In-Network Caching: An Efficient Content Distribution Strategy for Mobile Networks

Shaowei Wang, Tianyu Wang, and Xun Cao

Abstract

The sharp increase in wireless devices yields a huge amount of mobile data traffic, which has made either the radio access network or the core network of current mobile communication systems seriously overloaded. In-network caching arises as a promising solution to this burning issue. By introducing content centric networking infrastructure, popular content files can be intelligently stored in the radio access network so that redundant transmissions through the core network can be significantly reduced, which can substantially alleviate the load of both the core network and the backhauls between the radio access network and the core network. In this article, we discuss what to cache, how to cache and how to evaluate the performance of a cache-enabled mobile network. We first discuss the instructive caching policies, then propose reasonable performance evaluation metrics for these caching policies. We present detailed numerical results demonstrating remarkable gains by the in-network caching technique. Finally, we discuss related research directions, opportunities and challenges.

Introduction

With the fast development of the Internet and the growing popularity of smart mobile terminals, data business has been increasing explosively in mobile communication networks. Global mobile data traffic has broken through 24.3 Exabytes per month [1], which is straining the network resource of mobile communication systems from two perspectives, the radio access network and the core network, and is also putting service providers under enormous pressure in the areas of operation and maintenance. As a response to the rising challenges facing mobile networks, great effort has been made to enhance system capacity including developing advanced signal processing techniques such as massive MIMO [2, 3], designing more efficient radio access schemes such as heterogeneous networks [4, 5] and cloud radio access networks [6], exploiting the novel spectrum utilization paradigm such as cognitive radio [7], and exploring promising network architecture such as locator/identifier split networking [8]. As new applications emerge in an endless stream, such as video communications [9], online games, electronic commerce and autonomous vehicles, the mobile network is also faced with other urgent issues including system latency, network safety and mass data processing [10].

To address the problems mentioned above, the fifth generation (5G) mobile communication system appears with original designs for both novel network architectures and data transmission techniques to provide users with guaranteed quality of service (QoS) or quality of experience (QoE). 5G is expected to achieve 1000 times higher system capacity, 10 times spectrum efficiency and energy efficiency, and 25 times average cell throughput than those of 4G. Specifically, 5G mobile network designs are interested in exploring application-aware approaches to achieve ubiquitous communications between not only people to people but also machine to machine wherever they are or whenever they are needed, by whatever intermediate they adopt. In brief, a desired vision for 5G is to implement seamless and pervasive connectivity between anybody, anything, and anytime all over the world.

In the past decade, the growing demand in mobile networks has transferred from traditional text messages and voice services to video streaming and popular content sharing. Therefore, the information-centric networking paradigm is deemed as a promising solution to making 5G mobile networks more affordable for content distribution and information sharing, of which the content distribution network (CDN) is a concrete realization where content files are intelligently cached at the base stations (BSs) to reduce duplicated downloads and thus to optimize the user perceived latency in the mobile communication system. Figure 1 is the illustration of a cache-enabled mobile communication system, where the system is abstracted into a three-tier network architecture: upper Internet layer, middle evolved packet core (EPC) layer and lower radio access network (RAN) layer. The Internet CDN servers generally provide content that would be requested by the users who get access to mobile networks. The EPC consists of the serving gateway (S-GW), packet data network gateway (P-GW) and mobility management entity (MME), while the RAN usually includes the BSs for radio access.

As can be seen from Fig. 1, caching content files at the BSs has great potential to improve the system performance of mobile networks. First, latency perceived by users could be reduced if the target files are fetched from the caches at BSs directly since the BSs are always much closer to the users than either the EPC gateways or the Internet CDN servers, which not only improves the QoE of users, but also makes the mobile network capable of supporting delay-sensitive and high data rate services such as virtual reality (VR) and aug-

Digital Object Identifier:
10.1109/MWC.2019.1800449

The authors are with Nanjing University.

mented reality (AR); second, the load of backhauls which connect RAN and EPC can be alleviated since users do not need the help of backhauls to get the requested files that are cached at BSs. As a result, the operation and maintenance cost of mobile networks can be reduced, as well as the system energy consumption.

While caching at BSs can significantly enhance the system performance of mobile networks, what to cache and how to cache are the thought-provoking problems that should be carefully investigated so as to exploit the potential of the caching technique at the mobile network edge. Statistical analysis indicates that only a few popular content files are repeatedly requested by users among all available content files, resulting in duplicate transmissions from the remote Internet and a considerable expense for mobile service providers. Obviously, the popularity of content files should be exploited to enhance network performance, based on which classic caching policies including the least recently used scheme, the least frequently used scheme and the first-in first-out scheme could be feasible ones for applications. By storing content files in the RAN side of the mobile network, massive demands from end users can be directly satisfied at the BSs equipped with caches, instead of accessing the EPC and the Internet CDN servers. However, the performance of the cache-enabled mobile network is worthy of attention from a variety of sides, such as the variations of content popularity in different times and locations, the heterogeneity of multimedia content, the affordable cache capacity of BSs, the density of users and the network states such as instantaneous radio access capability of BSs.

