Hierarchical Caching for Statistical QoS Guaranteed Multimedia Transmissions over 5G Edge Computing Mobile Wireless Networks

Xi Zhang and Qixuan Zhu

Abstract

The 5G mobile wireless networks are expected to provision distinct delay-bounded QoS guarantees for a wide range of multimedia services, applications, and users with extremely diverse requirements. On the other hand, how to efficiently support multimedia services over 5G wireless networks has imposed many new challenging issues not encountered before in 4G wireless networks. Bringing data contents closer to mobile users, the caching-based content-centric edge computing network is a promising candidate network architecture and technique to efficiently guarantee QoS for time-sensitive multimedia transmissions over 5G mobile wireless networks. However, how to collaboratively integrate in-network caching and edge computing for multimedia transmissions over 5G wireless networks has been neither well understood nor thoroughly studied. To remedy these difficulties, in this article we propose hierarchical-caching-based content-centric network architectures and their three different implementation/control mechanisms over 5G edge computing mobile multimedia wireless networks, where popular multimedia data contents can be cached at different network tiers (e.g., routers, base stations, and mobile devices) to guarantee the statistical delay-bounded QoS for multimedia transmissions while minimizing redundant transmissions. We propose three hierarchical edge caching mechanisms: random hierarchical caching, which maximizes the average effective capacity based on the users' request frequency (i.e., data popularity); proactive hierarchical caching, which maximizes the cache hitting rate by predicting the popularity of data contents and caching the data contents with high popularity nearby mobile users; and game-theory-based hierarchical caching, where three network caching tiers are formulated as three game players in a cooperative game to maximize their aggregate effective capacity under the constraints of caching expenses at each tier. We develop three concrete algorithms to implement and control the three hierarchical caching mechanisms, respectively, for statistical delay-bounded QoS guaranteed multimedia transmissions. Using simulations and numerical analyses, we show that our proposed three hierarchical caching mechanisms significantly outperform other existing schemes in supporting the statistical delay-bounded QoS provisioning over 5G edge computing mobile wireless networks.

Introduction

With the rapid maturing of fourth generation (4G) standardization and the ongoing worldwide deployment of 4G cellular networks, extensive research activities on fifth generation (5G) mobile wireless communication technologies have emerged in both the academic and industrial communities. The main objective of upcoming 5G wireless networks focuses on the enabling techniques and mechanisms to ensure that various contemporary wireless applications can be promptly and satisfactorily accomplished at any time and any place, and in any manner [1, 2]. One of the most important services in 5G wireless applications lies in supporting rich multimedia services, including video conferences, mobile TV, music downloading, online gaming, and even 3D immersive media transmissions. These wireless multimedia services are all bandwidth-/computation-intensive and time-sensitive, and thus used to be confined to wired networks such as the Internet, but are now making forays into mobile devices and wireless networks. Mobile video streaming currently accounts for almost 70 percent of mobile data traffic and is expected to have a 500-fold increase over the next 10 years. Full high definition (FHD) video has also been widely delivered through popular media such as YouTube, and ultra high definition (UHD) and 3D video content will eventually take over in the not so distant future.

However, supporting delay-bounded quality of service (QoS) [3] for these highly bandwidth-intensive and time-sensitive multimedia services over the emerging 5G wireless networks with constrained wireless resources imposes many new challenging problems not encountered in 4G networks. For the last several years, telecommunications academia and industry have made a plethora of efforts on investigating various advanced and sophisticated wireless network architectural techniques such as *content-centric networks* [4–8] for guaranteeing delay-bounded QoS provisioning. The critical design issue for multimedia wireless services is how to efficiently guarantee timely multimedia data transmissions within specified delay bounds under the constrained radio spectrum and power supply. Because of the highly time-varying wireless channels, the *deterministic* delay-bounded QoS requirements for high-volume and diverse multimedia wireless traffic are usually hard to guar-

Digital Object Identifier:
10.1109/MWC.2018.1700327

Xi Zhang (corresponding author) and Qixuan Zhu are with Texas A&M University.

1536-1284/18/$25.00 © 2018 IEEE

antee. Alternatively, the *statistical* delay-bounded QoS provisioning theory has been proposed and shown to be a powerful technique to characterize and implement the delay-bounded QoS guarantee for wireless real-time traffic [9].

The caching-based content-centric network is among the most promising candidate techniques and architectures to support the statistical QoS guarantees for time-sensitive multimedia transmissions over 5G edge computing wireless networks by bringing data contents nearer to mobile users. To support multimedia data transmissions in 5G edge computing mobile wireless networks, in this article we propose three hierarchical caching mechanisms, each of which caches popular data contents at three different caching tiers of the edge of wireless networks: *Tier 1*: routers, *Tier 2*: cellular base station (BS) or WiFi access point (AP), and *Tier 3*: mobile devices, to guarantee the statistical delay-bounded QoS requirements while minimizing duplicate multimedia traffic in core networks. Since the caching schemes at these three wireless network caching tiers have their corresponding advantages and disadvantages, in terms of caching expenses, memory space constraints, power and spectrum efficiencies, caching coverage areas, and so on, we develop three hierarchical caching implementation mechanisms to characterize and compare their performance trade-offs of the caching schemes at distinct caching tiers. These three hierarchical caching mechanisms aim to address different caching purposes/goals under our proposed edge-computing-based wireless network framework, which are summarized as follows.

Random Hierarchical Caching Mechanism: This optimizes the three tiers' hierarchical caching mechanism by considering each data content to be randomly cached at each caching tier with a given probability depending on the data content's popularity, aiming to maximize the average effective capacity over three caching tiers at the wireless network edge. The advantage of our developed random hierarchical caching mechanism is that the three tiers' caching probabilities are jointly optimized to maximize the average effective capacity of a data content. On the other hand, in this mechanism, the cache hitting rate as one of the important criteria in content-centric networks is not considered.

Proactive Hierarchical Caching Mechanism: This maximizes the overall cache hitting rate, as the advantage over the random hierarchical caching mechanism, for the three caching tiers of the wireless network edge by estimating the popularity of each data content and proactively caching the data content with high popularity in the nearby neighborhood of mobile users. However, the disadvantage of this mechanism is that it needs to record the mobile users' request histories to derive the accurate data popularity estimation, resulting in extra memory cost.

Game-Theory-Based Hierarchical Caching Mechanism: This formulates the caching schemes at three caching tiers as a cooperative game in order to enable each caching-tier player to bid for the number of data contents and the caching lifespan under the caching expense constraint, maximizing the aggregate effective capacities for all three caching tiers. As an advantage over the above two mechanisms, the game-theory based hierarchical caching mechanism can maximize the overall payoff for the three caching tiers through optimizing the caching-tier players' cooperative bidding strategies. However, this mechanism needs more time to achieve the convergence to the Nash equilibrium than the above two mechanisms during the transient stage of the convergence.

