IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON GREEN COMMUNICATIONS ON WIRELESS NETWORKS

Received June 19, 2019, accepted July 17, 2019, date of publication July 26, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931449

Assessing the Energy Consumption of Proactive Mobile Edge Caching in Wireless Networks

MING YAN[1], (Member, IEEE), CHIEN AUN CHAN[2], (Member, IEEE), WENWEN LI[3], LING LEI[1], ANDRÉ F. GYGAX[4], AND CHIH-LIN I[3], (Senior Member, IEEE)

[1]School of Information and Telecommunications Engineering, Communication University of China, Beijing 100024, China
[2]Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
[3]Green Communications Research Center, China Mobile Research Institute, Beijing 100053, China
[4]Department of Finance, Faculty of Business and Economics, The University of Melbourne, Parkville, VIC 3010, Australia

Corresponding author: Ming Yan (yanm@cuc.edu.au)

This work was supported in part by the Fundamental Research Funds for the Central Universities of China under Grant 2018CUCTJ078 and Grant CUC18A002-2.

ABSTRACT Multiaccess edge computing and caching (MEC) is regarded as one of the key technologies of fifth-generation (5G) radio access networks. By bringing computing and storage resources closer to the end users, MEC could help to reduce network congestion and improve user experience. However, deploying many distributed MEC servers at the edge of wireless networks is challenging not only in terms of managing resource allocation and distribution but also in regard to reducing network energy consumption. Here, we focus on the latter by assessing the network energy consumption of different cache updating and replacement algorithms. First, we introduce our proposed proactive caching (PC) algorithm for mobile edge caching with Zipf request patterns, which could potentially improve the cache hit rates compared to other caching algorithms such as least recently used, least frequently used, and popularity-based caching. Then, we present the energy assessment models for mobile edge caching by breaking down the total network energy consumption into transmission and storage energy consumption. Finally, we perform a comprehensive simulation to assess the energy consumption of the PC algorithm under different key factors and compare with that of conventional algorithms. The simulation results show that improving cache hit rates by using the PC algorithm comes at the expense of additional energy consumption for network transmission.

INDEX TERMS Wireless edge caching, energy consumption, 5G, multiaccess edge computing, proactive caching.

I. INTRODUCTION

Emerging technologies (e.g., virtual reality, augmented reality, three-dimensional (3D) videos/games, and autonomous driving) require high bandwidth and extremely low latency to guarantee quality-of-service (QoS), high user quality of experience (QoE) [1], [2] and safety. The emergence of these applications has required rapid development of fifth-generation (5G) wireless networks. It is expected that 5G precommercial data terminals, smartphones, and other products will be released in the first half of 2019. By 2020, telecom operators are expected to realize large-scale deployments of 5G base stations [3].

In the logical architecture of 5G access networks, the baseband functionality of a cellular base station (BS) will be divided into two parts—a centralized unit (CU) and a distributed unit (DU) [3], [4]. The DU can be deployed in a macro or a small cell BS such as the micro, pico or femtocell [3], [4]. This two-level network architecture allows different deployment scenarios of multiaccess edge computing and caching (MEC) servers. For example, the MEC data servers can be deployed in a CU or a DU depending on the requirements of services and applications as well as usage patterns of local users [5]. Distributed caches are deployed very close to the end users, services and content items are delivered from the wireless edge caches instead of going through the backbone network to provide high bandwidth and low latency performance to the end users [6]. Although the power consumption of a single edge server is relatively low, a very large number of MEC servers is expected to be deployed in 5G networks. Therefore, the energy consumption

The associate editor coordinating the review of this manuscript and approving it for publication was Ilsun You.

of maintaining these caches at the edge of the wireless networks is challenging for mobile operators and has yet to be fully investigated [7], [8].

Because edge storage capacity is relatively small compared to a cloud data center, only a limited number of content items can be cached in the wireless edge caches. Thus, the most popular content items need to be identified and stored in the MEC servers to improve content request hit rates [5], [9]. Existing research shows that the popularity of web content varies by time and geography [10], [11], which requires high-frequency refreshing of cached content according to an update and replacement algorithm. However, the refresh rate and the size of the replacement content items require additional transmission energy to be consumed by the network [12]. Therefore, the design of an effective caching algorithm to minimize the transmission energy consumption while maximizing the cache hit rates remains an open research question.

To address the above challenges, we first propose a proactive cache updating algorithm for MEC based on a 5G network architecture using big data analysis in our previous work [13]. To reveal the performance of different algorithms, we simulate the MEC network architecture and calculate the cache hit rate and the number of cache content items in different simulation scenarios by extending our previous work. We then assess the transmission and storage energy consumption of edge caching under different configurations. We make the following contributions to the literature:

- We investigate the performance of different caching algorithms by analyzing the cache hit rate and the number of cache content items that need to be transported in different simulation scenarios.
- We investigate energy consumption of edge caching under different configurations such as different content refresh periods and cache size limitations.
- We compare the transmission and storage energy consumption of the conventional algorithms with that of the proposed algorithm.

The rest of the paper is organized as follows. Section II discusses related work on the energy consumption of different networks and caching strategies. In Section III, we present the CU/DU logical architecture and MEC server deployment scheme in 5G wireless networks. Section IV first introduces the conventional content update strategies. We then propose a proactive cache update and replacement algorithm based on big data prediction. In Section V, we first compare the performance of PC to conventional algorithms in different simulation scenarios. We then simulate the energy consumption of network transmission and storage of caches of our proposed caching algorithm. Finally, we compare the energy consumption of the proposed algorithm with that of conventional algorithms. Section VI concludes the paper.

II. RELATED WORK

Despite significant research on resource allocation and distribution of edge caching, a deeper understanding of what constitutes an effective and energy-efficient design of edge caching strategies is necessary. In [7], [8], the authors offered energy-optimal edge content cache and dissemination designs for both hot spot and rural areas, respectively. Due to different types of base stations that were deployed in different areas with different population densities, different edge caching strategies were designed to minimize the overall energy consumption. In [12], the authors considered that content data can be stored in both base stations and user devices and analyzed the energy consumption in both backhaul and access networks under two different caching strategies. Two optimization problems were proposed to minimize the total energy consumption for these two caching strategies while satisfying some predefined QoS constraints. To minimize the energy consumption of the MEC servers in 5G cellular networks, the authors in [14] considered the MEC servers' energy consumption, backhaul network capacities and content popularity distributions, and formulated a joint optimization framework under a given average download latency. Simulation results showed that the proposed solution could obtain better performance in terms of energy efficiency gains compared to conventional caching placement strategies. The authors in [15] minimized the energy consumption of a clustered device-to-device caching network under a random probabilistic caching scheme, where files were independently cached according to a specific probability distribution.