In this article, we discuss diverse cache techniques that could be employed in mobile networks, and uncover potential weaknesses and challenges for future research. We focus on the following issues:

- What and how to cache: a detailed consideration on what and how to cache the requested files in mobile networks, especially in the RAN side, is given, including feasible and efficient caching policies and their individual advantages and weaknesses.
- How to evaluate the caching performance: feasible evaluation metrics for the cache-enabled mobile network are discussed from different perspectives such as delay and access cost.
- Open questions and solutions: software-defined networking is proposed as a possible scheme to integrate cache techniques with practical mobile networks; on the other side, achieving trade-offs among radio resource, storage capacity and computing capability should be explored in depth so as to further exploit the potential of caching techniques in mobile networks.

Instructive Cache Placement Policies

As a promising technique to cope with the heavy backhaul overload issue in mobile networks, caching in the RAN attracts a lot of attention. In-network caching is the most important component for the emerging information-centric networking facing mobile networks, where replication-based multiparty communication and interaction models decouple senders and receivers, providing diverse design choices and features. Possible technical research directions for wireless content caching in 5G mobile networks can be found in [11], where the involved stakeholders including end users, network operators and device vendors are analyzed. In [12], the related caching techniques in the forthcoming 5G networks including both the EPC and the RAN have been investigated, as well as the various advantages and relevant opportunities and challenges, and a novel edge caching scheme for content-centric networking architecture is proposed. In [13], the fundamental limits of caching are introduced from the perspective of the global caching gain that depends on the aggregate global cache size. The performance gain heavily depends on the cache deployment strategy decided by mobile service providers. The state-of-the-art studies show that cache placement policies need to be designed elaborately in practical mobile communication systems. We summarize the main caching methods that could be instructive for deploying cache in mobile networks.

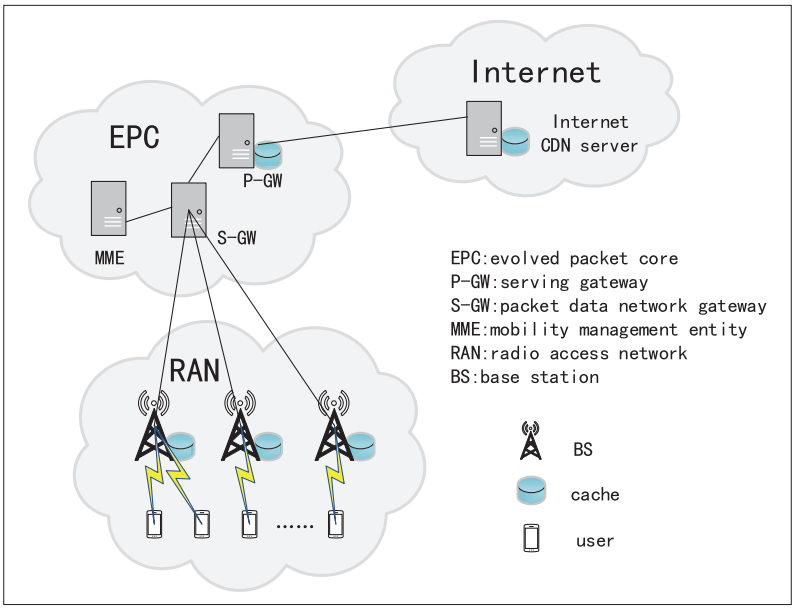

FIGURE 1. Illustration of cache-enabled mobile networks.

Random Caching Policy (RCP): The content files are randomly cached at each BS equipped with limited cache capacity. Since the RCP requires no statistical information about content files, it is easy to implement and is widely used in the Internet CDN. Another advantage of the RCP is that the diversity of content files can be guaranteed since these files are stored at each BS with the same probability. Thus, the RCP is suitable for the following scenarios. The requests of users are diverse, random, and have little relationship with personal preferences; the number of files that can be shared is relatively small; the cache capacity of BSs is relatively large so that a large proportion of content files are available for the users in the RAN side. Except for these scenarios, the RCP generally does not work well since it takes no prior information about content files into consideration.

Popularity-Driven Caching Policy (PCP): Investigations show that only a fraction of content files are download duplicately and contribute most of the traffic load in mobile networks. Thus, if popular content files, for example, music videos and exciting sports videos, are stored at the BSs in the RAN, redundant downloads can be reduced significantly and consequently the backhaul overload can be

Placing the content files according to their popularity is a promising strategy for mobile service providers, which means that the most popular files are always stored as many as possible at the BSs. However, such a popularity-driven caching policy may seriously deteriorate the QoE of the users requesting the files that cannot be found at the BSs.

alleviated. Network traffic analysis shows that the relative frequency of content files follows Zipf's law [14] which illustrates the relationship between the relative probability of a request from users for those content files and their ranking in popularity. Specifically, the relative probability of content requests generally follows Zipf-like distribution, that is, the probability of the jth most popular content file is $j^{-\gamma}$, where γ is the Zipf component indicating the skew of the popularity distribution of content files.