The rest of this article is organized as follows. The following section establishes the system models for our proposed content-centric-based hierarchical caching models in 5G edge computing wireless networks. Following that, we develop the random hierarchical caching mechanism and its implementation algorithm. Then we propose the proactive hierarchical caching mechanism and its implementing algorithm. After that, we design the cooperative-game-based hierarchical caching mechanism and its implementation algorithms. Then we evaluate the performance of our proposed schemes through simulations and numerical analyses. The article concludes in the final section.

The System Models

We propose the hierarchical-caching-based content-centric architectures over 5G edge computing mobile multimedia wireless networks, which cache the frequently requested multimedia data contents at the edge of the wireless networks to guarantee the statistical delay-bounded QoS requirements and minimize the redundant data transmissions in the Internet cloud and core networks. Figure 1 shows the system model for our proposed hierarchical caching architecture for statistical QoS guaranteed multimedia transmissions over 5G edge computing mobile wireless networks. As shown in Fig. 1, the edge of the wireless networks consists of the following three hierarchical and non-overlapped caching tiers [10]: (1) Tier 1: routers, (2) Tier 2: cellular BS and WiFi APs, and (3) Tier 3: mobile devices. We summarize the features, advantages, and disadvantages for caching schemes in these three tiers as follows:

Tier 1: Caching at Routers: The Tier 1 caching scheme caches data contents in network routers, which have the advantage of large memory space, low caching expense, and broad wireless network area coverage. On the other hand, in this caching scheme, each cached data content needs to be delivered to mobile users through a cellular BS or WiFi AP, which is constrained by the cellular network bandwidth, WiFi coverage area, and queueing length on the multihop delivery path, imposing difficulty in guaranteeing the transmission delay.

Tier 2: Caching at Cellular BS or WiFi APs: Using the caching scheme at Tier 2, data contents are cached at the cellular BS or WiFi AP of a wireless cell, where the cached data contents have a shorter delivery path than caching at routers on Tier 1, reducing the number of hops on the delivery path. Other advantages include large memory space, low caching expense, and high transmit power of data. If caching at the WiFi AP, this scheme saves the precious spectrum resource of the cellular network and provides better connection quality than cellular communication. On the other hand, if data contents are cached in the cellular BS, transmission rate is constrained by cellular network bandwidth, which imposes the multimedia data's delay-bounded QoS guarantee problems for caching at Tier 2. Moreover, this

Our proposed random hierarchical-caching mechanism characterizes the dynamics of data contents entering and leaving the caches of each tier by the Markov Chain model with exponential arrival and departure processes and characterizes the caching system state as the number of cached data contents at each tier.

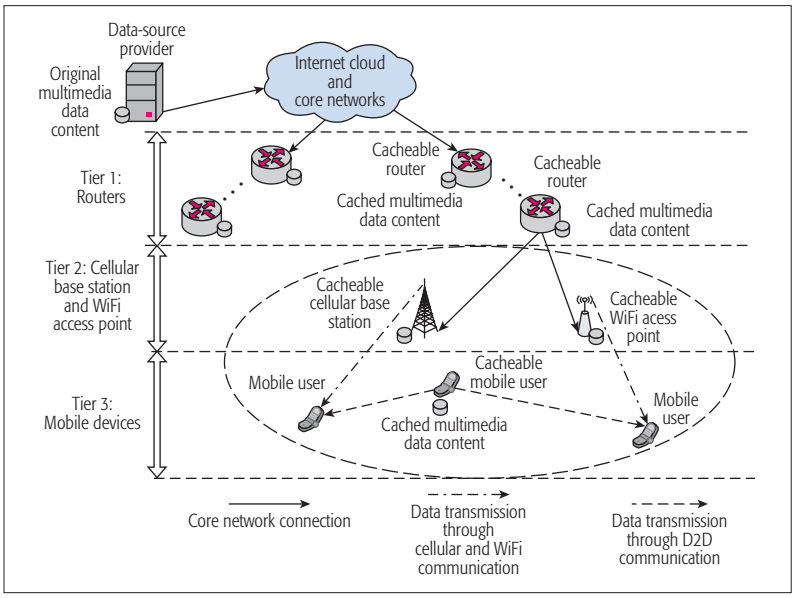

FIGURE 1. The framework of hierarchical caching in 5G edge computing wireless networks: caching in routers, cellular-BS/WiFi-APs, and mobile devices.

caching scheme is constrained by the cellular BS or WiFi AP coverage area, which may need relay nodes to forward the cached data contents to the data-requesting mobile users.

Tier 3: Caching at Mobile Devices: Using the caching scheme in mobile devices, mobile users can directly download their requested data contents from peer mobile users who hold the same contents. This direct data content exchange between peer mobile users in a device-to-device (D2D) communication (bypassing BS) fashion leads to multi-fold advantages. Caching through mobile devices can:

- Avoid broadcast transmissions/traffic, which significantly reduces interference and power consumption
- Further reduce transmission delays
- Increase the area coverage of data contents delivery through opportunistic contacts
- Reduce redundant transmissions and improve the scalability of networks

These advantages all play very important roles in supporting multimedia wireless transmissions over 5G mobile wireless networks. But the caching expense for caching at Tier 3 is higher than the above two schemes due to the limited caching memory and battery of each mobile device.

To characterize, compare, and balance the performance trade-offs among these three caching schemes, we propose three hierarchical caching mechanisms with each mechanism integrating the three caching schemes in routers, cellular-BS/WiFi-APs, and mobile devices. Our proposed three hierarchical caching mechanisms dynamically cache multimedia data contents at all three tiers at the edge of wireless networks according to different goals under different problem formulations; however, they are closely related and connected due to sharing the same system model of layered cache stations/hardware and addressing common caching problems:

- What to cache: Which data content is selected to cache?
- Where to cache: Which caching tier caches data contents?
- How long to cache: How long is the data content's caching lifespan?

Random Hierarchical-Caching Mechanism and Algorithm

Our proposed random hierarchical caching mechanism characterizes the dynamics of data contents entering and leaving the caches of each tier by the Markov chain model with exponential arrival and departure processes and characterizes the caching system state as the number of cached data contents at each tier. The random hierarchical caching mechanism aims at maximizing the average effective capacity by deriving the optimal data contents departure rates of leaving each tier's caches. We apply the effective capacity theory [9, 11] to characterize this statistical delay-bounded QoS provisioning, in upper-bounding the probability of delay bound violation in multimedia wireless transmissions. In the effective capacity theory, a transmitter needs to limit its traffic rate to a certain maximum value in order to ensure that the receivers's delay-bound violation probability is no more than a given value, which is the receiving node's delay-bounded QoS requirement.

