2024 IEEE 44th International Conference on Distributed Computing Systems Workshops (ICDCSW) | 979-8-3503-5471-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCSW63686.2024.00015

Hierarchical Caching for Digital Twins Communications Over mURLLC-Based 6G Distributed-Computing Mobile Networks Using FBC

Xi Zhang and Qixuan Zhu

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*}

***Abstract*—The *massive ultra-reliable and low-latency communications* (mURLLC) services are emerging as a new traffic type to support massive numbers of mobile users (MUs) demanding the stringent delay and error-rate bounded *quality-of-services* (QoS) requirements over 6G. Among multiple 6G mURLLC services, *digital twins* (DT) has been widely envisioned as a major intelligent application to support efficient interactions between physical and virtual objects. Moreover, multi-tier caching, which is one of the key distributed computing techniques, stores the frequently-demanded data items at different wireless network tiers to efficiently reduce mURLLC streaming delay and data move. However, how to efficiently cache mURLLC-based DT data items at different caching tiers of wireless networks and how to statistically upper-bound both *delay* and *error-rate* for DT communications remain challenging problems. To overcome these difficulties, in this paper we propose a multi-tier caching mechanism to support DT communications over 6G mobile networks. First, we propose the DT data adaptive collection scheme applying finite blocklength coding (FBC) to dynamically encode a physical object into its virtual representation according to the current network and wireless channel statuses. Second, we develop inter-tier and intra-tier collaborative caching mechanisms, where DT data items are selectively cached at different wireless network caching tiers according to their popularities including: router tier, massive-multiple-input-multiple-output (MIMO) base-station tier, and mobile device tier. Third, our proposed inter-tier collaborative caching mechanisms maximize the aggregate ϵ-effective capacity across all three caching tiers, and our proposed intra-tier collaborative caching mechanisms minimize the sum of data transmission delay for all DT data items cached in each caching tier. Finally, we numerically validate and evaluate our developed multi-tier hierarchical caching schemes over 6G DT-enabled mobile networks.**

***Index Terms*—6G, mURLLC, DT, statistical delay and error-rate QoS bounded provisioning, ϵ-effective capacity, multi-tier caching, distributed computing, FBC.**

I. Introduction

THE next-generation wireless networks are expected to support emerging target monitoring and controlling applications, such as robotics and autonomous vehicles, as the key data traffics to guarantee mobile users' (MUs') stringent transmission delay and error-rate bounded quality-of-service (QoS). However, the complexity of these applications grows as they integrate automation and intelligence, and the associated management costs have already become a significant challenge for network operators. To address this issue, *digital twins (DT)* has been widely recognized as an emerging tool to enable the monitoring, controlling, or even the optimization of large-scale physical systems in real-time. The DT consists of (i) a physical entity/object, (ii) a virtual/digital representation corresponding to the physical entity/object, and (iii) the wireless communication links between the physical and virtual twins.

One of the main challenges for DT is how to establish two directional real-time and high-reliability communications links between the physical and virtual twins for achieving timely interactions between them. Due to the massive geographical coverage scale and complex environments of mobile users (MUs), the new challenges for disseminating DT data content items with stringent low-delay and low-error-probability requirements need to be overcome. To overcome these difficulties, *massive ultra-reliable and low-latency communications (mURLLC)* [1] has been proposed as a key technique to generate the ultra real-time, reliable, and high data-rate wireless networking environments for supporting immersive and inter-operable DT data streaming. On the other hand, the physical-virtual synchronization of DT for complex heterogeneous services, such as the metaverse and virtual reality (VR)/augmented reality (AR) in 6G, requires intensive computational operations and significant consumption of networking resources. To address these issues, *multiple-tier computing* (MTC) techniques [2–4] have also been developed to support DT data communications by providing distributed computation, processing, and storage capabilities at different tiers of wireless networks. Using MTC-based *edge computing architectures*, wireless networks store the frequently-demanded DT data at different wireless network tiers along network-edge devices to efficiently reduce DT streaming delay and data move [5]. Leveraging the 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

This work of Xi Zhang and Qixuan Zhu was supported in part by the U.S. National Science Foundation under Grants CCF-2142890, CCF-2008975, ECCS-1408601, and CNS-1205726.

979-8-3503-5471-3/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDCSW63686.2024.00015

Authorized licensed use limited to: Texas A M University. Downloaded on October 09,2024 at 02:54:49 UTC from IEEE Xplore. Restrictions apply.