Consider a cache-enabled mobile network, where the cache capacity of each BS is limited and the requests for content files are sent to BSs continuously. The popularity of these files follows a distribution $\{z_j\}$ with $z_1 \geq z_2 \geq \ ... \geq z_J$. Obviously, z_1 is the most popular file while z_J is the least popular one. According to Zipf's law, we have

$$z_j = \frac{j^{-\gamma}}{\sum_{j=1}^{J} j^{-\gamma}}$$

and $\Sigma_{j=1}^{J} z_j = 1$. In reality, most users are usually interested in the popular content files and only a portion of the content files are frequently requested by the users. It reveals that placing the content files according to their popularity is a promising strategy for mobile service providers, which means that the most popular files are always stored as many as possible at the BSs. However, such a popularity-driven caching policy may seriously deteriorate the QoE of the users requesting the files that cannot be found at the BSs.

Hit-Probability Based Optimal Caching Policy (HCP): The hit probability is defined as the probability that the requests from users are directly processed at the local BSs in the RAN, not fetched from the EPC via backhauls. Higher hit probability leads to lower access cost, as well as better QoE for users. Developing an efficient caching strategy from the perspective of maximizing the total hit probability of all users is reasonable for practical mobile networks where users are randomly located around BSs equipped with capacity-limited cache. From the viewpoint of radio coverage, a given user can be served by one or more BSs, or none of the BSs can serve this user. Denote the number of BSs that can provide radio coverage for the user by θ, which is a random variable related to various network parameters, the radio coverage models that can be utilized to characterize the distribution of θ include signal-to-interference-plus-noise-ratio (SINR) model, Boolean model and overlaid 2-network model [15]. The SINR model describes the coverage quality at the origin in terms of SINR, while the Boolean model is suitable for noise-limited cases, that is, the interference is relatively small compared to noise. As for the overlaid 2-network model, it assumes that two or more networks run in parallel with different infrastructure and orthogonal resources by the same provider to serve the users in a given area. The coverage probability θ of the ith network can be described as $P_i := \mathbf{P}(\theta = i)$.

We define the caching probability of content file j as $f_j := \mathbf{P}(j \in \Omega_n)$, where Ω_n denotes the set of the cached files at BS n. If a user can download the requested file from one of the BSs that can provide radio cover for him/her, we regard it as a hit. Thus, if a user requests file j, the hit probability can be calculated mathematically as $1 - \Sigma_{i=0}^{\infty} P_i (1 - f_j)^i$, where $\Sigma_{i=0}^{\infty} P_i (1 - f_j)^i$ indicates the probability that none of the BSs covering the user stores the file he/she requests. However, we cannot know which content files the users would request in advance. Generally, popular files are more probable to be requested while other files are requested with a lower probability. We can model the distribution of the request probability by Zipf-like distribution [14], which means that each file has a probability z_j to be requested by users. Then, the total hit probability for a typical user can be defined as $H(f_1, f_2, \ ... \ , f_J) := 1 - \Sigma_{j=1}^{J} z_j \Sigma_{i=0}^{\infty} P_i (1 - f_j)^i$.

Apparently, the total hit probability of a typical user depends on the request probability z_j, the coverage probability P_i and the caching probability f_j. With given Zipf component γ and chosen coverage model, there exists an optimal caching probability for all the content files to maximize the total hit probability of the typical user. Denote the cache capacity of each BS by L_n, mathematically, the optimal caching probability of each file f_j can be obtained by solving the following optimization problem:

$$\begin{aligned} &\max_{f_j} H(f_1, f_2, \ldots, f_J) \\ s.t. \quad &C_1 : \sum_{j=1}^{J} f_j \leq L_n, \forall n \in \mathcal{N}, \\ &C_2 : 0 \leq f_j \leq 1, \forall j \in \mathcal{J}. \end{aligned} \tag{1}$$

When we get the optimal caching probability of each file f_j, we can decide the cache placement with joint considering f_j and cache capacity. This strategy is also described as probabilistic placement policy in [15].

Metrics for Cache Placement Performance

Employing caching techniques in mobile networks can yield quite a few benefits, such as saving operation expense, relieving backhaul burden and improving system energy efficiency. To measure the gains for a cache-enabled mobile network, we need a general metric that can precisely evaluate the performance improvement from different perspectives. The reason is as follows. Different types of performance indexes may require very different system configuration parameters even network architecture with big difference; however, it is impossible to design cache policies with different system parameters for a given mobile network, not to mention network architecture. We need to achieve a trade-off among different performance indicators so as to design a feasible cache configuration scheme for the practical mobile network.

Consider a cache-enabled mobile network with N BSs and K users. The set of BSs and users are denoted as $\mathcal{N} = \{1, 2, \ ... \ , N\}$ and $\mathcal{K} = \{1, 2, \ ... \ , K\}$, respectively. The requests from users can be satisfied in two ways:

- If the requested files are just stored at the BSs that can also provide radio coverage to the users, the users can directly fetch the content files from these local BSs.
- If the requested files are not found in the caches at the BSs which provide radio coverage to the users, the BSs forward the requests to the EPC via backhaul links. The latter is usually costly, as discussed above.