Random Hierarchical Caching Mechanism

In the random hierarchical caching mechanism, we consider the scenario that popular multimedia data contents are randomly cached at each wireless network tier with given probabilities. Once the popularity of a multimedia data content is larger than or equal to a pre-defined threshold, it enters the caches of routers (i.e., Tier 1), cellular-BS/WiFi-APs (i.e., Tier 2), or mobile devices (i.e., Tier 3) following the Poisson process with arrival rates λ_r, λ_b, and λ_d, respectively. According to each caching tier's caching capacity and caching expense, the data contents stored at different tiers have different caching lifespans, where after this lifespan period the cache stations delete their cached data contents to replace the old data contents with new ones. Thus, data contents depart their cached tiers due to the expiration of their caching lifespans. Based on these different lifespans, data contents depart the caches following the exponential distribution with departure rates μ_r, μ_b, and μ_d ($\mu_r \neq \mu_b \neq \mu_d$), for caching at routers, cellular-BS/WiFi-APs, and mobile devices, respectively. Since the cached data contents are entering and leaving the three caching tiers randomly all the time following the above defined arrival and departure processes, the total number of cached data contents on each caching tier varies randomly over time. We can model this random caching mechanism at three caching tiers as a Markov stochastic process. To quantitatively model and solve this stochastic problem, we formulate this random hierarchical caching mechanism by a 3D Markov chain model as shown in Fig. 2 with the system state being denoted by (s_r, s_b, s_d), where s_r is the number of data contents cached at routers on Tier 1, s_b is the number of data contents cached at cellular BS or WiFi APs on Tier 2, and s_d is the number of data contents cached at mobile devices on Tier 3.

We define random variables T_r, T_b, and T_d as caching lifespans for caching in routers, cellular BS

or WiFi APs, and mobile devices, respectively, and notice that T_r, T_b, and T_d follow the exponential distributions with departure rates μ_r, μ_b, and μ_d, respectively. Let π_r, π_b, and π_d be the probabilities of a data content to be cached at Tier 1, Tier 2, and Tier 3, respectively, which are derived from the 3D Markov chain model. Since the transmissions for multimedia data contents need to guarantee the statistical delay-bounded QoS, we characterize the allowed maximum delay bound variation probability for each tier by θ_r, θ_b, and θ_d, respectively. The effective capacity under the QoS measurement in terms of QoS exponents θ_i, $\forall i \in \{r, b, d\}$, is denoted by $E(\theta_i)$, and each tier's caching expense under the lifespan of T_i, $\forall i \in \{r, b, d\}$, is denoted by $c_i(T_i)$. The objective for our random hierarchical caching mechanism is to optimize the departure rates μ_r, μ_b, and μ_d to maximize the average effective capacity under the caching expense constraint $c^{\max}$ given as follows:

$$\max_{\mu_r,\mu_b,\mu_d} \{\pi_r E(\theta_r) + \pi_b E(\theta_b) + \pi_d E(\theta_d)\}$$
$$\text{subject to: } C1: \pi_r c_r(T_r) + \pi_b c_b(T_b) + \pi_d c_d(T_d) \le c^{\max},$$
$$C2: \pi_r + \pi_b + \pi_d = 1, \quad (1)$$

where C1 and C2 are two constraints for the optimization problem specified in Eq. 1.

The Algorithm to Implement Random Hierarchical-Caching Mechanism

To solve Eq. 1, we need to derive the probabilities expressions of π_r, π_b, and π_d for data contents being cached at Tier 1, Tier 2, and Tier 3, respectively. We also need to derive their corresponding effective capacities $E(\theta_r)$, $E(\theta_b)$, and $E(\theta_d)$, corresponding to π_r, π_b, and π_d, in these three tiers, respectively. Because the derivations of effective capacities $E(\theta_r)$, $E(\theta_b)$, and $E(\theta_d)$ are already given in [9, 11], in this random hierarchical caching mechanism, we only need to derive π_r, π_b, and π_d by applying the probability generating function [12, 13] as shown in Algorithm 1.

Proactive Hierarchical Caching Mechanism and Algorithm

As one of the important criteria to measure the efficiency of a cache mechanism, *cache hitting rate* is defined as the rate at which a requested data content by mobile users can be found within the wireless network's cache stations at any caching tier. To maximize the cache hitting rate and the average effective capacity over all multimedia traffic transmissions, the wireless network needs to predict which multimedia data contents will be requested in the near future and thus to proactively cache the data contents with likely highest requested probability at nearby mobile users. We propose the proactive hierarchical caching mechanism to estimate the requested probability for each data content so that we can proactively cache the most frequently requested data contents at the nearest caching tier to mobile users while caching the less frequently requested data contents at the relatively distant caching tiers to mobile users. Using this proactive hierarchical caching mechanism, we can maximize the cache hitting rate with the overall statistical delay-bounded QoS requirement guaranteed.

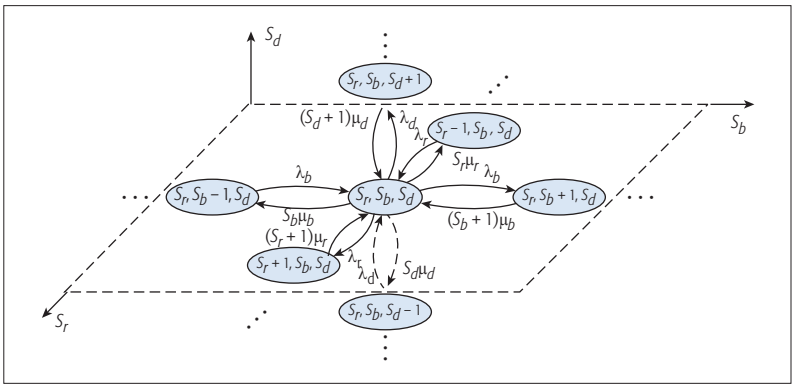

FIGURE 2. The three-dimensional Markov chain model for modeling the three hierarchical caching in routers, cellular-BS/WiFi-APs, and mobile devices. To simplify the diagram, self-transitions and their probabilities are omitted in this figure.

Proactive Hierarchical Caching Mechanism

The current sophisticated networks can exploit users' context information, anticipate users' demands, and leverage their predictive abilities to achieve significant resource saving to guarantee statistical QoS requirements and cost/energy expenditures. Endowed with this predictive ability, users are scheduled in a more efficient manner, and resources are pre-allocated more intelligently by proactively serving predictable peak-hour demands during off-peak times [15]. That is, when the multimedia wireless network serves mobile users' requests before their deadlines, the corresponding multimedia data contents are stored in the network cacheable nodes, and when a request is actually initiated, the cached data content is pulled out directly from the cache stations instead of accessing the core networks. For this purpose, our proposed techniques should be developed to seek optimal trade-offs between predictions that result in content being retrieved that users ultimately never request and requests not anticipated in a timely manner. Hence, the objective of our caching mechanism is to predict, anticipate, and infer future events in an intelligent manner, which is a complex problem. To overcome this challenge, other mobile users' requests as well as their social relationships can be leveraged to build reliable statistical models.

