier 3 (i.e., $f^{(\rm T2)} < f_d(\triangle t) < f^{(\rm T3)}$), we propose to cache it at the cellular BS or WiFi AP at Tier 2. In this inter-tier collaborative caching mechanism of caching Tier 2, a data content item can be cached at the cellular BS with probability $\pi_{\rm b}^{(\rm BS)}$; or can be cached at the WiFi AP with probability $\pi_{\rm b}^{(\rm AP)}$. We propose that if the MU $k$ requests the data content item $d, \forall d$, and another MU, denoted by $u$, $\forall u \in \mathcal{M}_k, u \neq k$, requests the same data content item, the MU $k$ forwards this data content item $d$ to MU $u$ through D2D communications.

Our proposed caching Tier 2's inter-tier collaborative caching mechanism aims at maximizing the average $\epsilon$-effective capacity for each MU if receiving the data from Tier 2. Thus, we construct the objective function for the intra-tier collaborative caching mechanism at Tier 2 as follows:

$$\max_{\pi_{\rm b}^{(\rm BS)},\pi_{\rm b}^{(\rm AP)}} \left\{EC_k^{(\rm b)}(\theta_k,\epsilon_k,\mathcal{P}_k)\right\} \quad (58)$$
$$\text{s.t.: C3}: \pi_{\rm b}(\lambda,\mu_{\rm b})D \leq S_b,$$
$$\text{C5}: \pi_{\rm b}^{(\rm BS)}+\pi_{\rm b}^{(\rm AP)} = \pi_{\rm b}(\lambda,\mu_{\rm b}),$$

where $EC_k^{(\rm b)}(\theta_k,\epsilon_k,\mathcal{P}_k)$, as initially defined in Section II-B, is the $\epsilon$-effective capacity for transmitting the data cached at Tier 2 to the MU $k$, and can be further derived as follows:

$$EC_k^{(\rm b)}(\theta_k,\epsilon_k,\mathcal{P}_k)$$
$$= \left\{\frac{1}{2}\left[EC_k^{\rm AP}(\theta_k,\epsilon_k,\mathcal{P}_{\rm AP})+EC_u^{\rm D2D}(\theta_u,\epsilon_u,\mathcal{P}_{\rm D2D})\right]\pi_{\rm D2D} +EC_k^{\rm AP}(\theta_k,\epsilon_k,\mathcal{P}_{\rm AP})(1-\pi_{\rm D2D})\right\}\pi_{\rm b}^{(\rm AP)}$$
$$+\left\{\frac{1}{2}\left[EC_k^{\rm BS}(\theta_k,\epsilon_k,\mathcal{P}_{\rm BS})+EC_u^{\rm D2D}(\theta_u,\epsilon_u,\mathcal{P}_{\rm D2D})\right]\pi_{\rm D2D} +EC_k^{\rm BS}(\theta_k,\epsilon_k,\mathcal{P}_{\rm BS})(1-\pi_{\rm D2D})\right\}\pi_{\rm b}^{(\rm BS)}, \quad (59)$$

where $\pi_{\rm D2D}$ is the probability that the requested data content item of MU $k$ is also requested by MU $u, \forall u \in \mathcal{M}_k, u \neq k$, and thus can be forwarded to MU $u$ by using D2D communications, which is given by

$$\pi_{\rm D2D} = \sum_{d=1}^{D} f_d(\triangle t)\left\{\sum_{\kappa=0}^{K-1}\left[p\left(\kappa,\pi l^2,k_0\right)\left(1-[1-f_d(\triangle t)]^{\kappa}\right)\right]\right\}. \quad (60)$$

In Eq. (59), $EC_k^{\rm BS}(\theta_k,\epsilon_k,\mathcal{P}_{\rm BS})$ and $EC_k^{\rm AP}(\theta_k,\epsilon_k,\mathcal{P}_{\rm AP})$ are given by Eq. (22) and Eq. (25), respectively; and $EC_u^{\rm D2D}(\theta_u,\epsilon_u,\mathcal{P}_{\rm D2D})$ can be derived by replacing $\theta_k$ and $\epsilon_k$ in Eq. (38) by $\theta_u$ and $\epsilon_u$, respectively.

### C. Intra-Tier Collaborative Caching Mechanism for Caching Tier 1 At Routers

When the popularity $f_d(\triangle t)$ of a DT data content item is higher than the threshold of Tier 1 but is lower than a threshold of Tier 2 (i.e., $f^{(\rm T1)} < f_d(\triangle t) < f^{(\rm T2)}$), we propose to cache it in the router at Tier 1. To maximize the caching gain at Tier 1, we need to develop the efficient algorithms to determine (i) where (in which router), (ii) what (what data content item need to be cached), and (iii) how long to cache the requested data content item in a router. The above problem (iii) is different from the derivation of $T_{\rm r}$ defined in Section IV, because $T_{\rm r}$ is the caching lifespan for Tier 1 while the above problem (iii) being the caching lifespan in one router and cached data content items being able to switch routers within Tier 1. As shown in Fig. 1, our caching mechanism model focuses on a set of MUs belonging to a BS connected to the data-source provider through a series of cacheable routers constituting a cacheable path in Tier 1. For problem (i), we denote the set of all cacheable routers along this cacheable path by $\mathcal{H} \triangleq \{1,2,\ldots,H\}$ with its index $h \in \mathcal{H}$. The distances, measured by the number of hops from the caching routers to the BS of the wireless cell and denoted by $\ell_h$, are ordered by $\ell_1 < \ell_2 < \cdots < \ell_H$. Regarding problems (ii) and (iii), we also define the *router's remaining caching lifespan* for the $d$th data content item at time $t$, denoted by $L_d(t)$ $(L_d(t) \leq T_{\rm r})$, as the residual time for the $d$th data content item to be saved in a cacheable router from time $t$. Similar to Sections V-A and V-B, we also need to derive the instantaneous popularity

**Algorithm 2:** Intra-Tier Collaborative Caching Algorithms for Caching in Cacheable Routers at Tier 1.

**initialize**: DT Data content item $d$ is only saved in the original data-source provider, $f_d(\triangle t)=0$, $L_d(t)=0$.
**while** MUs request to visit/download the data content item $d$ **do**
  Update $f_d(\triangle t)$.
  **if** $f_d(\triangle t) > f^{(\mathrm{T1})}$ and $T_{\mathrm{r}}$ has not expired **then**
    Cache/refresh data content item $d$ in a router according to the updated popularity profile $f_d(\triangle t)$. (The larger $f_d(\triangle t)$ is, the closer cacheable router to the wireless cell is chosen to cache the data content item $d$.)
    Update $L_d(t)$ according to $f_d(\triangle t)$ and delete the data content item $d$ from the previous router.
  **else**
    Set $L_d(t)=0$ and delete the data content item $d$ from any caches at cacheable routers.
  **end if**
**end while**

$f_d(\triangle t), \forall d$, of the $d$th data content item in this section, as the revisited probability by MUs during the last $\triangle t$ time period from the current observing time.