Figure 2 is an illustration of cutting down transmission redundancy by employing caching techniques. Figure 2a is a hierarchical topology of a mobile network without deploying caches while

Fig. 2b is the cache-enabled mobile network where caches are deployed at both the EPC side and the RAN side, and the BSs are connected to the EPC via wired or wireless backhaul with limited capacity. The colorized rectangles represent the caches that are filled with content files. The dotted lines represent the content requests, where different colors indicate different requested files. As shown in Fig. 2a, without any in-network content caching, if a user sends a file request, it will require a backhaul link from the EPC to the BS, leading to a lot of traffic demands with duplications. In contrast, as illustrated in Fig. 2b, the vast majority of requests from users can be satisfied locally by the BSs equipped with caches in the RAN. The backhaul burden is significantly alleviated with the reduction of duplicate transmissions, as well as the core network. Moreover, most users are spared from suffering long delay of content delivery. In other words, both users and mobile service providers can benefit from the caching scheme shown in Fig. 2b. We elaborate this point taking the following two aspects for example and conclude that different metrics can be merged in a universal framework so that we can focus on a simple but efficient one to enhance system performance for mobile networks.

User-Perceived Latency: Latency is the key for the delay-sensitive services provided by 5G and beyond mobile networks, where utra-reliable and low latency communications business, such as 360-degree video and AR/VR, self-driving cars and smart grid, should be supported. As is illustrated in Fig. 2, two kinds of user perceived latency exists in mobile networks: wireless transmission delay, denoted as $D^T_{k,n}$, indicating the transmission delay between user k and its associated BS n; and backhaul delay, denoted as $D^B_{k,n}$, indicating the delay generated by the BSs in the RAN to visit the EPC for file downloading. Generally, wireless transmission delay $D^T_{k,n}$ is related to the size of file j (denoted by S_j) that user k requests, and the achievable transmission rate $r_{k,n}$ between user k and its associated BS n. The mathematical formulation of $D^T_{k,n}$ can be defined as $D^T_{k,n} = S_j/r_{k,n}$. The backhaul delay $D^B_{k,n}$ is affected by various factors including the average link distance between the BSs and the EPC, the density of BSs and the average traffic load of the mobile network. Backhaul delay usually can be described as a random variable following exponential distribution. As aforementioned, the download of files for users is operated as the following procedure. For a given user k requesting file j, if its associated BS n just caches file j, user k can obtain the requested file directly in the RAN, and their perceived latency is $D^T_{k,n}$. However, if its associated BS n does not cache file j, the request sent to BS n from user k will be further transferred to the EPC. File j would be downloaded from the EPC through the backhaul to BS n, then user k can get it from BS n. Obviously, this roundtrip generates extra backhaul delay, and the total delay for user k will be changed to $D^T_{k,n} + D^B_{k,n}$. For simplicity, we use a binary variable $\rho_{k,n}$ to indicate whether extra backhaul delay is involved, with $\rho_{k,n} = 1$ implying the existence of backhaul delay and $\rho_{k,n} = 0$, otherwise. Thus, the delay for user k requesting file j if the user is associated with BS n can be denoted by $D_{k,n} = D^T_{k,n} + \rho_{k,n}D^B_{k,n}$. The latency metric is straightforward; however, delay jitter of wired/wireless backhauls cannot be completely avoided since diverse and complex factors in practical mobile networks make latency hard to model accurately.