We propose to first determine a given mobile user's *influential user set*, which is a set consisting of the highest correlation users for this given mobile user, according to four centrality metrics: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality [15]. For example, the influential user set can be a set of users who share common hobbies, ages, genders, education backgrounds, and so on, who are connected by biological relationships, and thus, the multimedia data contents in which they are interested may hit the same dataset. Then we predict the probability distribution for this given user's multimedia data content requests based on the request histories of the users in the influential user set so that we proactively cache this user's possible requested contents in different caching tiers according to its request's probability distribution. To predict the probability distribution of a mobile user's request, we model the data contents' requests as a Chinese restaurant process, which characterizes customers

Step 1: According to Fig. 2, derive the balance equations [14] for the three-dimensional Markov Chain model with caching schemes on Tier 1, Tier 2, and Tier 3, respectively, where the summation of probabilities for entering each state is equal to the summation of probabilities for leaving this state.
Step 2: Define the probability generating function [12]

$$G_i(z) = \sum_{q=1}^{s_i^{\max}} \pi_i z^q, \forall i \in \{r, b, d\},$$

and its first order derivative $G'_i(z) = \partial G_i(z)/\partial z$, where z is the parameter in z-transform and $s_i^{\max}$ is the maximum caching space in the caching tier i such that $s_i \le s_i^{\max}$.
Step 3: Multiply the balance equations derived in **Step 1** by z^q and add them up to obtain a differential equation.
Step 4: Solve the differential equation in **Step 3** to obtain the explicit expressions of $G_i(z)$ and $G'_i(z)$, $\forall i \in \{r, b, d\}$, respectively.
Step 5: Based on [12], compute the average number $\mathbb{E}_r\{q\}$ of data contents cached at Tier 1, which is equal to $G'_r(1)$; compute the average number $\mathbb{E}_b\{q\}$ of data contents cached at Tier 2, which is equal to $G'_b(1)$; and compute the average number $\mathbb{E}_d\{q\}$ of data contents cached at Tier 3, which is equal to $G'_d(1)$.
Step 6: Define $\mathbb{E}\{q\} \triangleq (\mathbb{E}_r\{q\} + \mathbb{E}_b\{q\} + \mathbb{E}_d\{q\})$ as the total number of data contents cached at all three tiers. We have $\pi_r = (\mathbb{E}_r\{q\}/\mathbb{E}\{q\})$, $\pi_b = (\mathbb{E}_b\{q\}/\mathbb{E}\{q\})$, and $\pi_d = (\mathbb{E}_d\{q\}/\mathbb{E}\{q\})$.
Step 7: Plug the derived π_r, π_b, and π_d into Eq. 1, and implement Linear Programming, we can obtain the optimal solutions μ_r, μ_b, and μ_d of Eq. 1.

ALGORITHM 1. Random hierarchical caching implementation algorithm.

1: **initialize:** data content g is only stored in the original data-source provider, $p(g) = 0$, $l_g(u) = 0$
2: **while** calculate $p(g)$ for the gth data content to predict its popularity **do**
3: **if** $p_r \le p(g) < p_b$ **then**
4: cache/refresh data content g in a router at Tier 1
5: update $l_g(u)$ and delete data content g from previous cache
6: **else if** $p_b \le p(g) < p_d$ **then**
7: cache/refresh data content g in a cellular BS or WiFi AP at Tier 2
8: update $l_g(u)$ and delete data content g from previous cache
9: **else if** $p(g) > p_d$ **then**
10: cache/refresh data content g in a mobile device at Tier 3
11: update $l_g(u)$ and delete data content g from previous cache
12: **else**
13: set $l_g(u) = 0$ and delete data content g from any caches at three caching tiers
14: **end if**
15: **end while**

ALGORITHM 2. Proactive hierarchical caching implementation algorithm.

seating stochastic process in a Chinese restaurant with infinite number of tables and infinite number of seats for each table. In the Chinese restaurant process, the first customer sits at the first table. Other customers choose tables one after another, choosing each occupied table with a probability proportional to the number of customers already there, or choosing an empty table. Suppose that there are $(u - 1)$ customers occupying Q tables in this restaurant, where $\forall u \in \{1, 2, ..., U\}$ is the index of customers and U is the total number of customers. Let Ω_u be the table that is selected by the uth customer, let $\{1, 2, ... , Q\}$ be a set which elements are all occupied tables indexed by q, and let n_q be the number of customers already sitting at the qth table. Given the existing $(u - 1)$ customers' table selections, we can write the probability for the uth customer selecting the gth table, $\forall g \in \{1, 2, ..., Q, (Q + 1)\}$, as $\Pr\{\Omega_u = g \mid \Omega_1, \Omega_1, ... , \Omega_{u-1}\}$ which is equal to $(n_q - \alpha)/(u - 1 + \beta)$ if $g \in \{1, 2, ..., Q\}$, or is equal to $(\beta + Q\alpha)/(u - 1 + \beta)$ if $g = (Q + 1)$, where $\Pr\{\cdot \mid \cdot\}$ represents the conditional probability, and α and β are the discount and strength parameters for this Chinese restaurant process, respectively. To apply the Chinese restaurant process in our proposed proactive hierarchical caching mechanism, the mobile users are customers, the requested multimedia data contents are tables, and multimedia data content disseminations among mobile users are modeled as the customers selecting tables. Although the Chinese restaurant process considers an infinite number of tables and seats, we only calculate the probabilities of a customer selecting the occupied tables and one empty table, which are both finite numbers. Thus, we can apply the Chinese restaurant process to estimate the data content's popularity by considering the finite number of data and finite number of users who request each datum.

Suppose that all mobile users in a cellular wireless cell are homogeneous; we define $p(g) \triangleq \sum_{u=1}^{U} \Pr\{\Omega_u = g \mid \Omega_1, \Omega_2, ..., \Omega_{u-1}\}$ as the estimated popularity (i.e., the requested probability by all mobile users) for the multimedia data content g, which is the average requested probability over all mobile users at the current moment (i.e., the moment that mobile user u makes the request). We set p_r, p_b, and p_d $(0 < p_r < p_b < p_d < 1)$ as the caching threshold for caching in routers, cellular-BS/WiFi-APs, and mobile devices, respectively. When $p(g) < p_r$, the data content g is not cached at any caching tiers and only exists in the data source provider. When $p_r < p(g) < p_b$, the data content g is cached at one of the routers. When $p_b < p(g) < p_d$, the data content g is cached at a cellular BS or a WiFi AP. When $p(g) > p_d$, the data content g is cached at one of the mobile devices. Therefore, we can derive our objective function for this proactive caching mechanism at the current moment by adapting three caching thresholds, p_r, p_b, and p_d, to maximize the overall cache hitting rate, while taking into account the data contents' QoS requirements at different tiers as follows:

$$\begin{aligned} \max_{p_r, p_b, p_d} \sum_{g=1}^{Q+1} \Bigg(& \int_{p_r}^{p_b} p(g) UE(\theta_r) dp(g) \\ & + \int_{p_b}^{p_d} p(g) UE(\theta_b) dp(g) \\ & + \int_{p_d}^{1} p(g) UE(\theta_d) dp(g) \Bigg) \\ \text{subject to: } & C1: 0 < p_r < p_b < p_d < 1; \\ & C2: 0 \le p(g) \le 1; \end{aligned} \tag{2}$$

where $C1$ and $C2$ are two constraints for the optimization problem specified in Eq. 2.