Another strand of research has investigated the cache performance of the optimized web caching strategies compared with the conventional nonpredictive methods such as least recently used (LRU), least frequently used (LFU) [16], [17] and popularity-based caching [18]. Simulation results showed that the hit rate of predictive content update strategies increased under different content request patterns [17]. Although 5G technology can provide high bandwidth for high-quality mobile video streaming, mobile users have to address the challenge of frequent handoffs between the 5G small cells. Some research proposed proactive content caching at the access edge to effectively maintain high-quality mobile video streaming for high-mobility 5G users moving among small cells [19], [20]. In [21], an integrated proactive content delivery scheme was proposed by exploiting both the availability of multiple service tiers and mobile user behavior prediction. The performance of the proposed scheme was then investigated to reveal the impacts of proactive window size, service-tier price ratio and traffic cost. For device-to-device (D2D) enabled networks, the authors of [22] proposed a proactive caching scheme. Their numerical results showed that up to 30% more users could be satisfied using this scheme compared to reactive caching [22]. In addition, the authors of [23] proposed a fog-to-fog data caching and selection method based on a data caching and selection strategy. The corresponding simulation results showed that this method could reduce the data retrieval latency and increase the file hit rate in 5G [23]. However, the effectiveness of different cache content updating algorithms to reduce the energy consumption in 5G

edge caching networks still requires more comprehensive assessments.

In our prior work, we investigated the energy assessment models of wireless access networks [24] and end-to-end wireless networks [25]. We proposed an energy model and assessed the access network energy consumed by different mobile services based on both the data and signaling traffic generated by those services [24]. We also developed a comprehensive service-specific end-to-end energy model to assess the energy consumption of each network segment, including the end-user devices, wireless access network, wireline core network and data center [25]. Here, we investigate the edge cache deployment scheme for 5G network architectures and extend these energy models to assess the energy consumption of 5G wireless edge caching with different cache content updating strategies.

III. CU/DU LOGICAL ARCHITECTURE AND MEC SERVER DEPLOYMENT SCHEME IN 5G

In 5G access networks, a CU/DU split is proposed to enable and enhance the cloud radio access network technology via several split options (3GPP TR 38.801). This split architecture provides centralization and distribution of control and capabilities depending on each situation of wireless networks [26]. For a CU/DU structure, a stack partition between the CU and the DU can also be optimally configured via big data analytics based on service patterns, fronthaul capability, frequency bands, user mobility, quality of experience [27].

Different from the fourth generation (4G) radio access architecture, the CU, DU and radio remote unit (RRU) form a gNodeB (gNB) BS, as shown in Figure 1. The 5G gNB is connected to the content delivery network (CDN) server or the data center through the network elements (e.g., routers and gateways.) in the core network. The CU is a centralized node, and it is connected with the DUs via a next-generation fronthaul interface (NGFI) [3]. The DU is connected to the RRUs via an NGFI. It can implement RF processing and baseband processing functionalities together with the RRU. A range of user equipment (UE) (e.g., smartphones, tablets, and enhanced mobile broadband devices) access the 5G network through these RRUs [26]. Using this architecture, a CU can support multiple DUs. Most of the control functionalities are centralized at the CU, while the fast scheduling of the air interface is performed at the DU [4], [6].

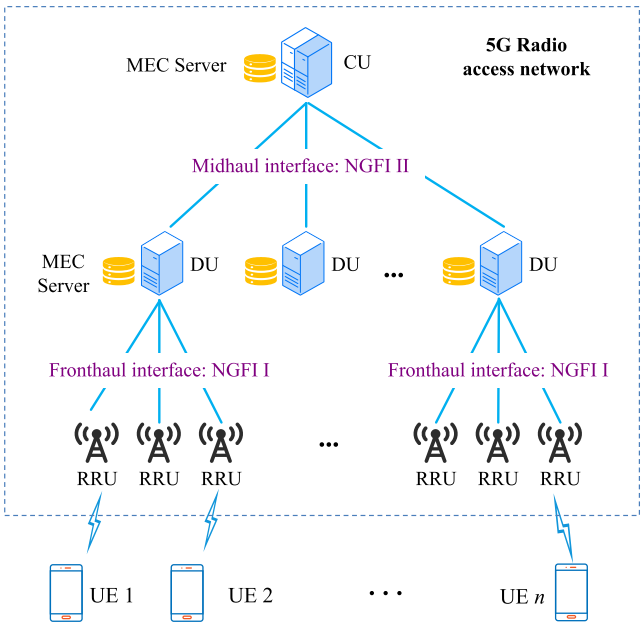


FIGURE 1. 5G radio access network architecture and MEC server deployment scheme.

The advantages of the CU/DU two-level architecture are summarized below:

- The hardware is more flexible than in existing wireless networks. The CU/DU can be a stand-alone device or integrated into a baseband unit (BBU) as a software module;
- The separation of CU and DU facilitates the coordination of performance and load management as well as real-time performance optimization. Network function virtualization (NFV) and software-defined networking SDN) are key technologies that make use of this architecture;
- Functional partitioning is configurable to meet the needs of different application scenarios, such as the variability of transmission delay [4].

The MEC servers can be flexibly configured through technologies such as SDN and NFV. As shown in Figure 1, the MEC server can be deployed at the CU level to serve all DUs connected to this particular CU. Users connected to these DUs can access the resources provided by this MEC server deployed at the CU. Alternatively, the MEC server can also be deployed in the DU. With this solution, each DU under a CU can have its storage scheme, and big data analysis can be used to provide personalized access services for different users connected to different DUs.

IV. CONTENT UPDATE STRATEGY OF EDGE CACHING

With the large number of MEC servers to be deployed in wireless networks, due to operational costs, the overall storage capacity of MEC servers will be lower than that of traditional data centers. Therefore, it is impractical to replicate all content items from the data center on the MEC servers. It is well known that the popularity of network video content follows the Zipf distribution (i.e., a relatively small number of the top popular video content items dominate most of the requests within a certain period of time [15]). Therefore, predictive analytics can be used to cache the most popular video content items in the next time period on the corresponding MEC servers to satisfy user requests (i.e., hit rate).

In practice, updating can be challenging as web content popularity is expected to change over time. Furthermore, each video content item has its life cycle: (1) growing in popularity, (2) reaching a peak, (3) declining in popularity and finally (4) reaching a low-level long-term equilibrium (i.e., long-tail

reduction in user requests). In addition, due to the heterogeneity of daily mobility patterns of users, the content request preferences in different regions also change throughout each day. To guarantee the requested hit rate of the cached contents and users' QoE, the cached content items need to be updated periodically. Different updating algorithms have their advantages and disadvantages. Theoretically, if the updating frequency is higher, the content request hit rate will be higher. However, a higher updating frequency will result in additional transmission overhead with a corresponding increase in energy consumption. Therefore, designing an effective cache updating algorithm and balancing the updating frequency and the corresponding network energy costs remains a major challenge.

Next, we evaluate the most common caching algorithms, i.e., LRU and LFU, before introducing our PC caching algorithm.

A. LEAST RECENTLY USED (LRU) ALGORITHM

The LRU algorithm assumes that the currently requested content is very likely to be requested in the next time period. Videos are first sorted according to the chronological order in which they have been requested in the previous period of time t. Top content items are cached based on the size of the cache. In the next time period t_1, if any new videos are requested, the new videos will replace those at the end of the queue [16].

The advantage of LRU is that the algorithm is simple and has high efficiency when the access content does not change much. The shortcoming of LRU is that it is vulnerable to random access noise, that is, the random access of the unpopular content items is mistaken for a large cache value, resulting in additional data transmission and storage overhead [16].