Due to the dynamics of the DT data-content popularity profile and the mobile wireless networks, a dynamic caching algorithm is needed. Thus, we develop a caching algorithm for routers where both the caching location and the router's remaining caching lifespan for a given data content item are associated with the data content item's instantaneous popularity profile. More specifically, a data content item with a larger instantaneous popularity is cached in the cacheable routers closer (a router with less $\ell_h$) to the wireless cell, and has a longer lifespan in the router. When receiving a data content item request from an MU, the network controller updates $f_d(\triangle t)$ of the $d$th data content item. Since the cacheable routers have different distances to the wireless cell (i.e., BS), a data content item with a larger $f_d(\triangle t)$ is cached in a router closer to the wireless cell (i.e., BS). This is because MUs are more likely to request a popular DT data content item afterwards, and thus caching the popular DT data in a nearby router can efficiently offload the traffic and reduce the transmission delay which is critically important to support statical QoS provisioning for transmitting 6G DT traffics in the core networks. If a data content item's $f_d(\triangle t)$ decreases to the popularity threshold for caching in routers $f^{(\mathrm{T1})}$, we set the data content item's remaining caching lifespan $L_d(t)=0$ and delete its cached copies in any cacheable routers. As a result, this data content item will only exist in the data-source provider. Then, **Algorithm** 2 details the caching location and router's remaining caching lifespan for our intra-tier collaborative caching algorithm at Tier 1. To simplify our description, we use the data content item $d$ as an example, and all data content items, $\forall d \in \{1,2,\ldots,D\}$, follow the same collaborative caching algorithm at Tier 1.

Note that the connections among routers or between a router and cellular-BS/WiFi-AP are wired links, and we do not investigate the $\epsilon$-effective capacity when data is transmitted within this Tier 1 because the $\epsilon$-effective capacity only exists in wireless links. Since delivering the data content items cached at Tier 1 also need to go through the cellular BS or WiFi AP, the wireless delivery pathes for a data content item cached at Tier 1 and Tier 2 are the same, and thus, the $\epsilon$-effective capacity for receiving a data content item cached at Tier 1 and Tier 2 are the same. As defined in C5 in Eq. (33), $\widetilde{\pi}_r^{(\mathrm{BS})}$ and $\widetilde{\pi}_r^{(\mathrm{AP})}$ are the probabilities that the data content item is cached at Tier 1 to be transmitted going through cellular BS and WiFi AP, respectively. The objective function for intra-tier collaborative caching mechanism at Tier 1 is shown as follows:

$$\max_{\widetilde{\pi}_{\mathrm{r}}^{(\mathrm{BS})},\, \widetilde{\pi}_{\mathrm{r}}^{(\mathrm{AP})}} \left\{ EC^{(\mathrm{r})}(\theta_k, \epsilon_k, \mathcal{P}_k) \right\} \tag{61}$$

$$\text{s.t.: C3: } \pi_{\mathrm{r}}(\lambda, \mu_{\mathrm{r}}) D \le S_{\mathrm{r}},$$

$$\text{C6: } \widetilde{\pi}_{\mathrm{r}}^{(\mathrm{BS})} + \widetilde{\pi}_{\mathrm{r}}^{(\mathrm{AP})} = \pi_{\mathrm{r}}(\lambda, \mu_{\mathrm{r}}),$$

where $EC_k^{(\mathrm{r})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ can be obtained by replacing $\pi_{\mathrm{b}}^{(\mathrm{AP})}$ and $\pi_{\mathrm{b}}^{(\mathrm{BS})}$ in Eq. (59) by $\widetilde{\pi}_{\mathrm{r}}^{(\mathrm{BS})}$ and $\widetilde{\pi}_{\mathrm{r}}^{(\mathrm{AP})}$, respectively.

## VI. Optimal Caching Scheme for Supporting Adaptive Blocklength to Minimize Total Transmission Delay

In Sections IV and V, we assume that the blocklength $n$ is a constant for all DT data-content items. In this section, we propose the data-adaptive caching schemes for DT data-content items if the blocklength of the $d$th DT data content item, denoted by $n_d, \forall d$, can be adjusted dynamically according to the $d$th data item's popularity $f_d(\triangle t)$ and MUs' statistical QoS requirements $\theta_k$ and $\epsilon_k$.

### A. DT Data Collection and Adaptation Schemes Based on MUs' Statistical QoS Requirements

Since MUs are frequently joining and leaving the cellular network and each MU has its unique DT data request, channel conditions, and statistical QoS requirements on delay exponent $\theta_k$ and decoding error probability $\epsilon_k$, we propose to dynamically adapt the DT data collection procedure model, i.e., the data encoding blocklength $n_d$, for the virtual representation of the $d$th DT data-content item based on MUs' statistical QoS requirements and its current popularity $f_d(\triangle t)$. The DT data content items' digital transformation scheme is shown by **Algorithm** 3. Based on **Algorithm** 3, we can derive the data-adaptive optimal caching scheme/policy to minimum the total transmission delay as detailed in the section that follows.