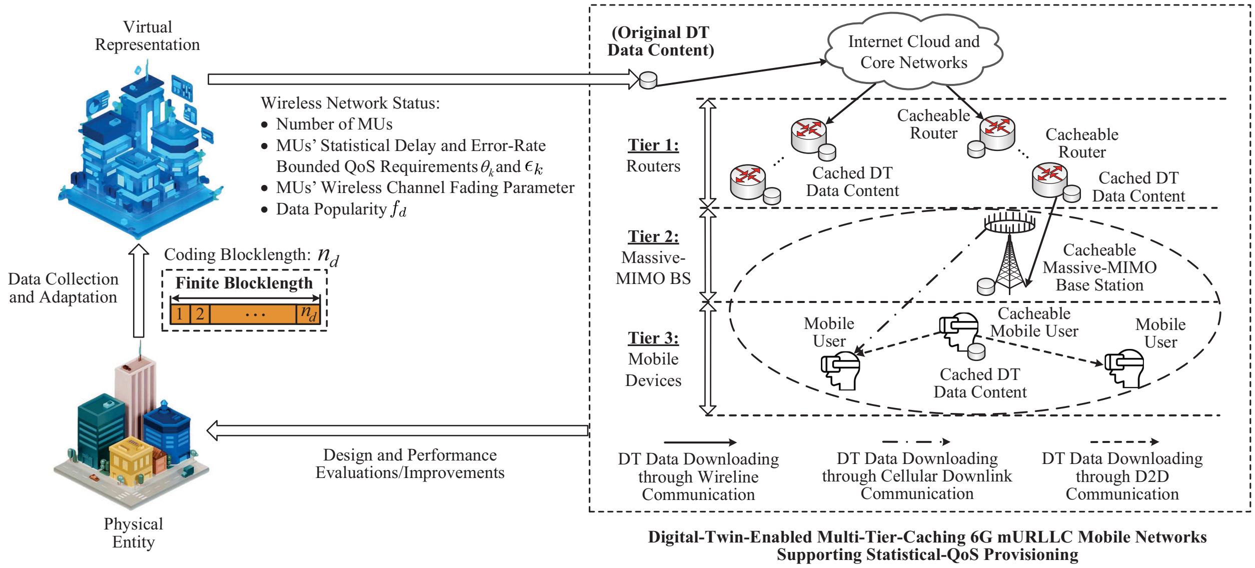
Virtual
Representation
Wireless Network Status:
• Number of MUs
• MUs' Statistical Delay and Error-Rate Bounded QoS Requirements θ_k and ϵ_k
• MUs' Wireless Channel Fading Parameter
• Data Popularity f_d
Coding Blocklength: n_d
Finite Blocklength
1 2 ⋯ n_d
Data Collection and Adaptation
Physical
Entity
Design and Performance Evaluations/Improvements
(Original DT Data Content)
Internet Cloud and Core Networks
Tier 1: Routers
Cacheable Router
Cacheable Router
Cached DT Data Content
Cached DT Data Content
Tier 2: Massive-MIMO BS
Cacheable Massive-MIMO Base Station
Tier 3: Mobile Devices
Mobile User
Cacheable Mobile User
Mobile User
Cached DT Data Content
DT Data Downloading through Wireline Communication
DT Data Downloading through Cellular Downlink Communication
DT Data Downloading through D2D Communication
Digital-Twin-Enabled Multi-Tier-Caching 6G mURLLC Mobile Networks Supporting Statistical-QoS Provisioning

Fig. 1. The distributed computing-based system architecture model for our proposed digital-twin-enabled multi-tier caching 6G mURLLC mobile wireless networks supporting statistical-QoS provisioning 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 (**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.

reduce the latency, thus improving QoS performances.

There exist some works investigating DT communications using distributed computations in 6G wireless networks. The work of [6] integrated DT communications with 6G wireless networks to migrate real-time data processing and computation to the edge plane, and then, developed a blockchain empowered federated learning framework running in the DT wireless networks for collaborative computing, improving the reliability and security of the system and enhancing data privacy. The authors of [7] proposed a fairness-aware latency mitigation framework in the DT aided edge computing with URLLC, which jointly optimizes various communications and computations parameters, namely, bandwidth allocation, transmission power, task offloading portions, and processing rate of user equipments and edge servers. A new paradigm DT network was developed by [8] for industrial Internet of Things (IIoT) systems through building network topology and the stochastic task arrival model. Then, the authors of [8] also formulated the stochastic computation offloading and resource allocation problem and minimized the long-term energy efficiency by leveraging Lyapunov optimization technique.