B. LEAST FREQUENTLY USED (LFU) ALGORITHM

The LFU algorithm eliminates data based on the historical request frequency of the data. It assumes that if the content item has been requested multiple times in the past, it will be requested more frequently in the future [16]. In LFU, each video content has a request count, all content items are sorted by their request counts, and content items with the same reference count are sorted by time.

In general, the efficiency of the LFU algorithm is better than LRU, and LFU can avoid the problem that the cache hit rate is reduced due to periodic or sporadic operations. However, the LFU algorithm needs to record historical request records of data. Once the data request mode changes, LFU needs a longer time to apply the new request mode. The disadvantage of the LFU algorithm is that historical data have a greater impact on future data; that is, old content that is no longer requested may accumulate a high frequency of request. In addition, a queue is required to record the request records of all content items. Each content needs to maintain a request count, so the algorithm's complexity is higher than for LRU [16].

C. POPULARITY-BASED CACHING ALGORITHM

In a content-centric networking architecture, the content is cached in the network nodes along its delivery path if caches are available. To manage the caches of the nodes effectively, the popularity-based caching strategy has been proposed to achieve a higher cache hit rate than the default caching strategy [18].

For popularity-based caching, the number of requests for each content item is counted by every node. Then, each node sorts the content items based on local statistics and caches the most popular content items. At the same time, the node notifies its neighbors to store the same content items. After receiving the notification message, the neighboring nodes determine whether to cache the content according to their own caching capabilities and constraints [18].

D. PROACTIVE CACHING (PC) ALGORITHM

To address the technical shortcomings of LRU and LFU, i.e., both algorithms cannot predict the request rate of new online content items, and they cannot track the rapid changes in content popularity, we propose a proactive cache (PC) updating and replacing algorithm based on a prediction from big data analytics. When the user behavior of content requests differs greatly, the PC algorithm considers the prediction based on historical request behavior to maximize the content request hit rate while minimizing the cache size requirements. In other words, PC is proposed to effectively improve the cache efficiency and minimize the network operating costs. All symbols and parameters used in the PC algorithm are described in Table 1.

TABLE 1. Description of symbols and parameters.

Symbol	Description
t	Time period. t_0: initial time period. t_1: next time period.
S_i	The ith content sorted with score.
m	Limit of the number of content items in one edge cache.
n	Number of all contents. In general, $n > m$.
R	Cache refresh period.
C	List of content items in the current cache.
C'	List of content items for the next time window.
$W_1 \sim W_4$	Weighting factors of the four time windows.
$C_1 \sim C_4$	Predicted sore lists in the next four time windows.
S_{ji}	Predicted sore of each content item in the jth time windows, $1 <= j <= 4$, $1 <= i <= n$.

As shown in Figure 2, based on the long-term historical data of user and content requests from the previous time periods, existing data mining algorithms can effectively predict the content items that are most likely to be requested during some future time period [28], [29]. However, two problems may arise. First, too many content items might be replaced when the prediction results are directly used to update the cache content. Second, the cache's efficiency relies heavily on the accuracy of the prediction. To address these problems, our PC algorithm determines the prioritization of the content items in the caching queue for the next time period based on the prediction results for future time windows and the

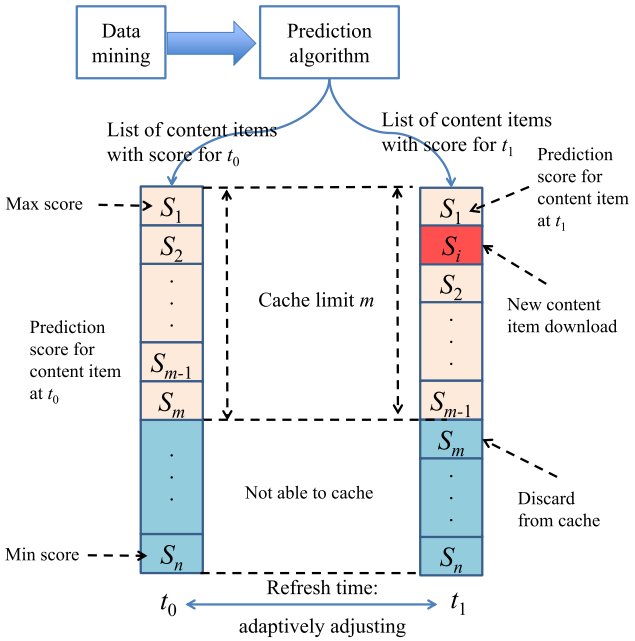


FIGURE 2. A schematic of the PC algorithm.

corresponding weighting factors. Hence, the PC algorithm replaces some expired or lowly requested content items with predicted content items that are expected to have high hit rates. Therefore, the PC algorithm mainly determines the priority and discards based on the order of the new content items that are expected to be requested in the next time period and the existing content items in the caching queue.

Table 1 denotes the cache refresh period by R. R can by optimized by mobile operators by changing the duration of the parameter (e.g., 15 minutes or longer). For example, R can be varied depending on network load requirements, e.g., finding a time window with low network load for transportation of content items so that it will not disrupt users' traffic. If the number of sliding time windows is 4, then the four time windows are $\{t \rightarrow t+R, t+R \rightarrow t+2R, t+2R \rightarrow t+3R, t+3R \rightarrow t+4R\}$. We then assume that the weighting factors of the list of content items to be requested in different time windows are $W_1 \sim W_4$. Generally, the corresponding weight factor of a closer time window will be greater due to different caching values. The weighting factors can be updated with the objective of reducing content transportation (therefore reducing transport energy consumption) while at the same maximizing the hit rates of caches. In our simulation, we set $W_1 \sim W_4$ to 1, 0.75, 0.5, and 0.25. However, it should be noted that the weight factors are dependent on the length of the time window, the similarity of user content request patterns, cache size and network conditions. The dependency of the weight factor on those factors mentioned above will be investigated in future work. The content items in the current cache list and the content items predicted to be requested in different time windows are combined to calculate the total weight of all content items, and the content items are rearranged in descending order relative to the total weight to obtain the new list. The PC algorithm is shown in Algorithm 1.

Algorithm 1 PC Algorithm

Input:
List of contents in current cache, C: $\{S_1, S_2, \ldots, S_m\}$; Content refresh rate, R (15 minutes, 30 minutes, etc.); Time windows, $\{t \rightarrow t+R, t+R \rightarrow t+2R, t+2R \rightarrow t+3R, t+3R \rightarrow t+4R\}$;
Weighting factors, $W_1 \sim W_4$.

Output:
List of content items in the cache for the next time window, C'.

1: Predict the probability of each content item being requested in the next 4 time windows, and obtain four lists of scores $C_1 \sim C_4$ (C_j: $\{S_{j1}, S_{j2}, \ldots, S_{jn}\}$);
2: **for** $(j = 1; j <= 4; j++)$
3 **for** $(i = 1; i <= n, i++)$
4 $S_i += S_{ji} * W_j$;
5: Sort content list in descending order relative to the total content weight, $S_1 \sim S_n$;
6: Determine the list of content items with top m scores, C', for the next time window, $t \rightarrow t+R$
7: Discard contents that are in the cache list C but not in the list C';
8: Transport the new contents in the list C';
9: At time $t+R$, shift the four time windows by R, and repeat 1 to 6 for the next C'.