### B. Optimal Caching Scheme for Selecting the Caching Route With the Minimum Total Transmission Delay

Based on Definition 2, $\epsilon$-effective capacity is the maximum data constant arrival rate under a given delay exponent and a given decoding error probability. Thus, the transmission delays for delivering the $d$th DT data to the $k$th MU if the data is cached at Tier 1, Tier 2, and Tier 3, denoted by $\omega_d^{(\mathrm{r})}$, $\omega_d^{(\mathrm{b})}$, and $\omega_d^{(\mathrm{m})}$, respectively, are given by $\omega_d^{(\mathrm{r})} = n_d / EC_k^{(\mathrm{r})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, $\omega_d^{(\mathrm{b})} = n_d / EC_k^{(\mathrm{b})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, and $\omega_d^{(\mathrm{m})} = n_d / EC_k^{(\mathrm{m})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, where $n_d$ is obtained by using **Algorithm** 3, $EC_k^{(\mathrm{r})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ can be obtained as described in the text following Eq. (61), $EC_k^{(\mathrm{b})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ is given by Eq. (59), and $EC_k^{(\mathrm{m})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ is given by Eq. (36). Note that under the adaptive blocklength scheme, $\epsilon$-effective capacities $EC_k^{(\mathrm{r})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, $EC_k^{(\mathrm{b})}(\theta_k, \epsilon_k, \mathcal{P}_k)$, and $EC_k^{(\mathrm{m})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ are functions of the blocklength $n_d$ according to Eq. (7). Thus, we can obtain the data-adaptive optimal caching scheme/policy for choosing the route with the minimum data transmission delay

**Algorithm 3:** DT Data-Content Items' Adaptive Data Collection Scheme.

1: **initialize**: The popularity $f_d(\triangle t)$ of the DT data content item $d$, MUs' channel fading parameter $m$, MUs' statistical QoS requirements $\theta_k$ and $\epsilon_k, \forall k$.
2: **if** $f_d(\triangle t) > f^{(\mathrm{T3})}$ **then**
3: The $d$th DT data-content item is cached at Tier 3.
4: Adjust the blocklength $n_d$ to maximize the $\epsilon$-effective capacity $EC_k^{\mathrm{D2D}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{D2D}})$ using Eq. (38).
5: **else** $f^{(\mathrm{T1})} < f_d(\triangle t) < f^{(\mathrm{T3})}$
6: The $d$th DT data-content item is cached at Tier 2 or Tier 1.
7: Adjust the blocklength $n_d$ to maximize the $\epsilon$-effective capacity $EC_k^{\mathrm{BS}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{BS}})$ using Eq. (22) or to maximize the $\epsilon$-effective capacity $EC_k^{\mathrm{AP}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{AP}})$ using Eq. (25).
8: **end if**
9: **output**: The optimal encoding blocklength $n_d$ for DT data collection and adaptation.

as follows:

$$\begin{cases} \min\limits_{\widetilde{\pi}_{\mathrm{r}}^{(\mathrm{BS})}, \widetilde{\pi}_{\mathrm{r}}^{(\mathrm{AP})}} \left\{\omega_d^{(\mathrm{r})}\right\}, \text{ if the } d\text{th data item is cached at Tier 1} \\ \qquad \left(\text{i.e., } f^{(\mathrm{T1})} \le f_d(\triangle t) < f^{(\mathrm{T2})}\right), \\ \min\limits_{\pi_{\mathrm{b}}^{(\mathrm{BS})}, \pi_{\mathrm{b}}^{(\mathrm{AP})}} \left\{\omega_d^{(\mathrm{b})}\right\}, \text{ if the } d\text{th data item is cached at Tier 2} \\ \qquad \left(\text{i.e., } f^{(\mathrm{T2})} \le f_d(\triangle t) < f^{(\mathrm{T3})}\right), \\ \min\limits_{l, \pi_{(\mathrm{m}),d}} \left\{\omega_d^{(\mathrm{m})}\right\}, \text{ if the } d\text{th data item is cached at Tier 3} \\ \qquad \left(\text{i.e., } f_d(\triangle t) \ge f^{(\mathrm{T3})}\right), \end{cases} \tag{62}$$

where $f_d(\triangle t)$ is specified in **Algorithm** 3 presented in the last section, and $f^{(\mathrm{T1})}$, $f^{(\mathrm{T2})}$, and $f^{(\mathrm{T3})}$ are defined in Section IV.

## VII. Performance Evaluations

We conduct the extensive numerical analyses to validate and evaluate our developed schemes. In Fig. 2(a), we set the total number of MUs to be $K = 20$, the cardinality of D2D communication set of the $k$th MU to be $|\mathcal{M}_k| = 5$, and the $j$th mobile user's caching capability to be $c_j = 1, \forall j$. Fig. 2(a) shows that given the caching probability at Tier 3 $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$, we can find an optimal D2D communication range $l$ of MUs that maximizes the cache hitting rate shown in Eq. (36), validating the existence of the optimal solution of Eq. (36). Therefore, the optimal aggregate $\epsilon$-effective capacity also exists. Fig. 2(a) also shows that given a D2D communication range $l$, the $\epsilon$-effective capacity for caching at Tier 3 is an increasing function of the caching probability at Tier 3 $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$. This is because a larger value of $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ implies a larger number of data content items which are cached in the mobile devices, and thus, leads to a larger probability to download the requested data item using D2D communications.

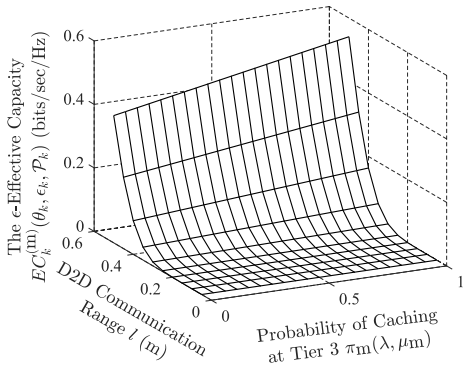
Fig. 2(b) shows a lower-bound on the cache hitting rate under different values of caching probability at Tier 3 $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$. We set the caching probability $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ as 0.5, 0.6, and 0.7, respectively. For each value of $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$, we plot a lower-bound on the cache hitting rate $P_c(k)$ under the total caching capability, denoted by $c_{\mathrm{t}} \triangleq \sum_{j=1}^{u^*} c_j$ to be 3 and 5, respectively. We also set the total number of MUs $K = 20$, the probability that the $d$th data content item is cached at a cache unit of a mobile device $\pi_{(m),d^*} = 0.2$, and the range parameter $\delta = 0.2$. Fig. 2(b) shows that $P_c(k)$ increases as caching probability $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ increases, because a larger value of $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ indicates that more mobile devices are able to download a requested data content item within Tier 3. Therefore, given a certain $c_{\mathrm{t}}$, a larger $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ yields a larger value of the lower-bound on $P_c(k)$. We can also observe that the lower-bound on $P_c(k)$ is a decreasing function of the total caching capability $c_{\mathrm{t}}$, because we have $\pi_{(\mathrm{m}),d^*} < 1$ in Eq. (48). Thus, a larger $c_{\mathrm{t}}$ results in a smaller $\pi_{(\mathrm{m}),d^*}^{c_{\mathrm{t}}}$.