However, how to optimize the DT data items caching scheme using the distributed computing techniques to guarantee both delay and error-rate bounded QoS performance for DT applications still remain challenging open problems. To address these challenges, in this paper we develop the adaptive encoding scheme to flexibly collect the digital data from its physical entity and design the collaborative and distritbuted multi-tier hierarchical caching mechanisms to store the frequently downloaded DT data items at different tiers of wireless networks for guaranteeing their communication statistical-QoS. First, we propose the collaborative multi-tier hierarchical caching mechanisms to support the DT transmissions, where we selectively store highly demanded DT data content items at different network edge caching tiers (i.e., routers, massive multiple-input multiple-output (massive MIMO) base station (BS), and mobile devices, respectively) according to popularities of the DT data items. Second, we develop the ϵ-*effective capacity* to measure the performance metrics for wireless communications to support the statistical delay and error-rate bounded QoS provisioning for a cached DT data's transmission. We also optimize the packet length of a DT data item according to the network status and wireless channel conditions. Third, we propose the inter-tier collaborative caching mechanisms to maximize the aggregate ϵ-effective capacity as the DT data *caching gain* across three caching tiers, and propose the intra-tier optimal caching algorithm to minimize the total transmission delay for all cached DT data items at each caching tier, respectively.

The rest of this paper is organized as follows. Section II establishes system models for our proposed DT-enabled multi-tier caching architectures. Section III develops the DT data adaptive collection scheme to encode the physical twin into a data packet with the finite blocklength. Section IV proposes inter-tier and intra-tier collaborative hierarchical caching mechanisms. Section V validates and evaluates our developed schemes through numerical analyses. This paper concludes with Section VI.

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 consist of the following three major components: (1) Physical Entity, (2) Virtual Representation, and

Authorized licensed use limited to: Texas A M University. Downloaded on October 09,2024 at 02:54:49 UTC from IEEE Xplore. Restrictions apply.

(3) Digital-Twin-Enabled Multi-Tier-Caching 6G mURLLC Mobile Networks under Statistical-QoS Provisions. Along with real-time data collection, data adaptation captures the actual mobile network status (e.g., number of MUs, statistical QoS of MUs' requested DT data items) 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 [9] 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 [10]. Finally, 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*. 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.

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 (3) Tier 3: mobile devices. We propose to cache the DT data items at different caching tiers according to the data item's popularity (i.e., requested frequency). Denote by $f^{(\mathrm{T1})}$, $f^{(\mathrm{T2})}$, and $f^{(\mathrm{T3})}$ the popularity thresholds for Tier 1, Tier 2, and Tier 3, respectively, and we set $f^{(\mathrm{T1})} < f^{(\mathrm{T2})} < f^{(\mathrm{T3})}$, for ensuring that we cache a more popular data item at the cache station closer to MUs. If the DT data item is cached at Tier 3, an MU retrieves this data item using device-to-device (D2D) communications If the DT data item is cached at Tier 2, an MU retrieves this data item using cellular downlink communications from the massive MIMO BS. If the DT data item is cached at Tier 1, an MU retrieves this data by wireline communications among routers/BSs and then cellular downlink communications from the massive MIMO BS to the MU.

III. DT Data Adaptive Collection and Hierarchical Caching Schemes Based on MUs' Statistical QoS Requirements

We consider a cellular wireless network with K MUs, indexed by k, requesting a total number of D DT data items, indexed by d. Define $\mathcal{D} \triangleq \{1, 2, \ldots, D\}$ as the set of all DT data items.

A. Statistical Delay-Bounded QoS for DT Data Transmission

The statistical delay-bounded QoS guarantees [11, 12] have been employed in analyzing the stochastic arrival and service processes over the time-varying wireless fading channels. One of the key parameter in statistical delay-bounded QoS guarantees is the *QoS exponent* for MU $k, \forall k$, denoted by $\theta_k > 0$, measuring the exponential decaying rate of the delay QoS violation probabilities, which can be defined as follows:

$$\theta_k \triangleq - \lim_{Q_{\mathrm{th},k} \to \infty} \frac{\log(\Pr\{Q_k(\infty) > Q_{\mathrm{th},k}\})}{Q_{\mathrm{th},k}} \tag{1}$$

where $Q_k(t)$ is the queue-length process for the kth MU converging in distribution to a random variable $Q_k(\infty)$ and $Q_{\mathrm{th},k}$ is the queue length threshold (bound).