Step 1: Each caching-tier player i determines its caching expense function $c_i(\cdot)$.
repeat
Step 2: Depending on its expense function $c_i(\cdot)$, each caching-tier player i derives the optimal number of data contents bid s_i^* and the optimal caching lifespan bid T_i^* in the current game round from Eq. 3 according to Karush-Kuhn-Tucker (KKT) conditions.
Step 3: Each caching-tier player i claims its current game round bids $s_i^{(n)} = s_i^*$ and $T_i^{(n)} = T_i^*$ to another two players.
Step 4: Using Eq. 3, all caching-tier players calculate the aggregate payoff of the current game round.
Step 5: According to the current game round bits in **Step 3**, each tier player i adapts its next game round bids to the better payoff performance level: $s_i^{(n+1)} = s_i^{(n)} + w_s(\Sigma_i s_i^{\max} - \Sigma_i s_i^{(n)})$ and $T_i^{(n+1)} = T_i^{(n)} + w_T(3T - \Sigma_i T_i^{(n)})$, where w_s and w_T are update step sizes for number of data contents and caching lifespan, respectively.
until The game states converge to the Nash Equilibrium and the overall payoff achieves the maximum.

ALGORITHM 3. Cooperative-game-theory-based hierarchical caching implementation algorithm.

The Algorithm to Implement the Proactive Hierarchical Caching Mechanism

Algorithm 2 details the caching localization among three wireless network tiers and caching lifespan for our proactive hierarchical caching mechanism. To simplify our description, we use the gth data content as an example to describe our proactive caching algorithm over three caching tiers. Let $l_g(u)$ be the caching lifespan for the gth data content when mobile user u makes a multimedia data content request.

Game-Theory-Based Hierarchical Caching Mechanism and Algorithm

The above random and proactive hierarchical caching mechanisms do not address, characterize, and compare the overall payoff for caching at three caching tiers, which is an important design criterion in 5G edge computing wireless networks. To optimize the overall payoff for caching at Tier 1, Tier 2, and Tier 3, we formulate the hierarchical caching problem as a cooperative gaming process where each caching tier bids for the number of cached multimedia data contents and caching lifespan under caching space constraints and collaborates with another two caching-tier players who bid for the same popular multimedia dataset, maximizing the aggregate effective capacity for all three tiers.

Game-Theory-Based Hierarchical Caching Mechanism

We define a given time period that consists of T time slots {1, 2, ..., T}, indexed by t. The total number of multimedia data contents in the time period T is $Q_{\max}$. The caching lifespan at each tier T_i, $\forall i \in \{r, b, d\}$, is specified above. Let $\rho_t(g)$ be the probability that the gth multimedia data content is requested by mobile users at time t. Let $I_{t,r}(g)$, $I_{t,b}(g)$, and $I_{t,d}(g)$ be the indicator function to indicate whether a cache hit has occurred at time t on Tier 1, Tier 2, and Tier 3, respectively, where $I_{t,i}(g) = 1$ or 0, $\forall i \in \{r, b, d\}$ indicates whether a cache hit has occurred or not. The number of cached data contents at each tier s_i, $\forall i \in \{r, b, d\}$ is constrained by the maximum allowed cache space $s_i^{\max}$. In our proposed game-theory-based hierarchical caching mechanism, caching-tier player i jointly bids for its optimal number of multimedia data contents s_i and caching lifespan T_i with the other two tier players, aiming to maximize the aggregate payoffs for all three caching tiers. We can write the objective function for this three-player cooperative game to maximize the aggregate payoffs (i.e., effective capacity minus caching expense) for all three tiers as follows:

$$\max_{\substack{T_r,T_b,T_d \\ s_r,\ s_b,\ s_d}} \left\{ \sum_{t=1}^{T} \sum_{g=1}^{Q_{\max}} \left[\begin{array}{l} \rho_t(g) I_{t,r}(g) E(\theta_r) + \rho_t(g) I_{t,b}(g) E(\theta_b) \\ + \rho_t(g) I_{t,d}(g) E(\theta_d) \end{array} \right] - \left[s_r \mathbb{E}\{c_r(T_r)\} + s_b \mathbb{E}\{c_b(T_b)\} + s_d \mathbb{E}\{c_d(T_d)\} \right] \right\}$$
$$\text{subject to: } C1: s_i \le s_i^{\max}, \forall i \in \{r, b, d\};$$
$$C2: 0 \le T_i \le T, \forall i \in \{r, b, d\};$$
$$C3: \rho_T(g) = \rho_0(g); \quad (3)$$

where $C1$, $C2$, and $C3$ are three constraints for the optimization problem shown in Eq. 3, $c_i(\cdot)$, $\forall i \in \{r, b, d\}$, is defined above, and $\mathbb{E}\{\cdot\}$ represents taking the expectation operation.

The Algorithm to Implement the Game-Theory-Based Hierarchical Caching Mechanism

Let s_i^* and T_i^*, $\forall i \in \{r, b, d\}$, be the optimal number of data contents and caching lifespan in the current game round, respectively, which are derived from Eq. 3. The caching-tier players denote their current game round bids for these two resources by $s_i^{(n)}$ and $T_i^{(n)}$ $\forall i \in \{r, b, d\}$, respectively. We develop our cooperative game algorithm for all three tier players to maximize the aggregate payoffs (i.e., effective capacity minus caching expense) in the edge computing wireless networks, as shown in Algorithm 3.