V. SIMULATION OF PERFORMANCE AND ENERGY CONSUMPTION OF EDGE CACHING

A. SIMULATION SETUP

We simulate the mobility patterns of 2,500 users on an 8 × 8, 64-cell playground (intercell-distance of 500 m) using the smoothly truncated Levy walks algorithm [30] with an average user movement speed of 20 km/h (which indicates that users are commuting). The algorithm simulates the mobility pattern of individual mobile users within the environment using preset probability distributions for travel distance, pause length, and change in travel direction [30]. Content requests from mobile users are modeled based on Poisson arrivals.

Related research has shown that the number of requests for web content follows Zipf's distribution [10], [15]. We assume that the maximum number of video contents that can be stored per edge cache is M and that the total number of video contents is N_f. All video content items are sorted according to popularity, from high to low, then the probability $P(i)$ of each content is subject to Zipf's distribution as follows:

$$P(i) = \frac{i^{-\alpha}}{\sum_{k=1}^{N_f} k^{-\alpha}} \quad (1)$$

where α indicates the similarity in content requests of different users. A smaller α indicates lower similarity. For example, if $\alpha = 0$, the probability that each video content is requested has a uniform distribution. As α increases, different users' requests have a higher similarity. In other words, the lower indexed content has a higher request probability [10]. We use Zipf's distribution to model the popularity of video content items in our simulation, and we model two different scenarios in terms of user similarity: low and high similarity, by setting α to 0.4 and 1.2, respectively.

We assume that the size of the video content pool is 500,000, the average size of each video content, S_f, is 15 Mbytes (approximately 1 minute of high definition video on YouTube) and the edge cache constraint is limited to 20% of the size of all contents. In other words, an edge cache can store up to the top 20% of the top popular contents.

The updating of stored content causes new content items to be transferred from the CDN server to the edge caches. The energy consumption of transmission of content items can be calculated using the following equation:

$$E_{transport} = (N_c E_c + N_e E_e + E_{bng} + E_{sw}) \times S_t \qquad (2)$$

where N_c and N_e are the numbers of core and edge routers in the core network and edge network, E_c, E_e, E_{bng}, and E_{sw} denote the energy per bit of the core router, the edge router, the broadband network gateway (BNG), and the Ethernet switch, respectively. S_t indicates the data size of the content item needing to be transported [25]. Table 2 lists the estimated energy per bit of different types of equipment in the wireline core network [31].

TABLE 2. Energy per bit of equipment in the wireline core network.

Type	Energy per bit (J/Gbit)
Core router	5.2557
Edge router	15.5536
BNG	11.3063
Ethernet switch	13.2056

Next, we assume that the power consumption per bit of caching in the MEC server, $P_{caching}$, is 6.25×10^{-12} W/bit. We then calculate the energy consumption for caching one content item for a certain period of time (for example, 1 hour) as below:

$$E_{caching} = P_{caching} \times S_f \times T \qquad (3)$$

where T is the caching duration. A summary of simulation parameters and assumptions is provided in Table 3.

B. PERFORMANCE OF PROACTIVE CACHING WHEN CACHING AT THE DUs

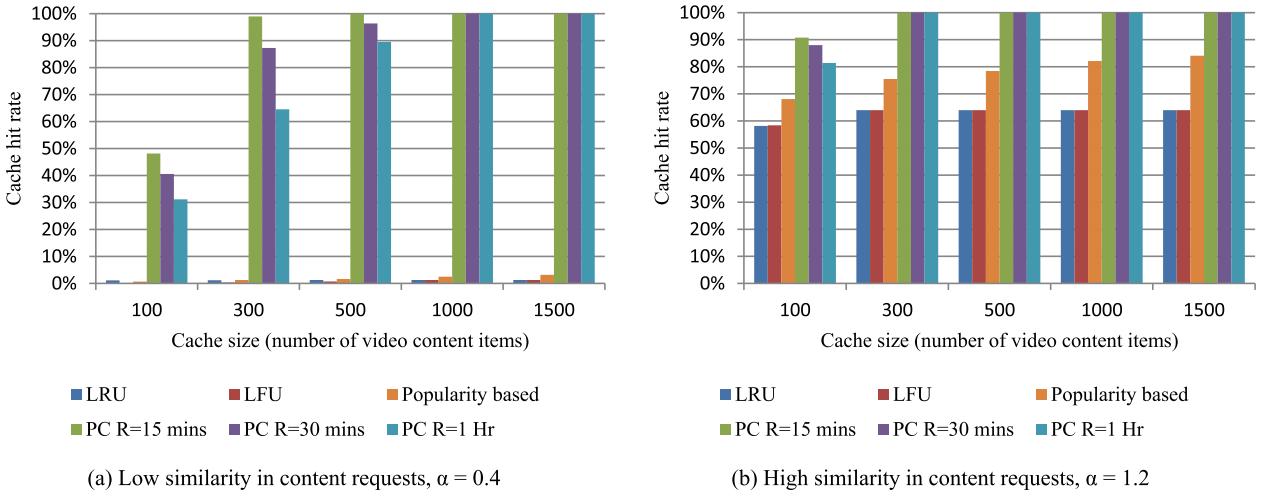
The number of active users connected to a wireless network varies at different times of the day. For instance, the busy hours refer to the hours with more active users and comparatively high traffic. In this simulation, we assume serving user demands during medium traffic loads with approximately 1,250 users (out of 2,500) being simultaneously active. Figure 3 shows the average cache hit rates of PC with different refresh periods compared to the conventional caching algorithms. We then analyze the impact of cache size on performance by varying its size. We find that when the cache capacity reaches 1,000, the average hit rate of the PC algorithm is close to 100%, so the maximum cache size in our simulation is set to 1,500. In addition, different user behavior similarities can also have an impact on performance. Here, we simulate low similarity in content requests ($\alpha = 0.4$) as shown in Figure 3(a) and high similarity in content requests ($\alpha = 1.2$) as shown in Figure 3(b).

TABLE 3. Simulation parameters and assumptions.

Parameter	Assumption
Radio access network architecture	Separated CU & DU
Number of DUs connected with a CU	3
Transmission delay	Ignored
Average data rate	1 Gbps
Number of 5G users	2500
Number of base stations	64 (8 × 8)
Intercell-distance	500 m
Average speed of users	20 km/h
α	0.4: low similarity 1.2: high similarity
N_f	500,000
S_f	15 Mbytes
N_c	5
N_e	2
$P_{caching}$	6.25×10^{-12} W/bit

Figure 3(a) shows that the average cache hit rates of LRU/LFU and the popularity-based caching are very low (approximately 1% for LRU/LFU and 3% for the popularity-based caching) even if the cache size reaches the maximum of 1,500 because when the similarity in user behavior is relatively low, the users' requests are relatively scattered. The caching strategies of LRU and LFU are not based on the prediction of user behavior, so they cannot effectively satisfy the requests of most users. Figure 3(a) also shows that PC algorithm improves the cache hit rate significantly compared to LRU, LFU, and popularity-based caching, which is because PC utilizes predictive analytics on user content requests and can precache popular video content at DU caches. Furthermore, as the refresh period R increases, the cache hit rate of the PC will decrease, especially when the cache size is small. For example, if a DU caches 100 video contents, the cache hit rate of PC is 48% when R = 15 minutes, and it decreases to 31% when R = 1 hour because a higher refresh rate enables timely prediction and update of cached content items to better meet user requests. However, comparing Figures 3(a) and 3(b), we observe that the average cache hit rates increase significantly as the user similarity in content requests increases, especially for LRU/LFU and the popularity-based caching.