B. Error-Rate Bounded QoS for DT Data Transmission

According to the wireless channel state, we propose the DT data adaptive collection scheme to convert the physical entity into its virtual representation by applying the adaptive source encoding and data collecting. We encode the DT physical object into its digital twin applying finite blocklength coding (FBC) for mitigating the transmission delay while guaranteeing the error-rate bounded QoS. We define the DT data adaptive collection scheme in the following definition.

Definition 1: We consider a wireless fading channel for transmitting the DT data, which uses input blockcode set $\mathcal{X}$ and output blockcode set $\mathcal{Y}$. We define that an (n_d, D, ϵ_k)-code, $\forall k \in \{1, 2, \ldots, K\}, \forall d \in \{1, 2, \ldots, D\}$, for a state-dependent memoryless channel consists of [13]:

- A set of DT data messages $\mathcal{D}$, which is encoded into an input blockcode set $\mathcal{X} = \{c_1, c_2, \cdots, c_D\}$ (i.e., $\mathcal{D} \mapsto \mathcal{X}$) with a total number of D codewords;
- The dth DT data codeword c_d, which is a data packet with n_d bits (i.e., the packet length is n_d), where n_d is a finite integer, $\forall d \in \mathcal{D}$;
- A fading wireless channel, with blockcode set $\mathcal{X}$ and output blockcode set $\mathcal{Y}$ (i.e., $\mathcal{X} \mapsto \mathcal{Y}$), transmitting the DT data item d to the kth MU, $\forall k$;
- An estimated DT message, denoted by $\widehat{d}$, decoded by the MU k using its received codewords $\widehat{c}_d \in \mathcal{Y}$ (i.e., $\mathcal{Y} \mapsto \mathcal{D}$);
- The decoding-error probability of the kth MU, denoted by ϵ_k, which is defined as

$$\epsilon_k \triangleq \frac{1}{D} \sum_{d=1}^{D} \Pr\left\{ d \neq \widehat{d} \right\}, \tag{2}$$

Authorized licensed use limited to: Texas A M University. Downloaded on October 09,2024 at 02:54:49 UTC from IEEE Xplore. Restrictions apply.

where $d \in \mathcal{D} = \{1, \cdots, D\}$ is the index of DT data message, and $\widehat{d}$ is the estimated message for d at MU k, with $\widehat{d} \in \mathcal{D}$;

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

C. Statistical Delay and Error-Rate Bounded QoS Provisioning for DT Data Transmission

To guarantee the mURLLC for MUs, we apply the concept of ϵ-effective capacity [1] for both statistical *delay* and *error-rate* bounded QoS provisioning for our proposed DT wireless networks through the following definition.

Definition 2: The ϵ-effective capacity, denoted by $EC_{k,d}(\theta_k, \epsilon_k, P_k)$, for the kth MU is defined as the maximum data-packet'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, denoted by θ_k, of the delay-bound violation probability and the non-vanishing decoding-error probability, denoted by ϵ_k, respectively, under the transmit power allocation P_k, which is given as follows [14]:

$$EC_{k,d}(\theta_k, \epsilon_k, P_k) = -\frac{1}{n_d \theta_k} \log\left\{\epsilon_k + (1-\epsilon_k) e^{-n_d \theta_k R_k(P_k)}\right\} \tag{3}$$

where n_d is the length of the packet for the dth DT data item defined in Definition 1 and $R_k(P_k)$ is the data communication rate for the kth MU receiving its requested DT data under the power allocation P_k.

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 θ_k and decoding error probability ϵ_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 dth DT data-content item based on MUs' statistical QoS requirements. We propose the adaptive DT data collection scheme for optimizing the length of the packet n_d as follows:

$$n_d^* = \arg\max_{n_d} \left\{EC_{k,d}(\theta_k, \epsilon_k, P_k)\right\} \tag{4}$$

where n_d^* is the optimal n_d obtained by the DT data collection scheme. If the requested dth data item is cached at Tier 3, we set $P_k = P_{\rm D2D}$, where $P_{\rm D2D}$ is the transmit power of D2D communications when an MU receives the requested data from a peer mobile device. If the requested dth data item is cached at Tier 2, we set $P_k = P_{\rm BS}$, where $P_{\rm BS}$ is the transmit power when an MU receives the requested data from the massive MIMO BS. If the requested dth data item is cached at Tier 1, this data item is delivered using wireline communications from routers to the massive MIMO BS, and then, the massive MIMO BS forwards this data item to the MU using cellular downlink communication with transmit power $P_k = P_{\rm BS}$. Denoting by f_d the popularity of the dth DT data item, we propose the following **Algorithm 1** for our DT data items hierarchical caching scheme.