Performance Evaluations

Figure 3a plots our objective function, which is the average effective capacity for the three schemes of caching in Tier 1, Tier 2, and Tier 3 in Eq. 1 with respect to the departure rate for caching in mobile devices μ_d and the departure rate for caching in router μ_r. In Fig. 3a, we set $s_r \in [0, 20]$, $s_b \in [0, 20]$, and $s_d \in [0, 10]$ to plot Eq. 1. According to Algorithm 1, the probability of caching at each caching tier $\pi_i = G_i'(1)/(G_r'(1) + G_b'(1) + G_d'(1))$, $\forall i \in \{r, b, d\}$. We can observe from Fig. 3a that there is always a pair of optimal value of μ_d and μ_r that maximizes the average effective capacity, showing the existence of the optimal solution for Eq. 1. Figure 3a also shows that for a given departure rate μ_d, the optimal solution of Eq. 1 will increase as the departure rate μ_r increases. This is because for a given μ_d, π_d increases as μ_r increases, indicating that more popular data contents are retrieved from the cacheable mobile user through the D2D wireless link, and the effective capacity for the D2D wireless link is larger than or equal to the effective capacity of the wireless link from the routers to mobile users. The constraint $C1$ for the maximization problem is shown in Fig. 3b, where the caching expense is due to the retention time.

To optimize the overall payoff for caching at Tier 1, Tier 2, and Tier 3, we formulate the hierarchical caching problem as a cooperative gaming process where each caching tier bids for the number of cached multimedia data contents and caching lifespan under caching space constraints and collaborates with another two caching-tier players who bid for the same popular multimedia dataset, maximizing the aggregate effective capacity for all three tiers.

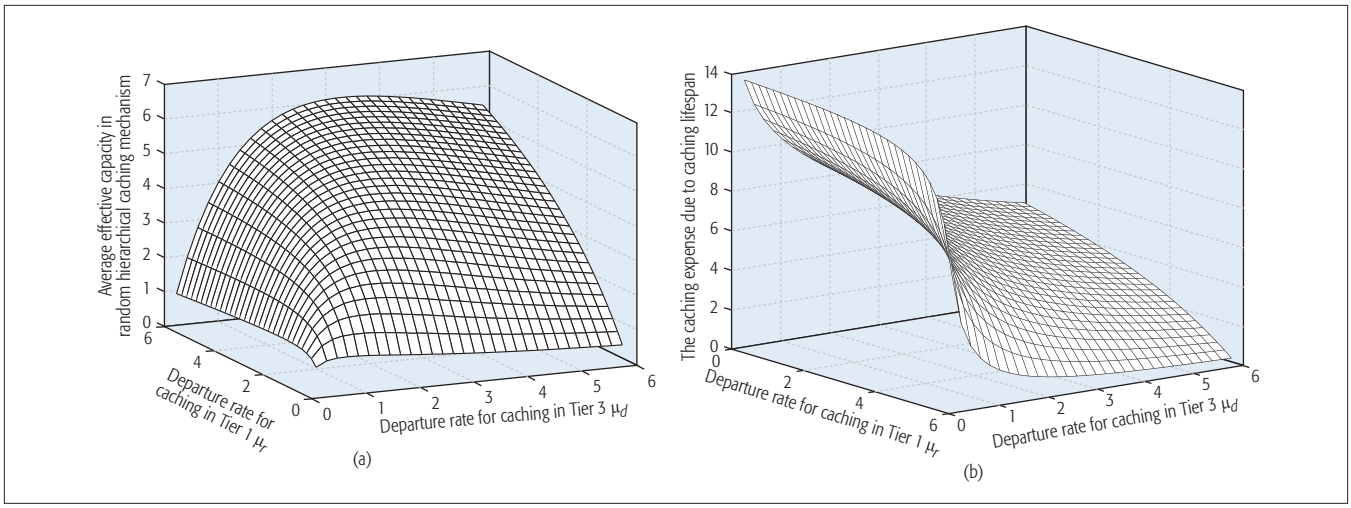

FIGURE 3. The performance for the random hierarchical caching mechanism with respect to the departure rate for caching in Tier 1 (routers) μ_r and the departure rate for caching in Tier 3 (mobile devices) μ_d: a) our objective function (i.e., average effective capacity) in Eq. 1; b) the constraint C1 in Eq. 1 for the objective function.

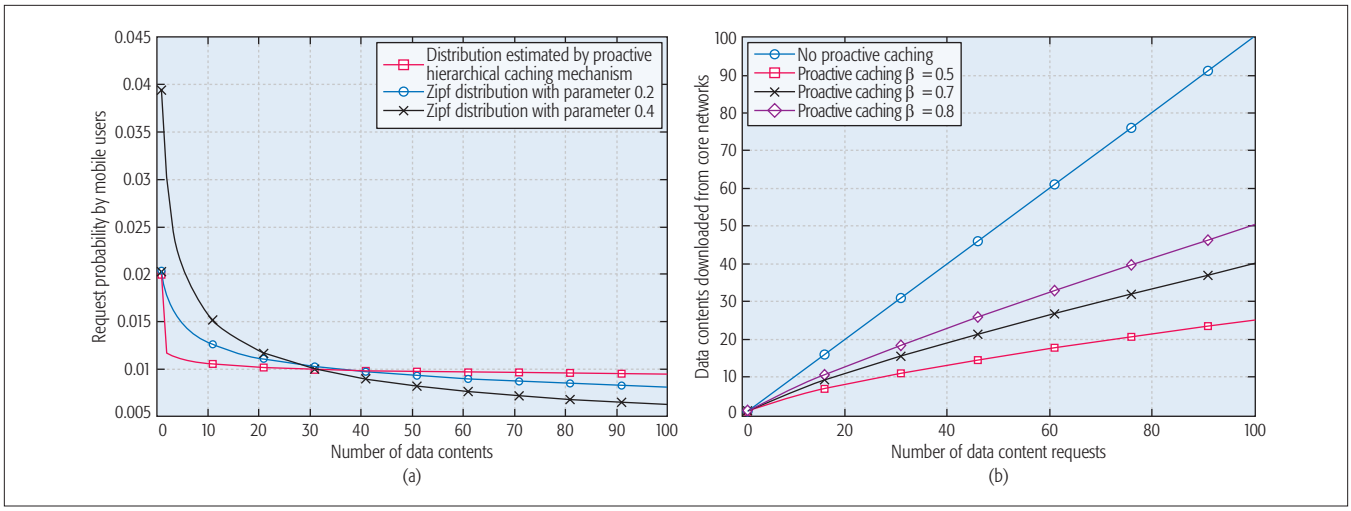

FIGURE 4. The performance comparisons for the proactive hierarchical caching mechanism: a) comparison of Zipf distribution of parameters 0.2 and 0.4 with the distribution estimated by the proactive hierarchical caching mechanism; b) the data contents served by core networks under different proactive hierarchical caching mechanism parameters.

We can observe from Fig. 3b that the caching expense decreases with the increase of μ_d and μ_r. Thus, for a given maximum expense constraint c^{max}, we can find all possible pairs of μ_d and μ_r according to Fig. 3b, and then find the optimal μ_d and μ_r according to Fig. 3a.