Fig. 2(c) plots a lower-bound on the cache hitting rate $P_c(k)$ under different total numbers of MUs $K$ and different values of unit memory space caching probability $\pi_{(\mathrm{m}),d}$. We set $K$ as 20 and 30, respectively. For each value of $K$, Fig. 2(c) plots the lower-bound of $P_c(k)$ under unit memory space caching probabilities $\pi_{(\mathrm{m}),d} = 0.2, 0.3$, and 0.4, respectively. We also set $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}}) = 0.5$, $\delta = 0.2$, and $c_{\mathrm{t}} = 5$. We observe from Fig. 2(c) that $P_c(k)$ is an increasing function of $K$ and $f_d(\triangle t)$, because a larger $K$ indicates a larger number of MUs in the wireless cell, and thus, there are more downloading requests of the $d$th data content item. Therefore, under a certain value of $\pi_{(\mathrm{m}),d}$, a cellular network with a larger number of users $K$ and a larger requesting probability $f_d(\triangle t)$ has a larger value of the lower-bound on $P_c(k)$. We also observe that the lower-bound on $P_c(k)$ is an increasing function of the unit memory space caching probability $\pi_{(\mathrm{m}),d}$. This is because the larger value of $\pi_{(\mathrm{m}),d}$ implies a larger probability to cache the $d$th data content item, leading to a larger probability that an MU can download the $d$th data content item from its neighbor peers.

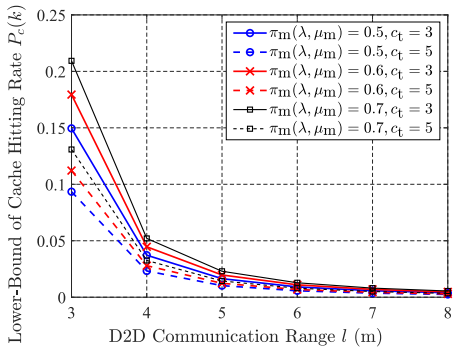
Fig. 2(d) plots the upper-bound on the cache hitting rate $P_c(k)$ under different values of total number of MUs $K$ and different values of Zipf distribution exponent $\eta$. We set $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}}) = 0.5$ and the total caching capability $\sum_{j=1}^{u^*} c_j \le 10$. Fig. 2(d) shows that the upper-bound on $P_c(k)$ is an increasing function of $K$, since a larger value of $K$ implies a larger density of MUs within a cellular network, and therefore, implies that the MU $k$ has more cooperation mobile devices that cache the data. We can also observe from Fig. 2(d) that the upper-bound on $P_c(k)$ is a decreasing function of $\eta$, since a larger value of $\eta$ implies a larger number of mobile devices caching the top frequently requested data content items, reducing the number of D2D pairs for downloading the less frequently requested data.

In Fig. 3(a), we plot the SNR of the $k$th MU $\gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}})$ under different values of distance $l_k$ between the $k$th MU and the massive MIMO BS. We set $M_T = 100$, $M_R = 2$, $K = 40$, $\lambda_{\mathrm{c}} = 10$m, $\tau_{\mathrm{ul,p}} = 16$, $\rho_{\mathrm{ul}} = 1$W, and $\mathbb{E}[\eta_{k,\alpha}^{(i)}] = 0.5$, the range of distance $l_k$ as [4,10] m, length of a codeword $n = 900$, and channel fading range $h_{k,\alpha}^{(q)} = [0.1, 1]$. Fig. 3(a) shows that there exists an optimal distance $l_k$ which maximizes the SNR, implying that a mobile user can adjust its position to obtain a better wireless channel. We also observe from Fig. 3(a) that although the SNR is an increasing function of $\mathcal{P}_{\mathrm{BS}}$, the optimal SNR is approximately the same for the optimal $l_k$ when increasing $\mathcal{P}_{\mathrm{BS}}$. This implies that at the optimal position of a mobile user, the transmit power allocation has a small impact on the SNR.

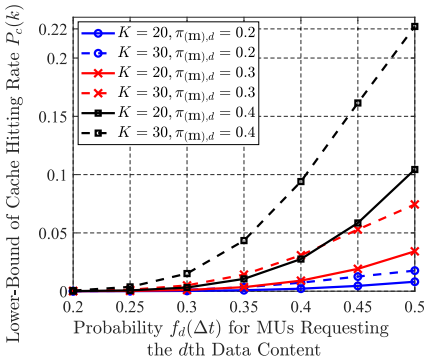
In Fig. 3(b), we plot the $\epsilon$-effective capacity under different values of distance $l_k$ between the $k$th MU and the massive MIMO BS. We set $w_k^{(q)} = 0.05$W. According to *Remarks on Theorem 2*, $\gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}}), \forall q$, are the same. Fig. 3(b) shows that the $\epsilon$-effective capacity decreases as the distance $l_k$ increases



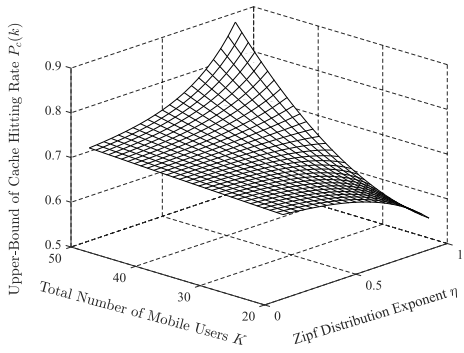

(a) The $\epsilon$-effective capacity for Tier 3 $EC_k^{(\mathrm{m})}(\theta_k, \epsilon_k, \mathcal{P}_k)$ under different values of caching probability at Tier 3 $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ and different values of D2D communication range $l$.