Algorithm 1 DT Data-Content Items' Hierarchical Caching Scheme

```
initialize: The popularity $f_d$ of the DT data content item $d$, popularity thresholds of Tier 1 $f^{(\rm T1)}$, Tier 2 $f^{(\rm T2)}$, and Tier 3 $f^{(\rm T3)}$, MU's statistical QoS requirements $\theta_k$ and $\epsilon_k$.
if $f_d$ is larger than the popularity threshold of Tier 3 (i.e., $f_d > f^{(\rm T3)}$) then
    The $d$th DT data-content item is cached at Tier 3.
    Adjust the optimal blocklength $n_d^*$ using Eq. (4) to maximize the $\epsilon$-effective capacity $EC_{k,d}(\theta_k, \epsilon_k, P_{\rm D2D})$
else if $f_d$ is less than the popularity threshold of Tier 3 but is larger than that of Tier 2 (i.e., $f^{(\rm T2)} < f_d < f^{(\rm T3)}$) then
    The $d$th DT data-content item is cached at Tier 2.
    Adjust the optimal blocklength $n_d^*$ using Eq. (4) to maximize the $\epsilon$-effective capacity $EC_{k,d}(\theta_k, \epsilon_k, P_{\rm BS})$.
else $f_d$ is less than the popularity threshold of Tier 2 but is larger than that of Tier 1 (i.e., $f^{(\rm T1)} < f_d < f^{(\rm T2)}$)
    The $d$th DT data-content item is cached at Tier 1.
    Adjust the optimal blocklength $n_d^*$ using Eq. (4) to maximize the $\epsilon$-effective capacity $EC_{k,d}(\theta_k, \epsilon_k, P_{\rm BS})$.
end if
output: The optimal encoding blocklength $n_d^*$ for DT data collection and adaptation and the optimal caching location.
```

IV. Inter-Tier and Intra-Tier Collaborative Hierarchical Caching Mechanisms

We propose both inter-tier and intra-tier collaborative hierarchical mechanisms for caching popular DT data items. The inter-tier collaborative hierarchical caching mechanisms aim at optimizing the caching lifespan of the cached DT data items. The intra-tier collaborative caching mechanisms aim at optimizing the wireless resources allocations (i.e., transmit power) within each caching tier.

A. Inter-Tier Collaborative Hierarchical Caching Mechanisms

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 item's caching lifespan is a function of its popularity. Denote by $L^{(\rm i)}(f_d), \forall {\rm i} \in \{\rm r, b, m\}$ the caching lifespan for the dth data item, which is a function of its popularity f_d, if this data item is cached at Tier 1, Tier 2, and Tier 3, respectively.

The inter-tier collaborative hierarchical caching mechanisms aim at optimizing the caching lifespan of each caching tier to maximize the aggregate ϵ-effective capacity of all MUs receiving the requested data items from all three caching tiers, which can be expressed as follows:

$$\max_{\substack{L^{(\rm r)}(f_d), L^{(\rm b)}(f_d)\\ L^{(\rm m)}(f_d)}} \left\{ \sum_{k=1}^{K} \left(\sum_{d \in \mathcal{D}^{(\rm r)}} EC_{k,d}(\theta_k, \epsilon_k, P_{\rm BS}) + \sum_{d \in \mathcal{D}^{(\rm b)}} EC_{k,d}(\theta_k, \epsilon_k, P_{\rm BS}) + \sum_{d \in \mathcal{D}^{(\rm m)}} EC_{k,d}(\theta_k, \epsilon_k, P_{\rm D2D}) \right) \right\} \tag{5}$$

Authorized licensed use limited to: Texas A M University. Downloaded on October 09,2024 at 02:54:49 UTC from IEEE Xplore. Restrictions apply.