We compare the Zipf distribution of parameters 0.2 and 0.4 with the distribution estimated by the proactive hierarchical caching mechanism using $\alpha = 0.5$ and $\beta = 0.5$ in Fig. 4a. Figure 4a shows that our estimated popularity for data contents approximates the Zipf distribution. Figure 4b plots the scenarios for implementing the proactive hierarchical caching mechanism under different parameter β and the scenario without implementing the proactive hierarchical caching mechanism. We can observe from Fig. 4b that using our proposed proactive hierarchical caching, we can reduce the redundant popular data content transmissions and offload the traffic at core networks, validating the proactive hierarchical caching mechanism.

Figure 5 plots each caching-tier player's bids for our proposed game-theory-based hierarchical caching mechanism, where players cooperate with each other to maximize the aggregated payoffs. We set the update step sizes w_s of all three tiers as 0.5 for bidding the number of data contents, and set the update step sizes w_T of all three tiers as 0.7 for bidding the caching lifespan. We also set $s_r^{max} = 53$, $s_b^{max} = 50$, $s_d^{max} = 48$, and $T = 29$. We can observe from Fig. 5 that the bids for number of data contents and caching lifespan for all three tiers can converge to constants as time passes, validating the existence of Nash equilibrium, and thus, the aggregate payoffs for all three tiers can converge to Nash equilibrium. We can also observe from Fig. 5 that the converged number of data contents for Tier 3 are larger than those of Tier 1 and Tier 2, but the converged caching lifespan for Tier 3 is the shortest in all three tiers. This is because the caching expense in Tier 3 is larger than the other two tiers, so mobile devices frequently update their cached data to reduce the memory and battery cost. On the other hand, the cache space in Tier 3 is larger than the other two tiers due to the massive mobile devices.

Figure 6a plots the average effective capacity comparisons among the three hierarchical caching mechanisms for a mobile user under different total number of cache space in all three wireless networks tiers. Suppose that there are totally 20 mobile users in this wireless network, and the cache spaces in Tier 1, Tier 2, and Tier 3 are 20, 35, and 45 percent, respectively, of the overall cache spaces in the wireless network edge. We can observe from Fig. 6a that when the total cache space is relatively small, the random hierarchical caching mechanism outperforms the other two mechanisms. This is because for small caching space, the Markov chain model in the random hierarchical caching mechanism is able to derive the optimal departure rates for cached data contents leaving their cached tiers. On the other hand, when the total cache space is relatively large, the game-theory-based hierarchical caching mechanism outperforms the other two mechanisms. This is because, constrained by a large number of caching spaces and corresponding calculation complexity, for the Markov chain model in the random hierarchical caching mechanism and the estimations of data contents' popularity in the proactive hierarchical caching mechanism it is difficult to derive the optimal departure rates and predict data contents' popularity. Therefore, the cooperative-game-based hierarchical caching mechanism is able to obtain the maximum average effective capacity due to its cooperation algorithm for three caching-tier players. Figure 6b plots the achievable statistical delay-bound comparisons among the three hierarchical caching mechanisms. We set the delay-bounded QoS exponent $\theta_r = \theta_b = \theta_d$ to simplify the performance evaluation calculations. We can observe from Fig. 6b that the proactive hierarchical caching mechanism yields the minimum achievable statistical delay-bound, which is because this mechanism can estimate the most popular data contents and caches them at Tier 3 to minimize the delay.

Conclusions

We have proposed three hierarchical caching schemes to cache the delay-bounded multimedia data contents in different caching tiers of 5G edge computing wireless networks: caching in routers (Tier 1), caching in the cellular base station or WiFi access point (Tier 2), and caching in mobile devices (Tier 3). To make use of advantages and overcome disadvantages of the three caching schemes, we have developed three caching mechanisms under different motivations with each mechanism integrating the three caching schemes in different tiers. Our first proposed random hierarchical caching mechanism optimizes the three caching schemes aiming to maximize the average effective capacity for all three caching tiers. Our second proposed proactive hierarchical caching mechanism can maximize the overall cache hitting rate for the entire wireless edge by estimating the popularity of each data content in order to cache the data content with high popularity nearby mobile users. Our third proposed game-theory-based hierarchical-caching mechanism formulates the three caching schemes in different caching tiers as a cooperative game in order to enable each caching-tier player to bid

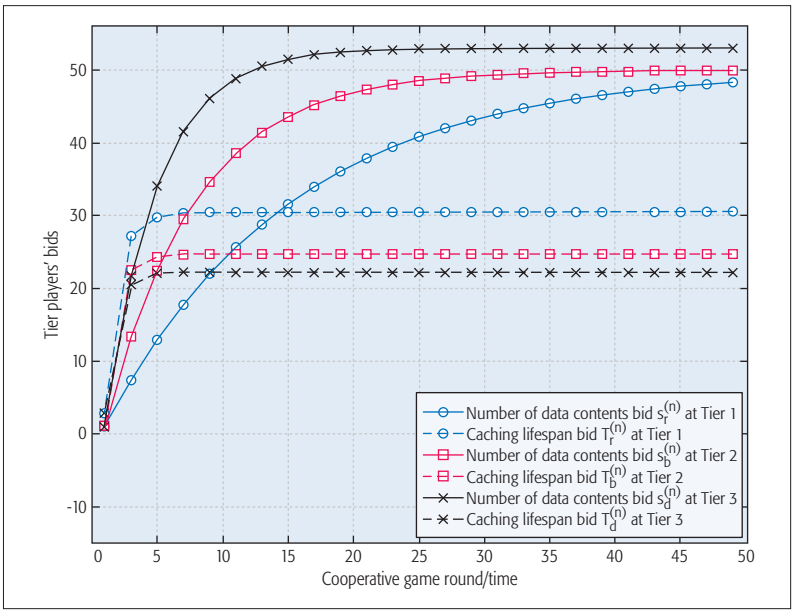

FIGURE 5. Caching-tier players' bids in the cooperative-game-based hierarchical caching mechanism.