(b) The lower-bound on the cache hitting rate $P_c(k)$ under different values of caching probability at Tier 3 $\pi_{\mathrm{m}}(\lambda, \mu_{\mathrm{m}})$ and different values of D2D communication range $l$.

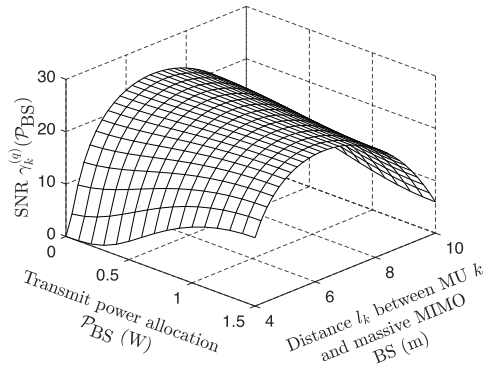



(c) The lower-bound on the cache hitting rate $P_c(k)$ under different total numbers of mobile users $K$ and different unit memory space caching probabilities $\pi_{(\mathrm{m}),d}$.

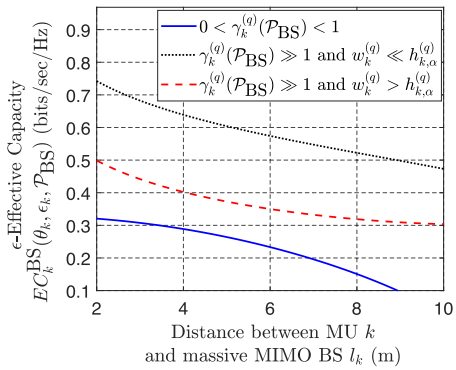



(d) The upper-bound on the cache hitting rate $P_c(k)$ under different total numbers of mobile users $K$ and different Zipf distribution exponents $\eta$.

Fig. 2. Comparisons of network performances for the intra-tier collaborative caching mechanism at Tier 3 on mobile devices.




(a) The SNR $\gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}})$ under different values of distance $l_k$ between the mobile user $k$ and the massive MIMO BS and different values of transmit power allocation $\mathcal{P}_{\mathrm{BS}}$.




(b) The $\epsilon$-effective capacity $EC_k^{\mathrm{BS}}(\theta_k, \epsilon_k, \mathcal{P}_{\mathrm{BS}})$ under different values of the distance $l_k$ and different values of the SNR for MU $k$ $\gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}})$.

Fig. 3. Performances comparisons for the intra-tier collaborative caching mechanism for Tier 2 on massive MIMO BS and MIMO WiFi AP.

because the large scale fading $\beta_k$ is a decreasing function of $l_k$. Fig. 3(b) validates that the $\epsilon$-effective capacity under the condition of $0 < \gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}}) < 1$ is always lower than that under the other two conditions, as it has the minimum SNR. Fig. 3(b) also validates that when $\gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}}) \gg 1$ and $w_k^{(q)} \ll h_{k,\alpha}^{(q)}$, the $\epsilon$-effective capacity is always larger than that with $\gamma_k^{(q)}(\mathcal{P}_{\mathrm{BS}}) \gg 1$ and $w_k^{(q)} > h_{k,\alpha}^{(q)}$ because an increased $h_{k,\alpha}^{(q)}$ leads to the larger $\epsilon$-effective capacity.

In Fig. 4, we compare the average transmission delay for receiving a DT data content item under our proposed multi-tier caching scheme with that of the random caching scheme and a caching scheme without employing the multi-tier mechanism. In the random caching scheme, defined as baseline caching scheme

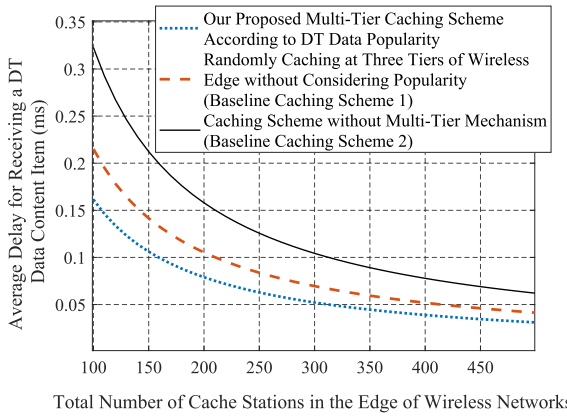



Fig. 4. Comparisons of the average transmission delay for receiving a DT data content item under our proposed multi-tier scheme with the existing baseline caching schemes.

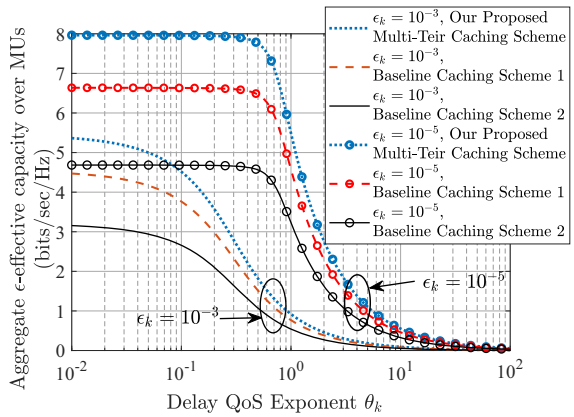



Fig. 5. Comparisons of the aggregate $\epsilon$-effective capacity over MUs for our proposed multi-tier caching scheme with the existing baseline caching schemes.

1, we assume that DT data content items are randomly cached at three tiers without considering their popularities. In the caching scheme without employing multi-tier caching mechanisms, defined as baseline caching scheme 2, we set that all cache stations are located at the massive MIMO BS and WiFi AP. We set that the popularity of DT data content items follows the Zipf distribution with the Zipf exponent 0.5 and there are a total of 1000 DT data content items. We also set that the transmission delay for downloading a data item from a mobile device using D2D communication, from a BS or WiFi AP, and from a router to an MU is 0.02 ms, 0.1 ms, and 0.4 ms, respectively. We can observe that the average transmission delay decreases as the total number of cache stations in wireless networks increases for all three caching schemes. We can also observe that our proposed scheme by caching the most popular content in Tier 3 can achieve the minimum average transmission delay.