$$\text{s.t.: } L^{(\mathrm{r})}(f_d), L^{(\mathrm{b})}(f_d), L^{(\mathrm{m})}(f_d) \le L^{\max}$$

where $EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{D2D}})$ can be obtained by using Eq. (3) and replacing P_k by P_{D2D}; $EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{BS}})$ can be obtained by using Eq. (3) and replacing P_k by P_{BS}; $\mathcal{D}^{(\mathrm{r})} \in \mathcal{D}$, $\mathcal{D}^{(\mathrm{b})} \in \mathcal{D}$, and $\mathcal{D}^{(\mathrm{m})} \in \mathcal{D}$ are three disjoint sets of DT data items cached at Tier 1, Tier 2, and Tier 3, respectively; and $L^{\max}$ is the maximum caching lifespan.

B. Intra-Tier Collaborative Caching Mechanisms

Based on Definition 2, ϵ-effective capacity is the maximum data constant arrival rate under a given delay exponent and a given decoding error probability. Thus, the transmission delay, denoted by $\omega_d^{(\mathrm{r})}$, $\omega_d^{(\mathrm{b})}$, and $\omega_d^{(\mathrm{m})}$, for transmitting the dth DT data item to the kth MU if this data item is cached at Tier 1, Tier 2, and Tier 3, respectively, are given by

$$\begin{cases} \omega_d^{(\mathrm{r})} = \omega_{\mathrm{R}}(f_d) + \dfrac{n_d^*}{EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{BS}})} \\ \omega_d^{(\mathrm{b})} = \dfrac{n_d^*}{EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{BS}})} \\ \omega_d^{(\mathrm{m})} = \dfrac{n_d^*}{EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{D2D}})} \end{cases} \tag{6}$$

where n_d^* is obtained by using Eq. (4), and $\omega_{\mathrm{R}}(f_d)$ is the transmission delay for wireline communications among routers and BSs depending on the data item's popularity f_d. Note that under the adaptive blocklength scheme, ϵ-effective capacities $EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{D2D}})$ and $EC_{k,d}(\theta_k, \epsilon_k, P_{\mathrm{BS}})$ are functions of the blocklength n_d according to Eq. (3). Thus, we propose the intra-tier collaborative caching schemes to minimize the sum of data transmission delay for all cached DT data items within each caching tier as follows:

$$\begin{cases} \min\limits_{P_{\mathrm{BS}}} \left\{ \sum\limits_{d \in \mathcal{D}^{(\mathrm{r})}} \omega_d^{(\mathrm{r})} \right\}, \text{ if the } d\text{th data item is cached at } \underline{\text{Tier 1}} \\ \qquad \left(\text{i.e., } f^{(\mathrm{T1})} \le f_d < f^{(\mathrm{T2})}\right), \\ \min\limits_{P_{\mathrm{BS}}} \left\{ \sum\limits_{d \in \mathcal{D}^{(\mathrm{b})}} \omega_d^{(\mathrm{b})} \right\}, \text{ if the } d\text{th data item is cached at } \underline{\text{Tier 2}} \\ \qquad \left(\text{i.e., } f^{(\mathrm{T2})} \le f_d < f^{(\mathrm{T3})}\right), \\ \min\limits_{P_{\mathrm{D2D}}} \left\{ \sum\limits_{d \in \mathcal{D}^{(\mathrm{m})}} \omega_d^{(\mathrm{m})} \right\}, \text{ if the } d\text{th data item is cached at } \underline{\text{Tier 3}} \\ \qquad \left(\text{i.e., } f_d \ge f^{(\mathrm{T3})}\right), \end{cases} \tag{7}$$

where $\mathcal{D}^{(\mathrm{r})}$, $\mathcal{D}^{(\mathrm{b})}$, and $\mathcal{D}^{(\mathrm{m})}$ are disjoint sets of DT data items cached at Tier 1, Tier 2, and Tier 3, respectively; and $f^{(\mathrm{T1})}$, $f^{(\mathrm{T2})}$, and $f^{(\mathrm{T3})}$ are defined in Section II.