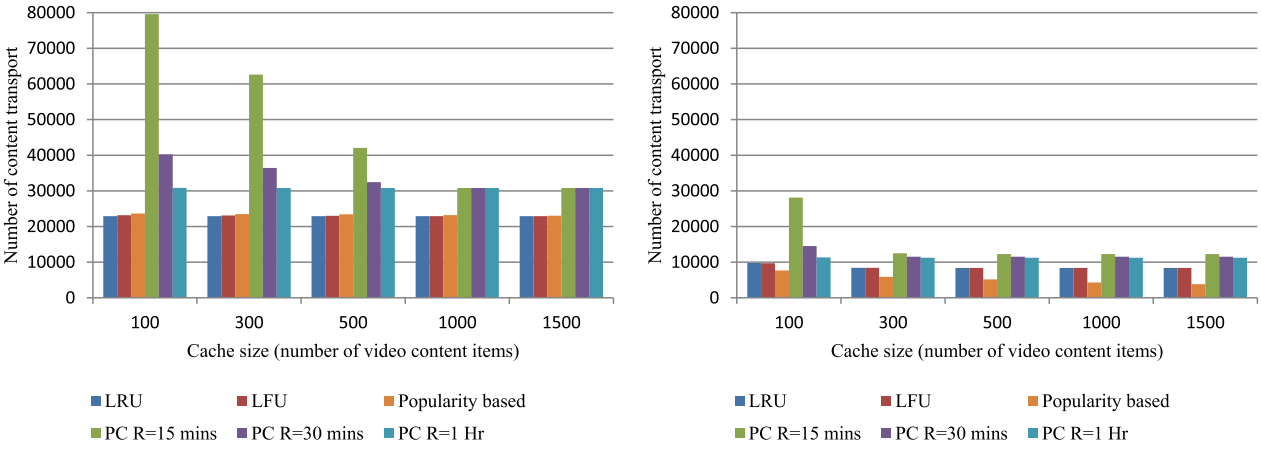
Figure 4 shows the number of content items, which need to be transported from CDN servers to DUs caches



(a) Low similarity in content requests, α = 0.4

(b) High similarity in content requests, α = 1.2

FIGURE 3. Cache hit rate vs. cache size when caching at DU with different similarity in content requests.



(a) Low similarity in content requests, α = 0.4

(b) High similarity in content requests, α = 1.2

FIGURE 4. Number of content items to be transported vs. different cache sizes when caching at DU with different similarity in content requests.

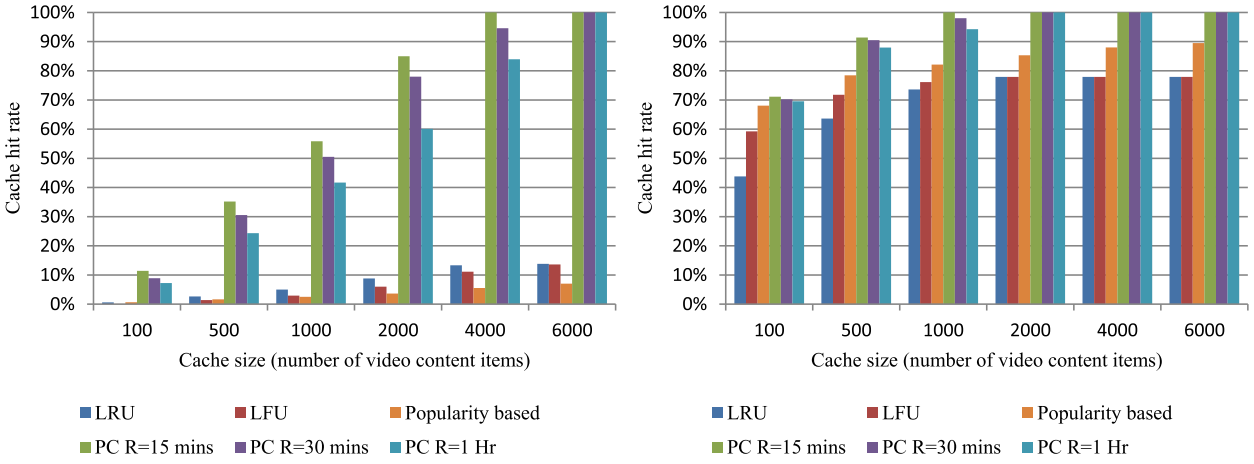
when adopting different caching algorithms. In general, the PC algorithm needs to transport more content items than LRU/LFU and the popularity-based caching due to the use of a proactive refresh method. Moreover, more content items need to be transported as the refresh period is short and, the number of content items that need to be transported for all algorithms decreases as the cache size increases because an increase in cache size allows more content items to be stored, and hence, reduces the requirement of transporting contents from the CDN server. Comparing Figures 4(a) and 4(b), we observe that the number of content items that need to be transported decreases dramatically as the similarity in user content requests increases because the greater similarity in user behavior reduces the additional transport requirements of new content items.

By observing the results shown in Figures 3 and 4, our proposed PC algorithm outperforms LRU, LFU, and the popularity-based caching in terms of cache hit rate. However, this comes at the cost of transporting additional video content items.

C. PERFORMANCE OF PROACTIVE CACHING WHEN CACHING AT THE CUs

In 5G access networks, the CUs are deployed at a relatively higher network hierarchical level than DUs. In general, the CU's cache capacity is much larger than that of the DU, and one CU can provide service for more users than one DU. In the simulation, we assume that a CU can support five DUs. Figure 5 shows the average cache hit rates of different algorithms with different cache sizes of a CU. We find that when the cache capacity of a CU reaches 6,000, the average hit rate of the PC algorithm is close to 100%, so the maximum cache size in our simulation is set to 6,000.

Similar to Figure 3(a), Figure 5(a) shows that the average cache hit rates of LRU, LFU, and popularity-based caching are very low (i.e., from less than 1% when the cache size is



(a) Low similarity in content requests, α = 0.4

(b) High similarity in content requests, α = 1.2

FIGURE 5. Cache hit rate vs. different cache sizes when caching at CU with different similarities in content requests.

100 to approximately 13.8% when the cache size increases to 6,000.) with low similarity in content requests ($\alpha = 0.4$). Due to the use of predictive analytics on user content requests, Figure 5(a) also shows that the PC algorithm improves the cache hit rate significantly compared to LRU, LFU and the popularity-based caching. Furthermore, the cache hit rate of PC decreases as the refresh period R increases. Additionally, by comparing Figures 5(a) and 5(b), we can see that the average cache hit rates increase significantly as the user similarity in content requests increases, especially for the conventional methods. As the cache size grows, the hit rate of the PC algorithm is more likely to reach 100% when the user similarity is high ($\alpha = 1.2$). Moreover, as the similarity in content requests increases, the performance gap between conventional algorithms and PC algorithms is significantly reduced.