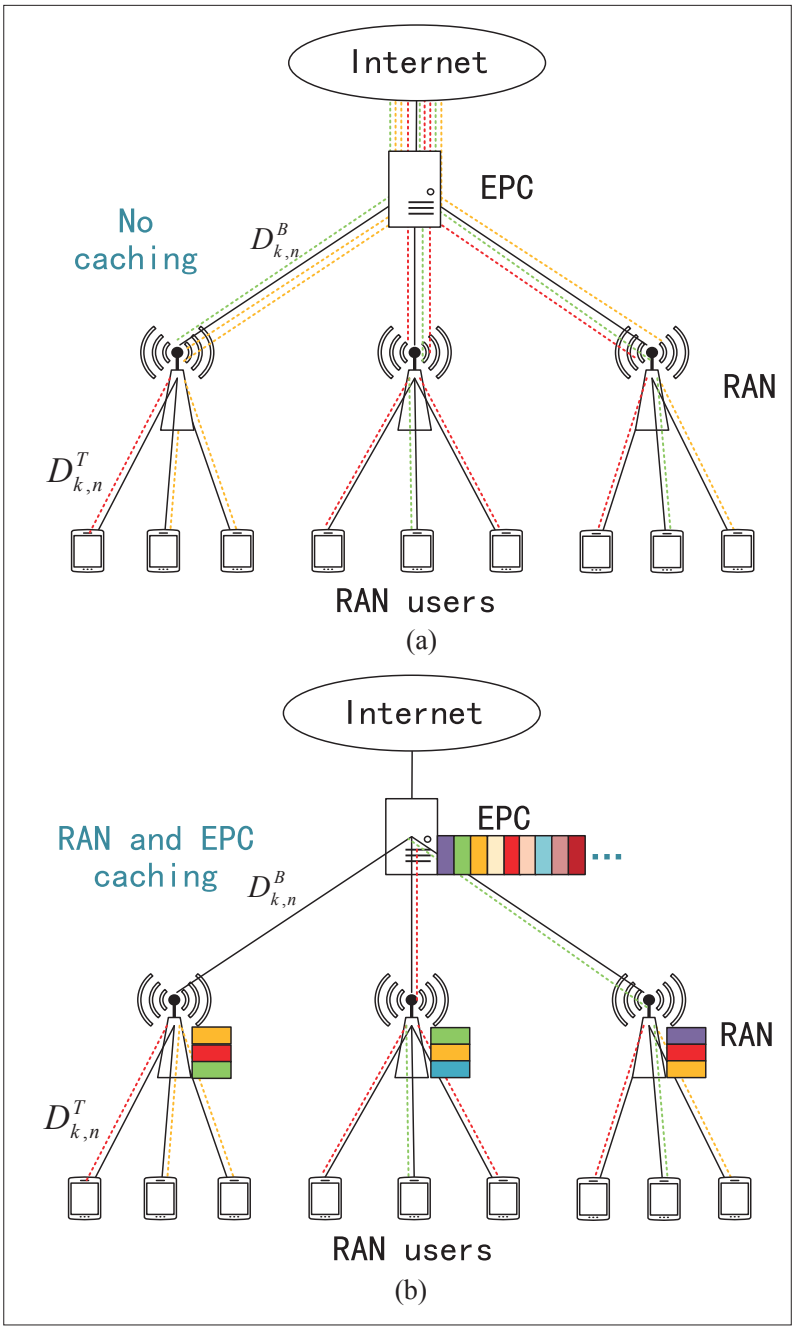

FIGURE 2. Comparison of a) no caching, and b) EPC and RAN caching.

Network Access Cost: From the viewpoint of service providers, caching at the BSs can alleviate the burden of backhauls, as well as the core network. We can simply call it access cost for the mobile network to serve a user, which is also deemed as a significative evaluation metric for caching performance. Again, there exists two kinds of access cost: the cost for transferring content files from the EPC to the BSs, denoted by c_0, and the cost for transferring content from the BSs to the users, denoted as $c_{k,n}$. $c_{k,n}$ is generally much less than c_0 in practical networks since it is always much cheaper for the users to get the target files from the BSs caching the files than fetching the files from the original server in the EPC if considering the additional consumption of the backhaul

In all scenarios, the HCP outperforms the other two policies since it not only takes into consideration the popularity of content files, but also focuses on the hit probability explicitly, making more requests be satisfied in the RAN side, which can decrease the user perceived latency and the network access cost jointly.

Parameter	Value
Radius of area	2 km
Total bandwidth	10 MHz
Transmission power budget	40 W
Thermal noise PSD	184 dBm/Hz
Path loss	$140.7 + 36.7\log_{10}(Distance)$
Zipf parameter	[0.5 1 1.5 2]
SINR threshold	[–10 0 10 20 40 100]
Number of users	[50 100 150 200 250]
Number of BSs	[20 25 30 35 40]
Number of files	[50 100 150 200 250]
Size of cache	[2 4 6 8 10]

TABLE 1. Simulation parameters.

resource and the core network resource. Consequently, the access cost for user k who requests file j, and is associated with BS n, can be defined as $c_{k,n} = c_{k,n} + \rho_{k,n}c_0$, where binary variable $\rho_{k,n}$ indicates whether the extra backhaul access cost exists or not, that is, $\rho_{k,n} = 1$ if the backhaul access cost exists, $\rho_{k,n} = 0$ otherwise. Notice that the access cost involves but is not limited to monetary expense, such as the total electric charge and broadband cost paid by users, the operation expense of service providers, and other incalculable expenses. The network access cost can also be the congestion levels on various network links, or just the system energy consumption.

We can observe that the metrics from different perspectives for cache performance measurement share the same starting point. Essentially, it is to make the requests from users be satisfied in the RAN side as many as possible so that the user-perceived latency and the network access cost can be reduced. Thus, although there exist various caching performance metrics, the starting point remains consistent among them, which may simplify system design for the cache deployment in practical mobile networks.

Performance Evaluation

Consider a cache-enabled mobile network where BSs are homogeneously deployed and users are randomly located around BSs uniformly. The Zipf-like distribution is exploited to characterize the content popularity. A user is considered to be covered by a BS as long as the SINR provided by the BS exceeds a given threshold. The list of parameters is given in Table 1. For the performance curves, these different parameters of interest are considered: SINR threshold, number of files, cache size of BSs, number of BSs, and number of users. The performance of the total hit probability, the average delay of the users and the network access cost are shown in Fig. 3.