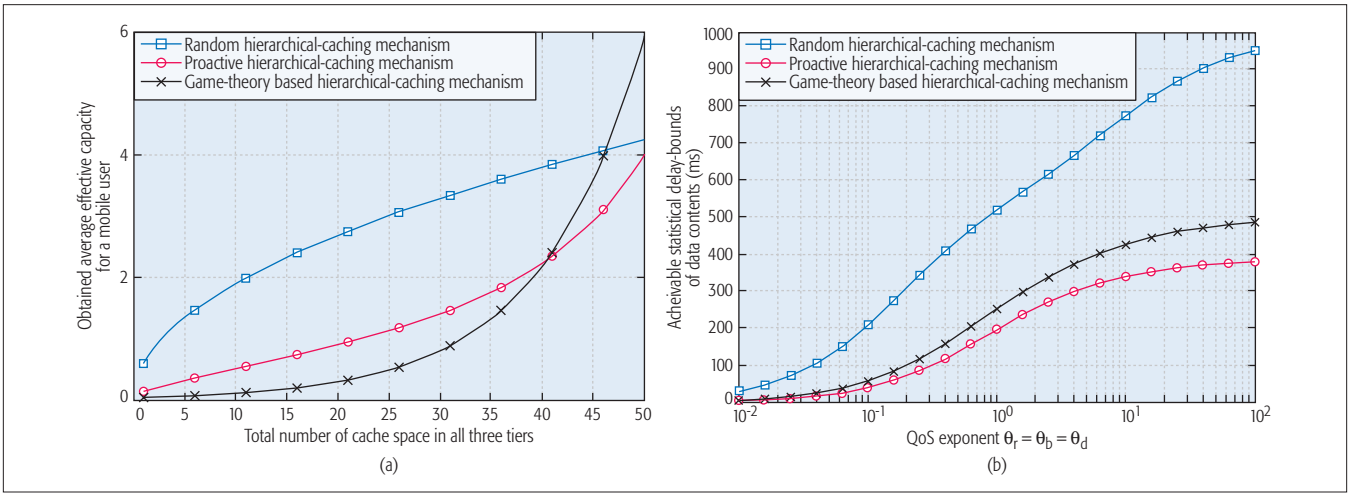

FIGURE 6. Performance comparisons for the random hierarchical caching mechanism, proactive hierarchical caching mechanism, and game-theory-based hierarchical-caching mechanism: a) the comparisons for average effective capacity for three hierarchical caching mechanisms; b) the comparisons for achievable statistical delay-bounds for the three hierarchical caching mechanisms.

We have proposed three hierarchical-caching schemes to cache the delay-bounded multimedia data contents in different caching tiers of 5G edge computing wireless networks. To make use of advantages and overcome disadvantages of the three caching schemes, we have developed three caching mechanisms under different motivations with each mechanism integrating the three caching schemes in different tiers.

for number of data contents and caching lifespan under caching expenses, maximizing the aggregate effective capacities for all three caching tiers.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation under Grants ECCS-1408601 and CNS-1205726, and the U.S. Air Force under Grant FA9453-15-C-0423.

References

[1] I. Chih-Lin *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.
[2] B. Bangerter *et al.*, "Networks and Devices for the 5G Era," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 90–96.
[3] H. Su and X. Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Provisionings over Cognitive Radio Wireless Networks," *IEEE JSAC*, vol. 26, no. 1, Jan. 2008, pp. 118–29.
[4] H. Liu, *et al.*, "On Content-Centric Wireless Delivery Networks," *IEEE Wireless Commun.*, vol. 21, no. 6, June 2014, pp. 118–25.
[5] X. Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 131–39.
[6] G. Xylomenos *et al.*, "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 2, Feb. 2014, pp. 1024–49.
[7] X. Zhang and Q. Zhu, "Information-Centric Network Virtualization for QoS Provisioning over Software Defined Wireless Networks," *Proc. IEEE MILCOM*, 2016, Baltimore, MD, Nov. 1–3, 2018.
[8] C. Liang, F. R. Yu, and X. Zhang, "Information-Centric Network Function Virtualization over 5G Mobile Wireless Networks," *IEEE Network*, vol. 29, no. 3, May/June 2015, pp. 68–74.
[9] J. Tang and X. Zhang, "Quality-of-Service Driven Power and Rate Adaptation over Wireless Links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, Aug. 2007, pp. 3058–68.
[10] X. Zhang and Q. Zhu, "Collaborative Hierarchical Caching over 5G Edge Computing Mobile Wireless Networks," *Proc. IEEE ICC 2018*, Kansas City, MO, May 20–24, 2018.
[11] D. Wu and R. Negi, "Effective Capacity: A Wireless Link Model for Support of Quality of Service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, Apr. 2003, pp. 630–43.
[12] A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd ed., Pearson/Prentice Hall, 2008.
[13] F. Mehmeti and T. Spyropoulos, "Performance Modeling, Analysis, and Optimization of Delayed Mobile Data Offloading for Mobile Users," *IEEE/ACM Trans. Networking*, vol. 25, no. 1, Jan. 2017, pp. 550–64.
[14] D. P. Bertsekas and R. G. Gallager, *Data Networks*, vol. 2.
[15] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 82–89.

Biographies

Xi Zhang [F] received his B.S. and M.S. degrees from Xidian University, Xi'an, China, and an M.S. degree from Lehigh University, Bethlehem, Pennsylvania, all in electrical engineering and computer science, and his Ph.D. degree in electrical engineering and computer science (electrical engineering-systems) from the University of Michigan, Ann Arbor. He is currently a full professor and the founding director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He is a Fellow of the IEEE for contributions to quality of service (QoS) theory in mobile wireless networks. He is an IEEE Distinguished Lecturer of both IEEE Communications Society and IEEE Vehicular Technology Society. He also received also received a TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He was with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hill, New Jersey, and AT&T Laboratories Research, Florham Park, New Jersey, in 1997. He was a research fellow with the School of Electrical Engineering, University of Technology, Sydney, Australia, and the Department of Electrical and Computer Engineering, James Cook University, Australia. He has published more than 320 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 Best Paper Awards at IEEE ICC 2018, IEEE GLOBECOM 2014, IEEE GLOBECOM 2009, IEEE GLOBECOM 2007, and IEEE WCNC 2010, respectively. One of his *IEEE Journal on Selected Areas in Communications* papers has been listed as the IEEE Best Readings Paper (receiving the highest citation rate among all IEEE transactions/journal papers in the area) on Wireless Cognitive Radio Networks and Statistical QoS Provisioning over Mobile Wireless Networking. He also received a TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He is serving or has served as an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *IEEE Transactions on Network Science and Engineering*, twice as a Guest Editor for the *IEEE Journal on Selected Areas in Communications* for two Special Issues, one on Broadband Wireless Communications for High Speed Vehicles and the other on Wireless Video Transmissions, an Associate Editor for *IEEE Communications Letters*, twice as Lead Guest Editor for *IEEE Communications Magazine* for two Feature Topics, one on Advances in Cooperative Wireless Networking and the other on Underwater Wireless Communications and Networks: Theory and Applications, and a Guest Editor for *IEEE Wireless Communications* for a Special Issue on Next Generation CDMA vs. OFDMA for 4G Wireless Applications, an Editor for Wiley's *Journal on Wireless Communications and Mobile Computing*, the *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 WCNC 2013, TPC Chair for IEEE INFOCOM 2017–2018 Workshops on Integrating Edge Computing, Caching, and Offloading in Next Generation Networks, and TPC/General Chair for numerous other IEEE/ACM conferences, symposia, and workshops.

Qixuan Zhu received her B.S. degree from Tianjin University of Technology and Education, China, and her M.S. degree from George Washington University, Washington, DC, all in electrical and computer engineering. She is currently pursuing a 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. She received the Best Paper Award at IEEE ICC 2018.