Figure 6 shows the number of content items that need to be transported from CDN servers to the CU caches. Similar to Figure 4, the PC algorithm needs to transport more content items than the conventional algorithms due to the use of a proactive refresh method. As the refresh period of the PC algorithm increases, the number of content items that need to be transported decreases dramatically. The number of content items that need to be transported using all three algorithms also decreases as the cache size increases. Comparing Figure 6(a) and 6(b), we can also see that the number of content items that need to be transported decreases dramatically as the similarity in user content requests increases.

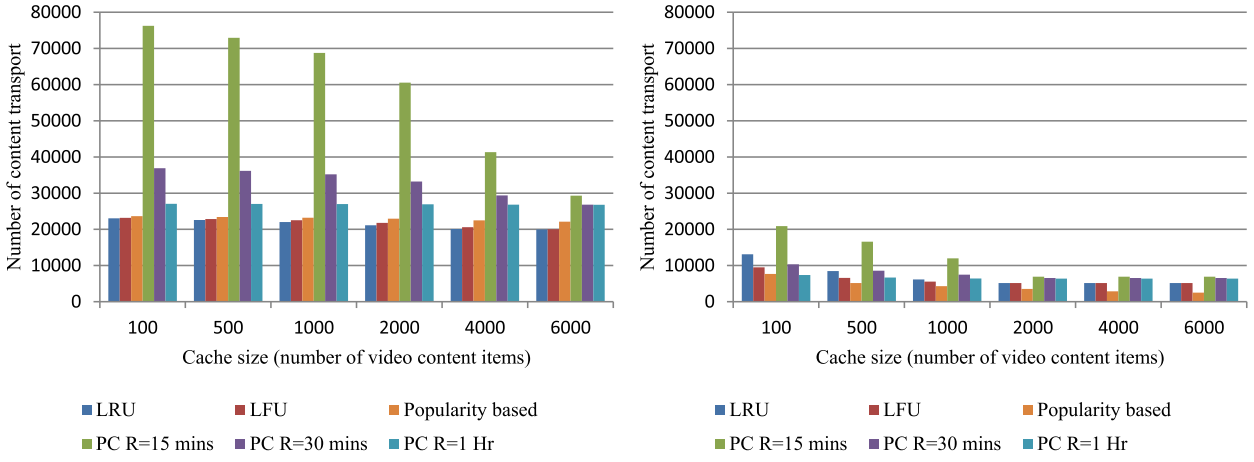
D. ENERGY CONSUMPTION OF CACHING AT THE DUs

If we deploy the edge caches at the DUs, the power consumption consists of two parts: the network transmission energy consumed by transmitting the content items that need to be updated from the CDN server to the caches, and the cache energy consumption of storing different numbers of video content items. We simulate the energy consumption using these cache updating algorithms, i.e., LRU, LFU, popularity-based caching and PC. We model the PC algorithm with three refresh cycles, by setting R to 15 minutes, 30 minutes, and 1 hour. The simulation results are shown in Figure 7.

As shown in Figure 7(a), when user similarity is low, the conventional algorithms such as LRU, LFU, and popularity-based caching cannot achieve a 100% user request hit rate even when storing up to 1,500 content items. The PC algorithm uses the proactive refresh technique, which can achieve a 100% hit rate when the cache usage is small. For example, if the refresh period R is set to 15 minutes, the hit rate can reach 100% when the number of cached content items is greater than or equal to 500. However, the extra cost is that more video content items need to be updated and transferred, which consumes more network transmission power.

In addition, when the PC algorithm is adopted, as the number of stored content items increases, the storage energy consumption increases linearly, but the transmission energy consumption decreases. Therefore, in the actual network deployment, the trade-off between cache energy consumption and the transport energy should be considered. Furthermore, as the refresh period R increases, more content items need to be stored in the DU caches to achieve a higher hit rate. For example, when R is 30 minutes, 1,000 content items need to be stored to achieve a 100% hit rate. In contrast, when R is 15 minutes, only 500 content items need to be stored. Moreover, as R increases, the transmission energy consumption decreases, thus the trade-off between R and total energy consumption needs to be jointly considered.

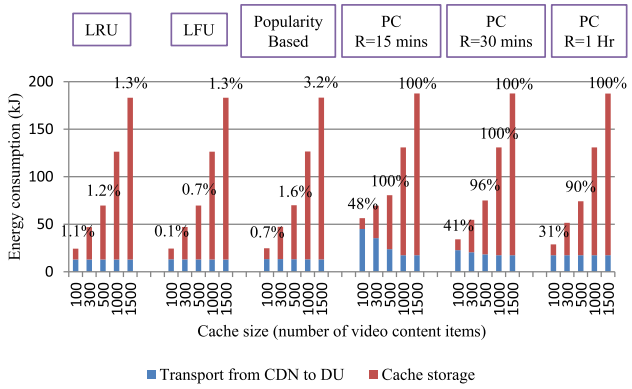
Comparing Figures 7(a) and 7(b), it can be seen that when the similarity in content requests increases, the transmission energy consumptions of all algorithms decreases, which is because the number of requests for popular content items increases significantly compared to the low similarity scenario, and hence, fewer content items need to be updated in each cache refresh cycle. Moreover, fewer content items



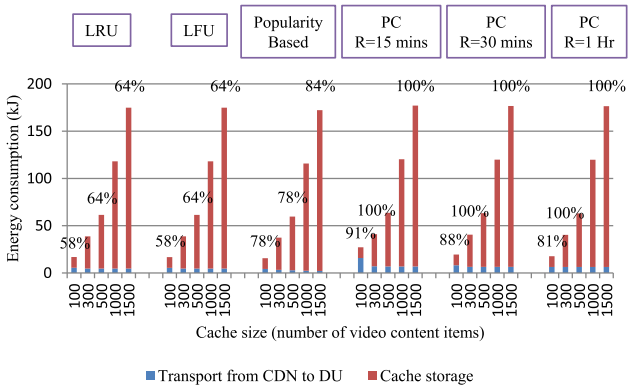
(a) Low similarity in content requests, α = 0.4

(b) High similarity in content requests, α = 1.2

FIGURE 6. Number of content items to be transported vs. different cache sizes when caching at CU with different similarities in content requests.