The SINR Threshold: As the SINR threshold increases, the total hit probability decreases due to the difficulty brought by the high SINR required to provide radio coverage the users for BSs. On the other hand, the total hit probability increases as the increasing of Zipf parameter γ which determines the diversity of content files. Specifically, the larger γ is, the fewer popular files account for most of the content files.

The Number of Users: The average delay and the total network cost of all users are proportional to the number of users for all three caching policies. It can be explained as follows. The number of users that can be served by a given set of BSs is limited due to the bandwidth and the power budgets of the BSs, so the growth of users makes more users not be served by these BSs, even though the requested files have been stored in the caches at these BSs. These users have to wait in line or download the files from the core network via backhauls, increasing delay and access cost.

The number of BSs: The average delay of users is inversely proportional to the number of BSs for the HCP and the RCP, because more files can be stored in the caches at the BSs so that more requests of the users can be afforded by these BSs in this condition. More requests are directly satisfied in the RAN side so the total delay can be reduced. For the PCP, the files stored at the BSs are fixed and thus the continuous growth of BSs does not have significant affect on its performance. We can also see that the average delay varies only slightly for the PCP.

The Number of Files: The growth of files increases the average delay of users, as well as the total network access cost. Since the number of BSs and the cache size are fixed, the total number of files is also fixed. It decreases the diversity of the cached files, which correspondingly increases the difficulty of hitting user requests.

The Size of Caches: When the cache size becomes larger, the total access cost is reduced significantly. Larger cache size means that more files can be stored at the BSs so that the stored files can be much more diverse, which increases the hit probability of users and leads to the reduction of access cost.

In all scenarios, the HCP outperforms the other two policies since it not only takes into consideration the popularity of content files, but also focuses on the hit probability explicitly, making more requests be satisfied in the RAN side, which can decrease the user perceived latency and the network access cost jointly.

Research Directions

Integration with Software-Defined Networking

Software-defined networking (SDN) could be a promising architecture for deploying the cache-enabled mobile network. SDN is described as a novel network paradigm for which the control plane is separated from the underlaying data plane to be responsible for the overall network behavior. SDN has many advantages over legacy methods. First, it is much more convenient to implement new ideas for the network via a software program since it is easier to manipulate than the traditional fixed commands in specific devices. Second, compared to the distributed management paradigm, the SDN owns a centralized control layer for network configuration, which frees the operator from configuring all demanding network devices individually. Moreover, it is possible to learn the global knowledge of network and deals with the forwarding traffic in a logically single location. In this regard, it is quite promising to make

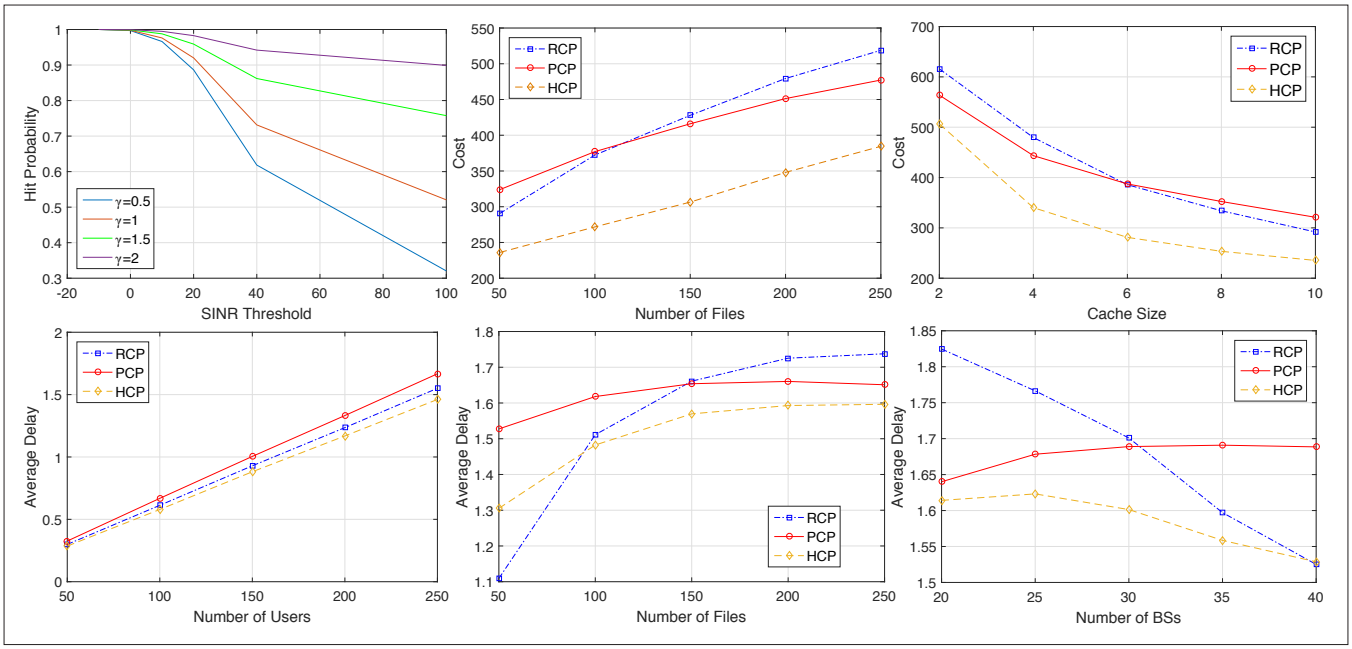

FIGURE 3. Hit probability, average delay and total cost of network with respect to SINR threshold, cache size of BSs, number of files, BSs and users.

a blueprint for a cache based software-defined networking architecture by combining cache management and SDN together to improve system performance.