We compare the aggregate $\epsilon$-effective capacity over MUs for our proposed multi-tier caching scheme with two existing baseline caching schemes in Fig. 5. Baseline caching schemes 1 and 2 are the same as those in Fig. 4. We assume that there are 10 MUs in the wireless network requesting 50 DT data contents, blocklength $n = 1000$, decoding error probabilities $\epsilon_k$ are the same for all MUs, and other parameters are the same as those in Fig. 4. We can observe that comparing with the two existing baseline caching schemes, our proposed multi-tier caching scheme can achieve the maximum aggregate $\epsilon$-effective capacity for both $\epsilon_k = 10^{-3}$ and $10^{-5}$. We can also observe that for the same caching scheme, the aggregate $\epsilon$-effective capacity under $\epsilon_k = 10^{-5}$ is larger than that of $\epsilon_k = 10^{-3}$. This is because the smaller $\epsilon_k$ indicates a better channel quality, and thus, the wireless network can achieve a larger aggregate $\epsilon$-effective capacity.

## VIII. Conclusions

We have proposed collaborative multi-tier hierarchical caching mechanisms for DT over mURLLC-based 6G mobile wireless networks, where popular DT data can be cached at different wireless network caching tiers (e.g., at routers (Tier 1), massive-MIMO-BS/WiFi-AP (Tier 2), and mobile devices (Tier 3), respectively). In particular, we have developed an inter-tier collaborative hierarchical caching scheme, which maximizes the aggregate $\epsilon$-effective capacity across all three caching tiers. Then, we have also developed the intra-tier collaborative hierarchical caching scheme at Tier 1, through updating cached data content items and caching lifespans according to the popularity of a DT data content item. The intra-tier collaborative hierarchical caching mechanism at Tier 2 optimizes the data content items' caching probabilities at the cellular BS and WiFi AP, respectively. The intra-tier collaborative hierarchical caching mechanism at Tier 3 optimizes the D2D communication distance and the data caching probability at each cache unit.

## Appendix A
## Proof of Theorem 1

*Proof:* To simplify the notation, we use $\gamma_k$ and $\overline{\gamma}_k$ to replace $\gamma_k(\mathcal{P}_k)$ and $\overline{\gamma}_k(\mathcal{P}_k)$, respectively, in this proof. Using Eq. (6), we can derive the expectation of decoding error probability as follows:

$$\mathbb{E}_{\gamma_k}[\epsilon_k(\gamma_k)] = \int_0^\infty Q\left(\frac{\log_2(1+\gamma_k) - \frac{\log_2 W_k}{n}}{\sqrt{\frac{2\gamma_k+\gamma_k^2}{n}}\Big/(1+\gamma_k)}\right) P_Z(\gamma_k) d\gamma_k$$

$$= Q\left(\int_0^\infty \frac{\log_2(1+\gamma_k) - \frac{\log_2 W_k}{n}}{\sqrt{\frac{2\gamma_k+\gamma_k^2}{n}}\Big/(1+\gamma_k)} P_Z(\gamma_k)\, d(\gamma_k)\right), \quad (63)$$

which can be further derived by considering the following two cases, respectively.

**Case 1.** When $\gamma_k \gg 1$, Eq. (63) can be further derived as:

$$\mathbb{E}_{\gamma_k}[\epsilon_k(\gamma_k)] \approx Q\left(\sqrt{n}\left[\int_0^\infty \log_2(1+\gamma_k) P_Z(\gamma_k) d\gamma_k - \frac{\log_2 W_k}{n}\right]\right)$$

$$\overset{(f)}{=} Q\left(\sqrt{n}\left[\log_2(1+\overline{\gamma}_k) - \frac{\log_2 W_k}{n}\right]\right) \quad (64)$$

where (f) holds by applying Taylor-series expansion over $\log_2(1+\gamma_k)$, taking integral for each term under the Nakagami-$m$ distribution, and using Taylor series again.

**Case 2.** When $0 < \gamma_k < 1$, by applying Taylor series that $1/[(1-x)^2] = \sum_{i=1}^\infty i x^{i-1}$ and $\log(1+x) = \sum_{i=1}^\infty [(-1)^{i+1} x^i]/i$, we have

$$\frac{\sqrt{\frac{2\gamma_k+\gamma_k^2}{n}}}{1+\gamma_k} = \frac{1}{\sqrt{n}}\sqrt{1-\frac{1}{(1+\gamma_k)^2}} \approx \frac{\sqrt{1-(1-2\gamma_k)}}{\sqrt{n}} = \sqrt{\frac{2\gamma_k}{n}} \quad (65)$$

and

$$\log_2(1+\gamma_k) \approx (\log_2 e)\gamma_k. \quad (66)$$

Substituting Eq. (65) and Eq. (66) into Eq. (63), we can further derive $\mathbb{E}_{\gamma_k}[\epsilon_k(\gamma_k)]$ as follows:

$$\mathbb{E}_{\gamma_k}[\epsilon_k(\gamma_k)] \approx Q\left(\sqrt{\frac{n}{2}}\left[(\log_2 e)\int_0^{\infty}\sqrt{\gamma_k}P_Z(\gamma_k)\,d\gamma_k - \frac{\log_2 W}{n}\int_0^{\infty}\frac{1}{\sqrt{\gamma_k}}P_Z(\gamma_k)\,d\gamma_k\right]\right) = Q\left(\sqrt{\frac{n}{2}}\left[(\log_2 e)\overline{\gamma}_k^{\frac{1}{2}} - \frac{\log_2 W_k}{n\sqrt{\overline{\gamma}_k}}\right]\right). \quad (67)$$

Combining Eq. (64) derived in **Case 1** with Eq. (67) derived in **Case 2**, we have

$$\mathbb{E}_{\gamma_k}[\epsilon_k(\gamma_k)] = \begin{cases} Q\left(\sqrt{n}\left[\log_2(1+\overline{\gamma}_k) - \frac{\log_2 W_k}{n}\right]\right), & \text{if } \overline{\gamma}_k \gg 1 \\ Q\left(\sqrt{\frac{n}{2}}\left[(\log_2 e)\overline{\gamma}_k^{\frac{1}{2}} - \frac{\log_2 W_k}{n\sqrt{\overline{\gamma}_k}}\right]\right), & \text{if } 0<\overline{\gamma}_k<1. \end{cases} \quad (68)$$

Substituting Eq. (68) back into Eq. (3), we obtain Eq. (7), completing the proof for Theorem 1. ■