V. Performance Evaluations

In Fig. 2, we compare the ϵ-effective capacity $EC_{k,d}\left(\theta_k, \epsilon_k, P_{\mathrm{BS}}\right)$ under the MIMO baseline BS technique with our proposed massive MIMO BS techniques. We set the number of antennas as $M_T = 10$ for the MIMO baseline BS, and set the number of antennas as $M_T = 80, 100, 120$, respectively, for our proposed massive MIMO BS techniques. We also set the number of antennas on each MU as $M_R = 2$, the delay QoS exponent as $\theta_k = 10^{-1}$, the decoding error probability as $\epsilon_k = 10^{-4}$. We also assume that the wireless channel follows the Nakagami-m fading with the fading parameter $m = 2$. Figure 2 shows that the ϵ-effective capacity $EC_{k,d}\left(\theta_k, \epsilon_k, P_{\mathrm{BS}}\right)$ is a monotonically increasing function of the number of antennas M_T on the MIMO baseline BS/massive MIMO BS under the same BS transmit power allocation, implying that a larger M_T can improve the communication performance.

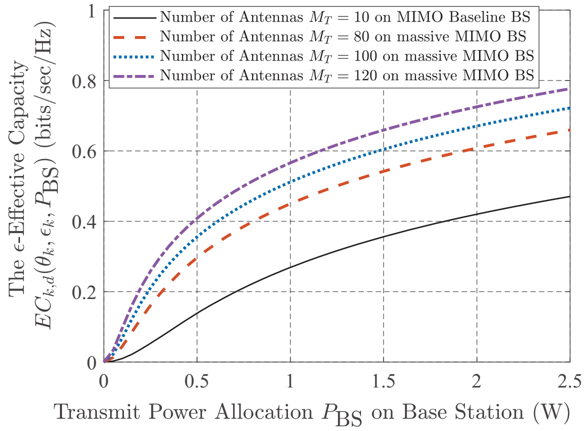

Fig. 2. Comparisons of the ϵ-effective capacity $EC_{k,d}\left(\theta_k, \epsilon_k, P_{\mathrm{BS}}\right)$ under the MIMO baseline BS schemes with our proposed massive MIMO BS schemes.

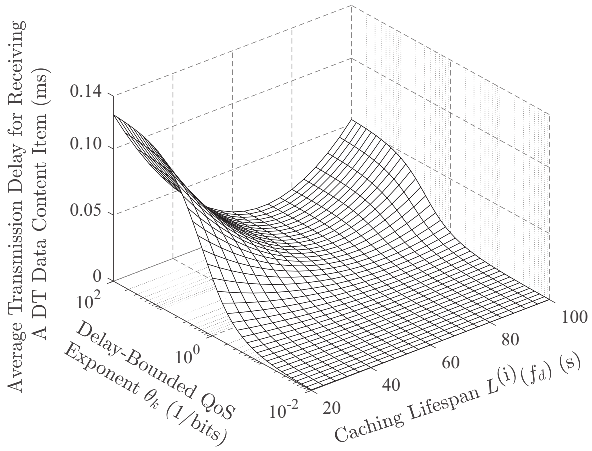

Fig. 3. Average transmission delay for receiving a DT data content item under different values of caching lifespan.

In Fig. 3, we show the average transmission delay for receiving a DT data content item under our proposed multi-tier caching scheme under different values of the caching lifespan $L^{(\mathrm{i})}(f_d), \forall \mathrm{i} \in \{\mathrm{r}, \mathrm{b}, \mathrm{m}\}$. 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 is 0.05 ms if an MU receives its requested DT data items using D2D communications, and we set that transmission delays are 0.02 ms and 0.25 ms if an MU receives its requested DT data

Authorized licensed use limited to: Texas A M University. Downloaded on October 09,2024 at 02:54:49 UTC from IEEE Xplore. Restrictions apply.

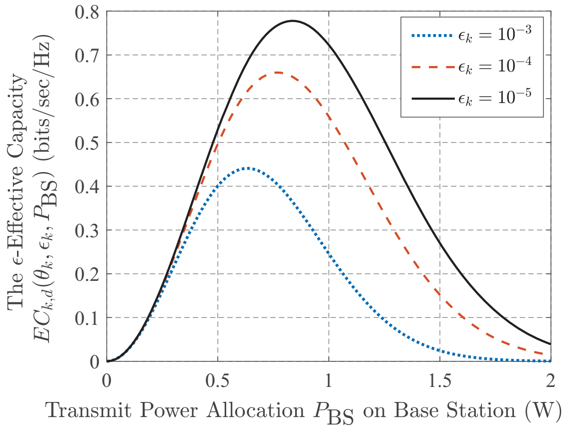

Fig. 4. Comparisons of the ϵ-effective capacity $EC_{k,d}(\theta_k, \epsilon_k, P_{\text{BS}})$ under different values of the decoding error probability ϵ_k and the transmit power allocation P_{BS}.