Figure 4 illustrates a cache-based SDN architecture for mobile networks. Since cache performance is affected by various factors, comprehensive analysis of them should be conducted to generate a global and efficient cache deployment policy. A real-time decision making module is necessary since the popularity of content files, the density of users and the network state are generally time-varying. As can be seen from Fig. 4, information collection and analysis including proactive learning and prediction of popularity of content files, the statistics and analysis for user density, and the management of the load balance of BSs can be conducted at the forwarding/switching plane. The network operators can perform resource allocation and decision making from the global view of the network in the control plane. Finally, efficient caching-and-delivery schemes are established to provide users with better QoS.

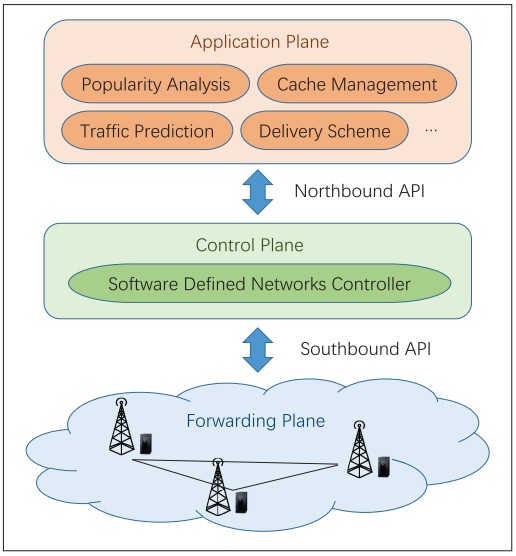

FIGURE 4. SDN-based cache scheme for mobile networks.

Trade-Off Among Radio Resource, Storage Capacity and Computing Capability

As discussed above, in-network caching attempts to reduce redundant transmissions in order to relieve the burden of the core network and provide a better user experience. Actually, this improvement is at the cost of deploying cache memories in both the RAN side and the EPC side. In one respect, many BSs that caches popular files need to be active for a long time so as to support caching effectively, which generates significant energy expense. On the other side, deciding what to cache and how to cache, including analyzing the popularity of files, traffic prediction and cache scheduling, needs a large amount of computing resources, which could be another bottleneck for mobile networks. Thus, there should be a trade-off among radio resources including the bandwidth and power budgets of BSs, storage capacity and computing capability of both the core network and the BSs. Figure 5 illustrates the trade-off among these requirements.

On the other hand, the discussed caching policies are only designed from the perspective of users and the BSs are assumed to be homogeneous in a mobile network. However, it is not always the case in practical scenarios since the cost of connecting different BSs is not the same for a typical user. A remote BS may lead to much higher access costs compared to a near one since it generally needs more bandwidth and power to serve the user. Moreover, the priorities of users may be different from the viewpoint of service providers. For a given user, the priority depends on

It is extremely important to decide which content should be cached in which BS, which BS should be associated with which user, and which BS should be turned off at which time period. Applying a cache policy independent of those factors is not an intelligent approach, and a joint caching deployment, user association and BS operation scheme should be carefully designed to overcome these challenges.

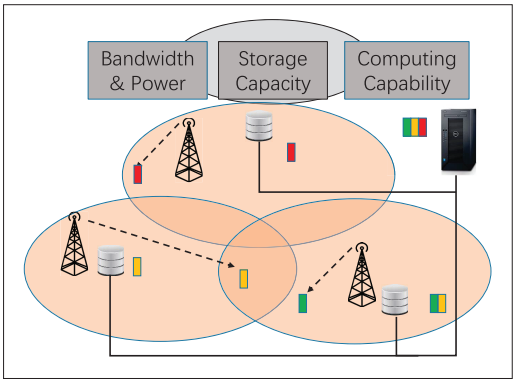

FIGURE 5. Trade-off among radio resource, storage capacity and computing capability in cache-enabled mobile networks.

the location of the user and the density of traffic distribution, the load of BSs and the core network. In particular, mobile networks are becoming highly dynamic. Access nodes and backhauls in mobile networks may be spontaneously created and be randomly switched off, affecting the life span of the cached files significantly. We should not blindly pursue saving the transmission bandwidth cost while ignoring the storage capacity and the computing capability. It is extremely important to decide which content should be cached in which BS, which BS should be associated with which user, and which BS should be turned off at which time period. Applying a cache policy independent of those factors is not an intelligent approach, and a joint caching deployment, user association and BS operation scheme should be carefully designed to overcome these challenges.