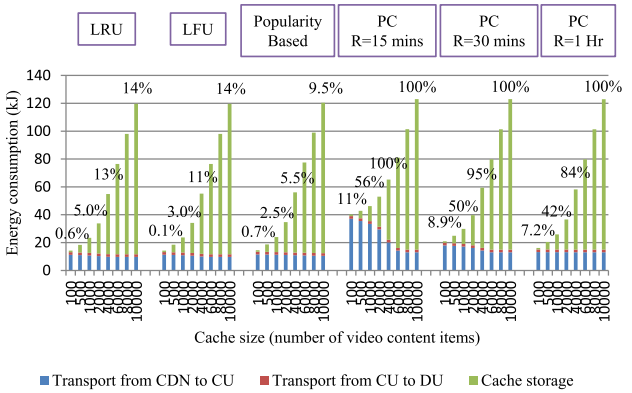


(a) Low similarity in content requests, α = 0.4

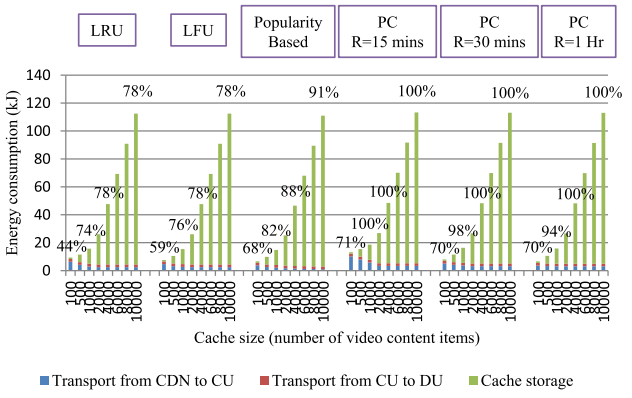


(b) High similarity in content requests, α = 1.2

FIGURE 7. Comparison of network energy consumption when caching at DU with different similarities in video content item requests. (Note: Numbers (%) above the bars indicate the average hit rate.)



(a) Low similarity in content requests, α = 0.4



(b) High similarity in content requests, α = 1.2

FIGURE 8. Comparison of network energy consumption when caching at the CU with different levels of similarity in video content items requests. (Note: Numbers (%) above the bars indicate the average hit rate.)

need to be stored on the DU to achieve a 100% hit rate. For example, when using the PC algorithm and setting R to 30 minutes, only 300 content items need to be stored with high similarity in content requests, and this number is 1,000 with low similarity in content requests.

E. ENERGY CONSUMPTION OF CACHING AT CUs

For edge caches deployed at the CUs, the total power consumption consists of three parts: (i) the transmission energy consumed by transmitting the content items that need to be updated from the CDN server to the CUs, (ii) the transmission

energy consumed by transmitting these content items from the CUs to the DUs, and (iii) the cache energy consumption of storing different numbers of video content items. The simulation results are shown in Figure 8. We observe that the PC algorithm, in theory, can relatively easily achieve a 100% hit rate but at the cost of higher transmission energy consumption compared to conventional algorithms such as the LRU and LFU.

Since the storage capacity of the CU is higher than that of the DU, when the user similarity is relatively high, the transmission energy from the CDN server to the CU is very small. Particularly, when the number of stored video content items is relatively large, the cache storage energy consumption is almost negligible compared to the transmission energy from the CU to DU.

In addition, by comparing Figures 7 and 8, caching at the CUs generally consumes less energy than caching at the DUs because each DU uses the same update and storage algorithm, which results in the same content items being stored in multiple DUs. However, this causes additional transfers of content items from the CDN server to the DUs, which consumes more storage and transmission power.

VI. CONCLUSION

The flexibility of the 5G network architecture has provided a platform for the deployment of MEC infrastructure. By deploying a large number of MEC servers, network content items can be cached in advance at the wireless edge to provide users with an ultimate QoE, extremely low latency and high bandwidth performance. However, due to limited storage capacity of edge caches, only top popular content items can be selected for storage to satisfy the QoS of mobile users. In addition, due to the differences in user similarity, the performance of the conventional cache updating algorithms, such as the LRU, LFU and the popularity-based caching, cannot meet the dynamic real-time changes of user requirements. In this paper, we introduce a predictive caching algorithm that utilizes big data analytics to predict user content requests and determine what content items need to be cached where in the network to achieve a better QoE. However, the cost of adopting this new algorithm is that more content items need to be updated and transferred, which consumes more network transmission energy. We performed a comprehensive simulation to assess the network transmission and storage energy consumption of our proposed PC algorithm under different refresh cycles. The simulation results provide useful insights for mobile operators to assess the trade-off between cache energy consumption and transport energy, and the trade-off between refresh cycles R and the total energy consumption of MEC. The key findings in this paper will provide a reference baseline for the energy-efficient deployment of edge caching in future wireless networks such as 5G and beyond.

REFERENCES

[1] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Netw.*, vol. 32, no. 2, pp. 78–84, Mar. 2018.

[2] M. Erol-Kantarci and S. Sukhmani, "Caching and computing at the edge for mobile augmented reality and virtual reality (AR/VR) in 5G," in *Ad Hoc Networks*. Cham, Switzerland: Springer, 2018, pp. 169–177.

[3] I. Chih-Lin, H. Li, J. Korhonen, J. Huang, and L. Han, "RAN revolution with NGFI (xHaul) for 5G," *J. Lightw. Technol.*, vol. 36, no. 2, pp. 541–550, Jan. 15, 2018.

[4] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098–3130, 4th Quart., 2018.

[5] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[6] A. Ksentini, P. A. Frangoudis, P. C. Amogh, and N. Nikaein, "Providing low latency guarantees for slicing-ready 5G systems via two-level MAC scheduling," *IEEE Netw.*, vol. 32, no. 6, pp. 116–123, Nov. 2018.

[7] S. Y. Lien, S. C. Hung, H. Hsu, and D. J. Deng, "Energy-optimal edge content cache and dissemination: Designs for practical network deployment," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 88–93, May 2018.

[8] T. T. Vu, D. T. Ngo, M. N. Dao, S. Durrani, and R. H. Middleton, "Spectral and energy efficiency maximization for content-centric C-RANs with edge caching," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6628–6642, Aug. 2018.

[9] H.-C. Hsieh, J.-L. Chen, and A. Benslimane, "5G virtualized multi-access edge computing platform for IoT applications," *J. Netw. Comput. Appl.*, vol. 115, pp. 94–102, Aug. 2018.

[10] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Unravelling the impact of temporal and geographical locality in content caching systems," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1839–1854, Oct. 2015.

[11] A. Brodersen, S. Scellato, and M. Wattenhofer, "Youtube around the world: Geographic popularity of videos," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 241–250.

[12] T. X. Vu, S. Chatzinotas, B. Ottersten, and T. Q. Duong, "Energy minimization for cache-assisted content delivery networks with wireless backhaul," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 332–335, Jun. 2018.

[13] M. Yan, C. A. Chan, W. Li, L. Lei, Q. Shuai, A. Gygax, and I. Chih-Lin, "Assessing the energy consumption of 5G wireless edge caching," in *Proc. IEEE Int. Conf. Commun.*, Shanghai, China, May 2019, pp. 1–6.

[14] Z. Luo, M. LiWang, Z. Lin, L. Huang, X. Du, and M. Guizani, "Energy-efficient caching for mobile edge computing in 5G networks," *Appl. Sci.*, vol. 7, no. 6, p. 557, May 2017.