## Appendix B
## Proof of Theorem 3

*Proof:* We define the SNR vector, denoted by $\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}})$, for all $M_R$ antennas deployed on MU $k$ as follows:

$$\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}}) \triangleq \left[\gamma_k^{(1)}(\mathcal{P}_{\text{BS}}), \ldots, \gamma_k^{(M_R)}(\mathcal{P}_{\text{BS}})\right] \quad (69)$$

when MU $k$ receives the cached DT data from the massive MIMO BS with the transmit power allocation $\mathcal{P}_{\text{BS}}$, where $\gamma_k^{(q)}(\mathcal{P}_{\text{BS}}), \forall q \in \{1, 2, \ldots, M_R\}$, is defined in Eq. (17) by replacing $\mathcal{P}_k$ by $\mathcal{P}_{\text{BS}}$ in Eq. (17). Since Theorem 2 shows that all $\gamma_k^{(q)}(\mathcal{P}_k), \forall q$, are the same, we re-define $\overline{\gamma}_k(\mathcal{P}_{\text{BS}}) \triangleq \gamma_k^{(q)}(\mathcal{P}_{\text{BS}}), \forall q$, given by Eq. (17) after replacing $\mathcal{P}_k$ by $\mathcal{P}_{\text{BS}}$. Extending the derivations for the $\epsilon$-effective capacity given in Eq. (3) over the single antenna channel into its massive-MIMO-channel version, we can derive the massive MIMO channel's $\epsilon$-effective capacity, denoted by $EC_k^{\text{MIMO}}(\theta_k, \epsilon_k, \mathcal{P}_{\text{BS}})$, for the $k$th MU downloading the DT data-content items from the massive-MIMO antennas equipped BS as follows:

$$EC_k^{\text{BS}}(\theta_k, \epsilon_k, \mathcal{P}_{\text{BS}}) \overset{\text{(g)}}{=} EC_k^{\text{MIMO}}(\theta_k, \epsilon_k, \mathcal{P}_{\text{BS}}) = -\frac{1}{n\theta_k}\log\left\{\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}})) + [1-\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}}))]e^{-\theta_k \log_2 W_k}\right\} \quad (70)$$

where (g) follows due to Eq. (21) and $\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}})$ is defined in Eq. (69). Using the FBC scheme, the decoding error probability of the $k$th MU under massive MIMO channel, denoted by $\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}}))$, can be obtained by extending the decoding error probability of Eq. (4) over the single antenna channel into its massive-MIMO-channel version of $\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}}))$ as follows:

$$\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}})) = \prod_{q=1}^{M_R}\epsilon_k\left(\gamma_k^{(q)}(\mathcal{P}_{\text{BS}})\right) = \prod_{q=1}^{M_R} Q\left(\frac{C\left(\gamma_k^{(q)}(\mathcal{P}_{\text{BS}})\right) - \frac{\log_2 W_k}{n}}{\sqrt{V\left(\gamma_k^{(q)}(\mathcal{P}_{\text{BS}})\right)/n}}\right) \overset{\text{(h)}}{=} \left[Q\left(\frac{C\left(\overline{\gamma}_k(\mathcal{P}_{\text{BS}})\right) - \frac{\log_2 W_k}{n}}{\sqrt{V\left(\overline{\gamma}_k(\mathcal{P}_{\text{BS}})\right)/n}}\right)\right]^{M_R} \quad (71)$$

where (h) holds due to *Remarks on Theorem 2*. Using *Remarks on Theorem 2* and extending Eq. (68) over the single antenna channel into its massive-MIMO-channel version, we can further derive the decoding error probability of $\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}}))$ over the massive-MIMO-channel by employing Eq. (5) and Eq. (71) as follows:

$$\epsilon_k(\boldsymbol{\gamma}_k(\mathcal{P}_{\text{BS}})) = \begin{cases} \left\{Q\left(\sqrt{n}\left[\log_2(1+\overline{\gamma}_k(\mathcal{P}_{\text{BS}})) - \frac{\log_2 W_k}{n}\right]\right)\right\}^{M_R}, \\ \qquad \text{if } \overline{\gamma}_k(\mathcal{P}_{\text{BS}}) \gg 1, \\ \left\{Q\left(\sqrt{\frac{n}{2}}\left[(\log_2 e)\sqrt{\overline{\gamma}_k(\mathcal{P}_{\text{BS}})} - \frac{\log_2 W_k}{n\sqrt{\overline{\gamma}_k(\mathcal{P}_{\text{BS}})}}\right]\right)\right\}^{M_R}, \\ \qquad \text{if } 0<\overline{\gamma}_k(\mathcal{P}_{\text{BS}}) < 1. \end{cases} \quad (72)$$

Defining the first and second parts of Eq. (72) as $Q_1$ and $Q_2$ as defined by the first and second parts of Eq. (23), respectively, and then plugging Eq. (72) into Eq. (70), we can thus prove Eq. (22) and further also prove Eq. (23), which complete the proof of Theorem 3. ■

## References


[1] X. Zhang, Q. Zhu, and H. V. Poor, "Neyman-Pearson criterion driven NFV-SDN architectures and optimal resource-allocations for statistical-QoS based mURLLC over next-generation metaverse mobile networks using FBC," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 3, pp. 570–587, Mar. 2024.

[2] K. Wang, W. Chen, J. Li, Y. Yang, and L. Hanzo, "Joint task offloading and caching for massive MIMO-aided multi-tier computing networks," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 1820–1833, Mar. 2022.

[3] K. Wang, D. Niyato, W. Chen, and A. Nallanathan, "Task-oriented delay-aware multi-tier computing in cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2000–2012, Jul. 2023.

[4] X. Zhang and Q. Zhu, "Collaborative hierarchical caching over 5G edge computing mobile wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2018, pp. 1–6.

[5] L. U. Khan, Z. Han, W. Saad, E. Hossain, M. Guizani, and C. S. Hong, "Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities," *IEEE Commun. Surv. Tuts.*, vol. 24, no. 4, pp. 2230–2254, Fourthquarter 2022.

[6] Y. Hui et al., "Digital-twin-enabled on-demand content delivery in Het-VNets," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14028–14041, Aug. 2023.