Conclusions

The caching technique shows its potential in improving the network performance of 5G and beyond mobile communication systems. In this article, we have made investigations on the promising caching techniques for mobile networks, including feasible caching policies and their individual advantages and weaknesses, and have introduced suitable metrics for performance evaluation of caching techniques from different perspectives such as the QoS of users and the operating expense of mobile networks. Numerical simulations demonstrate that applying in-network caching strategy in mobile networks can reduce latency and access cost significantly. We have also discussed future research directions such as designing a software-defined networking based cache framework for mobile networks and obtaining a trade-off among radio resource, storage capacity and computing capability to further improve system performance.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (61671233, 61801208, 61627804); the Jiangsu Science Foundation (BK20170650); the Postdoctoral Science Foundation of China (BX201700118, 2017M621712); the Jiangsu Postdoctoral Science Foundation (1701118B); and the open research fund of the National Mobile Communications Research Laboratory, Southeast University (2019D02).

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2016–2021," White Paper, Feb. 2017.
[2] F. Jiang *et al.*, "Stair Matrix and Its Applications to Massive MIMO Uplink Data Detection," *IEEE Trans. Commun.*, vol. 66, no. 6, June 2018, pp. 2437–55.
[3] M. Feng, S. Mao, and T. Jiang, "Joint Frame Design, Resource Allocation and User Association for Massive MIMO Heterogeneous Networks with Wireless Backhaul," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, Mar. 2018, pp. 1937–50.
[4] R. Q. Hu and Y. Qian, "An Energy Efficient and Spectrum Efficient Wireless Heterogeneous Network Framework for 5G Systems," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 94–101.
[5] X. Du *et al.*, "A Routing-Driven Elliptic Curve Cryptography Based Key Management Scheme for Heterogeneous Sensor Networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, Mar. 2009, pp. 1223–29.
[6] Q. Shen, Z. Ma, and S. Wang, "Deploying C-RAN in Cellular Radio Networks: An Efficient Way to Meet Future Traffic Demands," *IEEE Trans. Veh. Tech.*, vol. 67, no. 8, Aug. 2018, pp. 7887–91.
[7] J. Dai and S. Wang, "Clustering-Based Spectrum Sharing Strategy for Cognitive Radio Networks," *IEEE JSAC*, vol. 35, no. 1, Jan. 2017, pp. 228–37.
[8] B. Feng *et al.*, "Locator/Identifier Split Networking: A Promising Future Internet Architecture," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 4, Fourth Quarter 2017, pp. 2927–48.
[9] Y. Xiao *et al.*, "Internet Protocol Television (IPTV): The Killer Application for the Next-Generation Internet," *IEEE Commun. Mag.*, vol. 45, no. 11, Nov. 2007, pp. 126–34.
[10] S. Yu *et al.*, "Networking for Big Data: A Survey," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 1, First Quarter 2017, pp. 531–49.
[11] G. Paschos *et al.*, "Wireless Caching: Technical Misconceptions and Business Barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 16–22.
[12] 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.
[13] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, May 2014, pp. 2856–67.
[14] L. Breslau *et al.*, "Web Caching and Zipf-Like Distributions: Evidence and Implications," *Proc. IEEE INFOCOM'99*, vol. 1, Mar. 1999.
[15] B. Blaszczyszyn and A. Giovanidis, "Optimal Geographic Caching in Cellular Networks," *Proc. IEEE ICC'15*, June 2015, pp. 3358–63.

Biographies

Shaowei Wang [S'06, M'07, SM'13] received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. He is currently a full professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. From 2012 to 2013, he was a visiting scholar/professor with Stanford University, Stanford, CA, USA, and The University of British Columbia, Vancouver, BC, Canada. His research interests include telecommunication systems, operations research and machine learning. He organized the Special Issue on Enhancing Spectral Efficiency for LTE-Advanced and Beyond Cellular Networks for *IEEE Wireless Communications*, and the Feature Topic on Energy-Efficient Cognitive Radio Networks for *IEEE Communications Magazine*. He is on the editorial board of *IEEE Communications Magazine*, *IEEE Transactions on Wireless Communications*, and *Springer Journal of Wireless Networks*. He serves/has served on the technical or executive committee of reputable conferences including IEEE INFOCOM, IEEE ICC, IEEE GLOBECOM, and IEEE WCNC, among others.

Tianyu Wang [S'11, M'16] received the Ph.D. degree from Peking University, Beijing, China, in 2011. He is currently an associate researcher with the School of Electronic Science and Engineering, Nanjing University, China. He has published more than 40 IEEE journal and conference papers, and received the Best Paper Award from the IEEE ICC'15, IEEE GLOBECOM'14, and ICST ChinaCom'12. His current research interests include network slicing, load balancing and machine learning in wireless networks. He is a co-first author of this article.

Xun Cao [S'10, M'12] received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2012. He held visiting positions with Philips Research, Aachen, Germany, in 2008, and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a visiting scholar with the University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. He is currently a full professor with the School of Electronic Science and Engineering, Nanjing University. His research interests include computational photography, image-based modeling and rendering, and VR/AR systems.