[15] R. Amer, M. M. Butt, H. ElSawy, M. Bennis, J. Kibilda, and N. Marchetti, "On minimizing energy consumption for D2D clustered caching networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[16] G. Hasslinger, J. Heikkinen, K. Ntougias, F. Hasslinger, and O. Hohlfeld, "Optimum caching versus LRU and LFU: Comparison and combined limited look-ahead strategies," in *Proc. 16th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw.*, Shanghai, China, May 2018, pp. 1–6.

[17] G. Hasslinger, K. Ntougias, F. Hasslinger, and O. Hohlfeld, "Performance evaluation for new Web caching strategies combining LRU with score based object selection," *Comput. Netw.*, vol. 125, pp. 172–186, Oct. 2017.

[18] C. Bernardini, T. Silverston, and O. Festor, "MPC: Popularity-based caching strategy for content centric networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 3619–3623.

[19] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: Proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[20] J. Qiao, Y. He, and X. S. Shen, "Proactive caching for mobile video streaming in millimeter wave 5G networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[21] J. Hu, Y. Lai, A. Peng, X. Hong, and J. Shi, "Proactive content delivery with service-tier awareness and user demand prediction," *Electronics*, vol. 8, no. 1, p. 50, Jan. 2019.

[22] A. Said, S. W. H. Shah, H. Farooq, A. N. Mian, A. Imran, and J. Crowcroft, "Proactive caching at the edge leveraging influential user detection in cellular D2D networks," *Future Internet*, vol. 10, no. 10, p. 93, Sep. 2018.

[23] I. A. Ridhawi, N. Mostafa, Y. Kotb, M. Aloqaily, and I. Abualhaol, "Data caching and selection in 5G networks using F2F communication," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Montreal, QC, Canada, Oct. 2017, pp. 1–6.

[24] M. Yan, C. A. Chan, W. Li, I. Chih-Lin, S. Bian, A. F. Gygax, C. Leckie, K. Hinton, E. Wong, and A. Nirmalathas, "Network energy consumption assessment of conventional mobile services and over-the-top instant messaging applications," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3168–3180, Dec. 2016.
[25] M. Yan, C. Chan, A. Gygax, J. Yan, L. Campbell, A. Nirmalathas, and C. Leckie, "Modeling the total energy consumption of mobile network services and applications," *Energies*, vol. 12, no. 1, p. 184, Jan. 2019.
[26] I. Bor-Yaliniz, M. Salem, G. Senerath, and H. Yanikomeroglu, "Is 5G ready for drones: A look into contemporary and prospective wireless networks from a standardization perspective," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 18–27, Feb. 2019.
[27] S. Han, I. Chih-Lin, G. Li, S. Wang, and Q. Sun, "Big data enabled mobile network design for 5G and beyond," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 150–157, Jul. 2017.
[28] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: Moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.
[29] M. B. Karimi, A. Isazadeh, and A. M. Rahmani, "QoS-aware service composition in cloud computing using data mining techniques and genetic algorithm," *J. Supercomput.*, vol. 73, no. 4, pp. 1387–1415, Apr. 2017.
[30] L. Cao and M. Grabchak, "Smoothly truncated levy walks: Toward a realistic mobility model," in *Proc. Int. Perform. Comput. Commun. Conf.*, Dec. 2014, pp. 1–8.
[31] A. Vishwanath, F. Jalali, K. Hinton, T. Alpcan, R. W. A. Ayre, and R. S. Tucker, "Energy consumption comparison of interactive cloud-based and local applications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 616–626, Apr. 2015.

MING YAN (M'19) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2002, and the M.S. and Ph.D. degrees in communication and information systems from the Communication University of China (CUC), Beijing, China, in 2006 and 2012, respectively. From 2014 to 2015, he was a Visiting Research Scholar with the Center for Energy-Efficient Telecommunications, The University of Melbourne, where he was involved in developing new energy models for mobile services. He is currently an Associate Researcher with the School of Information and Telecommunications Engineering, CUC. His research interests include future wireless systems, green technologies in wireless communication systems, mobile wireless networks, and mobile multimedia broadcast technologies.

CHIEN AUN CHAN (M'10) received the Ph.D. degree in electrical engineering from The University of Melbourne (UoM), VIC, Australia, in 2010. In 2011, he joined the Centre for Energy-Efficient Telecommunications, UoM, where he was involved in developing new modeling and energy-efficient techniques for cloud applications and mobile services. Since 2017, he has been a Research Fellow with the Department of Electrical and Electronic Engineering, UoM, where he is involved in developing future wireless systems, smart wearable systems, big data network analytics, and mobile edge computing systems. His research interests include mobile wireless networks, the Internet of Things, green communications and networking, big data network analytics, cloud applications and services, network service virtualization, and optical networks.

WENWEN LI received the M.S. degree in telecommunications engineering from Xidian University, in 2008. She joined the Terminal Technology Department, China Mobile Research Institute, as the Project Manager. In 2013, she joined the Green Communications Research Center, where she was involved in energy-efficient communications and wireless interface protocol of TD-SCDMA/TD-LTE/WLAN for network end-to-end green ecosystems. She holds five patents. She has authored extensively in her research fields.

LING LEI received the Ph.D. degree in electrical engineering from Beihang University, China, in 2012. She is currently a Lecturer with the School of Information and Telecommunications Engineering, Communication University of China (CUC), Beijing. She had been a Visiting Scholar with the University of California, Berkeley, and the Chinese University of Hong Kong. Her research interests include wireless communication technologies, mobile wireless networks, and mobile multimedia broadcast technologies.

ANDRÉ F. GYGAX received the Ph.D. degree in finance from the University of Melbourne, VIC, Australia. He is currently with the Department of Finance, University of Melbourne, as a Faculty Member. He is a Fellow of the Center for Business Analytics, Melbourne Business School, a Research Associate with the Center for Energy-Efficient Telecommunications, Melbourne School of Engineering, and a Research Associate with the Melbourne Networked Society Institute. His work focuses on social and physical networks in theory and practice with applications in finance, economics, sociology, and environmental sciences.

CHIH-LIN I (SM'03) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA. She has almost 30 years of experience in the wireless communication area. She was with world-class companies and research institutes, including the Wireless Communication Fundamental Research Department of AT&T Bell Labs; Headquarters of AT&T as the Director of Wireless Communications Infrastructure and Access Technology; ITRI of Taiwan as the Director of Wireless Communication Technology; and Hong Kong ASTRI as the Vice-President and the Founding GD of the Communications Technology Domain. She was an Elected Board Member of the IEEE Communications Society. She received the IEEE TRANSACTIONS ON COMMUNICATIONS Stephen Rice Best Paper Award. She was a winner of the CCCP National 1000 Talent Program. She is currently the China Mobile Chief Scientist of Wireless Technologies in charge of advanced wireless communication research and development efforts of the China Mobile Research Institute. She established the Green Communications Research Center of China Mobile, spearheading major initiatives, including 5G key technology research and development, high-energy efficiency system architecture, technologies, and devices, green applications, and C-RAN and soft base station. She was the Chair of the Communications Society Meeting and Conferences Board, and the Founding Chair of the IEEE WCNC Steering Committee. She is currently the Chair of FuTURE Forum 5G SIG, an Executive Board Member of GreenTouch, and a Network Operator Council Member of ETSI NFV.