[7] K. Zhang, J. Cao, S. Maharjan, and Y. Zhang, "Digital twin empowered content caching in social-aware vehicular edge networks," *IEEE Trans. Comput. Social Syst.*, vol. 9, no. 1, pp. 239–251, Feb. 2022.

[8] D. Van Huynh, S. R. Khosravirad, A. Masaracchia, O. A. Dobre, and T. Q. Duong, "Edge intelligence-based ultra-reliable and low-latency communications for digital twin-enabled metaverse," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1733–1737, Aug. 2022.

[9] N. Apostolakis, L. E. Chatzieleftheriou, D. Bega, M. Gramaglia, and A. Banchs, "Digital twins for next-generation mobile networks: Applications and solutions," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 80–86, Nov. 2023.

[10] L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-twin-enabled 6G: Vision, architectural trends, and future directions," *IEEE Commun. Mag.*, vol. 60, no. 1, pp. 74–80, Jan. 2022.

[11] H. X. Nguyen, R. Trestian, D. To, and M. Tatipamula, "Digital twin for 5G and beyond," *IEEE Commun. Mag.*, vol. 59, no. 2, pp. 10–15, Feb. 2021.

[12] S. Shukla and A. A. Abouzeid, "Optimal device-aware caching," *IEEE Trans. Mobile Comput.*, vol. 16, no. 7, pp. 1994–2007, Jul. 2017.

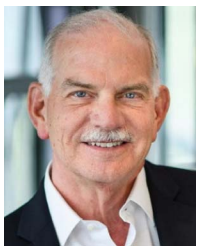
[13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[14] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 118–129, Jan. 2008.
[15] 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.
[16] H. Su and X. Zhang, "Clustering-based multichannel MAC protocols for QoS provisionings over vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 56, no. 6, pp. 3309–3323, Nov. 2007.
[17] 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.
[18] X. Zhang, J. Tang, H.-H. Chen, and S. Ci, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Commun. Mag.*, vol. 44, no. 1, pp. 100–106, Jan. 2006.
[19] X. Zhang and Q. Zhu, "Scalable virtualization and offloading-based software-defined architecture for heterogeneous statistical QoS provisioning over 5G multimedia mobile wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 12, pp. 2787–2804, Dec. 2018.
[20] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*. New York, NY, USA: Wiley, 2004.
[21] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K: Cambridge Univ. Press, 2016.
[22] X. Zhang, Q. Zhu, and H. V. Poor, "Massive-MIMO channel capacity modeling for mURLLC over 6G UAV mobile wireless networks," in *Proc. IEEE Annu. Conf. Inf. Sci. Syst.*, 2022, pp. 49–54.
[23] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
[24] O. Raeesi, A. Gokceoglu, Y. Zou, E. Björnson, and M. Valkama, "Performance analysis of multi-user massive MIMO downlink under channel non-reciprocity and imperfect CSI," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2456–2471, Jun. 2018.
[25] X. Zhang and Q. Zhu, "Statistical delay and error-rate bounded QoS provisioning over massive-MIMO based 6G mobile wireless networks," in *Proc. IEEE Glob. Commun. Conf.*, 2022, pp. 353–358.
[26] X. Zhang and Q. Zhu, "Distributed mobile devices caching over edge computing wireless networks," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2017, pp. 127–132.
[27] X. Zhang and Q. Zhu, "P2P caching schemes for jointly minimizing memory cost and transmission delay over information-centric networks," in *Proc. IEEE Glob. Commun. Conf.*, 2016, pp. 1–6.
[28] J. Jiang, S. Zhang, B. Li, and B. Li, "Maximized cellular traffic offloading via device-to-device content sharing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 82–91, Jan. 2016.
[29] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.

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

He is currently a Full Professor and the Founding Director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is a Fellow of the IEEE for contributions to quality of 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, in 1997. He was a Research Fellow with the School of Electrical Engineering, University of Technology, Sydney, NSW, Australia, and the Department of Electrical and Computer Engineering, James Cook University, Townsville, QLD, Australia. He has published more than 400 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 six Best Paper Awards at IEEE GLOBECOM 2020, 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/JOURNALS papers in the area) on Wireless Cognitive Radio Networks and Statistical QoS Provisioning over Mobile Wireless Networking. He is an IEEE Distinguished Lecturer of both IEEE Communications Society and IEEE Vehicular Technology Society. He received the TEES Select Young Faculty Award for Excellence in Research Performance from the College of Engineering at Texas A&M University, College Station, TX, in 2006, and the Outstanding Faculty Award from Texas A&M University, in 2020.

Prof. Xi Zhang is serving or has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, twice as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for two special issues on "Broadband Wireless Communications for High Speed Vehicles" and "Wireless Video Transmissions", an Associate Editor for IEEE COMMUNICATIONS LETTERS, twice as the Lead Guest Editor of *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 of Wiley's JOURNAL ON WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, JOURNAL OF COMPUTER SYSTEMS, NETWORKING, AND COMMUNICATIONS, and Wiley's JOURNAL ON SECURITY AND COMMUNICATIONS NETWORKS, and an Area Editor for Elsevier's JOURNAL ON COMPUTER COMMUNICATIONS, among many others. He is serving or has served as the TPC Chair for IEEE GLOBECOM 2011, TPC Vice-Chair for IEEE INFOCOM 2010, TPC Area Chair for IEEE INFOCOM 2012, Panel/Demo/Poster Chair for ACM MobiCom 2011, General Chair for IEEE ICDCS 2024 Workshop on "Digital Twin-Enabled 6G Multi-tier Distributed Computing Systems", 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.

**Qixuan Zhu** received the B.S. degree from the Tianjin University of Technology and Education, Tianjin, China, and the M.S. degree from The George Washington University, Washington, DC, USA, all in electrical and computer engineering. She is currently working toward the Ph.D. degree under the supervision of Professor Xi Zhang in the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. She received the Best Paper Award from IEEE ICC 2018. She also received the Hagler Institute for Advanced Study Heep Graduate Fellowship Award and Dr. Christa U. Pandey '84 Fellowship from Texas A&M University.



**H. Vincent Poor** (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University, in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign, Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. During 2006 to 2016, he served as the Dean of the 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, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications*. (Cambridge University Press, 2022). 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 IEEE Alexander Graham Bell Medal in 2017.