items from a BS and a router, respectively. If a data item is cached in a cache station of three caching tiers and is within its caching lifespan, this data item can be delivered to an MU from this cache station. Otherwise, if the data item is deleted after its caching lifespan, this data item will be downloaded through Internet cloud and core networks, yielding a larger transmission delay. We can observe from Fig. 3 that there always exists an optimal caching lifespan that minimizes the average transmission delay. Since the transmission delay is a decreasing function of the ϵ-effective capacity, the existence of the average transmission delay indicates the existence of the optimal solution for Eq. (5). We can also observe from Fig. 3 that transmission delay monotonically increases as the delay-bounded QoS exponent θ_k increases. This is because θ_k measures the stringency of the delay-bounded QoS and thus a channel with a smaller θ_k (i.e., loose delay-bounded QoS provisions) can support a larger data arrival rate, resulting in a larger ϵ-effective capacity and thus a smaller average transmission delay.

In Fig. 4, we compare the ϵ-effective capacity $EC_{k,d}(\theta_k, \epsilon_k, P_{\text{BS}})$ of the kth MU for our proposed multi-tier caching scheme under different values of the decoding error probability ϵ_k and transmit power allocation P_{BS}. We set that blocklength $n_d = 1000$ and delay-bounded QoS exponent $\theta_k = 10^{-2}$. We can observe from Fig. 4 that there always exists an optimal transmit power P_{BS} that maximizes the ϵ-effective capacity $EC_{k,d}(\theta_k, \epsilon_k, P_{\text{BS}})$. This implicates that there always exists an optimal transmit power minimizing each transmission delay in Eq. (6), since the transmission delay is a monotonically decreasing function of the ϵ-effective capacity. Therefore, we are able to derive the optimal solution for the sum of data transmission delay given by Eq. (7). We can also observe that for the same power allocation, the aggregate ϵ-effective capacity increases as the decoding error probability ϵ_k decreases. This is because the smaller ϵ_k indicates a better wireless channel quality, and thus, the wireless network can achieve a larger ϵ-effective capacity.

VI. Conclusions

We have proposed collaborative multi-tier hierarchical caching mechanisms for DT over mURLLC-based 6G mobile wireless networks, where a popular DT data item can be cached at different wireless network caching tiers according to its popularity. We have developed the adaptive data collection scheme for encoding the physical twin into its virtual twin. Moreover, we have designed an inter-tier collaborative hierarchical caching scheme, which maximizes the aggregate ϵ-effective capacity across all three caching tiers. Then, we have also developed the intra-tier collaborative hierarchical caching scheme to minimize the transmission delay for downloading a cached DT data item from each caching tier through optimizing the wireless resources allocation (i.e., transmit power) according to the wireless channel condition.

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 6G metaverse mobile networks using FBC," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 3, pp. 570–587, 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 Transactions on Communications*, vol. 70, no. 3, pp. 1820–1833, March 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 Journal on Selected Areas in Communications*, vol. 41, no. 7, pp. 2000–2012, 2023.

[4] X. Zhang and Q. Zhu, "Collaborative hierarchical caching over 5G edge computing mobile wireless networks," in *Proc. IEEE International Conference on Communications (ICC) 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 Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2230–2254, Fourthquarter 2022.

[6] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Low-latency federated learning and blockchain for edge association in digital twin empowered 6G networks," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5098–5107, 2021.

[7] D. Van Huynh, V.-D. Nguyen, S. R. Khosravirad, G. K. Karagiannidis, and T. Q. Duong, "Distributed communication and computation resource management for digital twin-aided edge computing with short-packet communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 10, pp. 3008–3021, 2023.

[8] Y. Dai, K. Zhang, S. Maharjan, and Y. Zhang, "Deep reinforcement learning for stochastic computation offloading in digital twin networks," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4968–4977, 2021.

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

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

[11] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 118–129, Jan. 2008.

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

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

[14] X. Zhang and Q. Zhu, "Statistical delay and error-rate bounded QoS provisioning over massive-MIMO based 6G mobile wireless networks," in *Proc. IEEE Global Communications Conference (GLOBECOM) 2022*, 2022, pp. 353–358.

Authorized licensed use limited to: Texas A M University. Downloaded on October 09,2024 at 02:54:49 UTC from IEEE Xplore. Restrictions apply.