Caching at the Wireless Edge: Design Aspects, Challenges, and Future Directions

Dong Liu, Binqiang Chen, Chenyang Yang, and Andreas F. Molisch

The authors introduce methods to predict the popularity distributions and user preferences, and the impact of erroneous information. They then discuss the two aspects of caching systems, content placement and delivery. They present the key differences between wired and wireless caching, and outline the differences in the system arising from where the caching takes place.

ABSTRACT

Caching at the wireless edge is a promising way to boost spectral efficiency and reduce energy consumption of wireless systems. These improvements are rooted in the fact that popular contents are reused, asynchronously, by many users. In this article we first introduce methods to predict the popularity distributions and user preferences, and the impact of erroneous information. We then discuss the two aspects of caching systems, content placement and delivery. We expound the key differences between wired and wireless caching, and outline the differences in the system arising from where the caching takes place (e.g., at base stations or on the wireless devices themselves). Special attention is paid to the essential limitations in wireless caching, and possible trade-offs between spectral efficiency, energy efficiency, and cache size.

INTRODUCTION

The main trend of improving spectral efficiency (SE) for fifth generation (5G) cellular networks is enhanced exploitation of spatial resources. By reducing the distance between base stations (BSs) and users, an order of magnitude increase in network throughput can be expected. However, the effectiveness of this approach relies on high-speed backhaul connectivity to every single BS, the capacity of which must exceed the aggregated data rate of all its served users, which often is not practical for the reason of cost.

An alternative approach to improving SE is to reduce the unnecessary traffic load by rethinking the characteristics of the traffic itself and coming back to the most basic questions. Should all the data be conveyed right after they are generated or even requested?

The major driver of the exponential traffic growth in today's wireless networks is mobile Internet, which inherits a large portion of traffic from the wired Internet. The majority of such traffic is content delivery, which is non-real-time in nature. In fact, it has been reported that video on demand (VoD), a major class of content dissemination traffic, will generate more than 69 percent of mobile data traffic by the end of 2019. Different from conventional communications, the content generation of this traffic is usually decoupled from the content delivery in terms of both source-destination and the time instance of generation request.

Another key feature of content delivery traffic is that a few popular contents account for most of the traffic load, and are requested by many users at different times. With predicted content popularity endowed by big data analytics, it is advantageous to cache popular contents *locally* before the requests truly arrive, directly at the wireless edge (e.g., at BSs or even end-user devices) [1, 2]. This is all the more attractive as the trend in rapid growth of storage capacity of devices enables such caching at relatively low cost.

Caching in the wired Internet is well established and has been shown to reduce latency and energy consumption. Similar benefits can be achieved by caching at the wireless edge, which can improve both SE and energy efficiency (EE) by precaching at users or further enabling device-to-device (D2D) communications, and by caching at small BSs (SBSs) to eliminate the backhaul bottleneck, possibly with a higher gain. This is because content is available locally, instead of requiring redundant traverse across backhaul links (not to mention transport across the Internet and mobile core network). However, the limitations of wireless networks need to be taken into account.

While the potential of local caching has been revealed by several recent investigations [1–5], the current article provides an overview of the state-of-the-art challenges and possible solutions of caching at the wireless edge, which differs from [6] that focused on caching at mobile core networks. To identify what is unique in wireless caching, we address the similarity and emphasize the difference between local caching in wired and wireless edge.

For both wired and wireless networks, caching consists of two closely-coupled problems: *content placement* and *content delivery*. The content placement problem includes determining the size and location of each cache, selecting which content from a library to place at which nodes, and how to download the content to these cache nodes. The content delivery problem is how to convey a content to a user that requests it.

Since local caching aims to improve the quality of experience (QoE) of users or improve network performance without compromising QoE,

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61120106002 and National Basic Research Program of China (973 Program) under Grant 2012CB316003. The work of D. Liu was supported by the Academic Excellence Foundation of BUAA for Ph.D. Students. The work of A. F. Molisch was supported by the National Science Foundation.

Dong Liu, Binqiang Chen, and Chenyang Yang are with Beihang University; Andreas F. Molisch is with the University of Southern California.

0163-6804/16/$25.00 © 2016 IEEE

the benefits and solutions of caching highly depend on the traffic characteristics (e.g., the users' demand profiles and quality requirements). Therefore, we first introduce *content popularity* and *user preference*, and discuss the QoE metrics of two representative types of content delivery traffic and the resulting challenge. Then we address the differences in content placement and content delivery in wireless and wired networks. Next, we compare the SE and EE gains from caching at the wireless and wired edge, and illustrate the fundamental trade-offs and design aspects of caching at the wireless edge by simulation for two example systems. Finally, concluding remarks are provided.

Key Features of Content Delivery Traffic

Content Popularity and User Preference

Using big data analytics, the statistical patterns of content requests, both in aggregate form and on a per-user basis, can be predicted and play a key role in the design of caching.

Content Popularity: By popularity we mean the ratio of the number of requests for a particular content to the total number of requests from users, usually obtained for a certain region during a given period of time. It has been reported that the popularity of content follows a Zipf distribution, a sort of power law distribution [7], which can be characterized by the content catalog size N_f and a skewness parameter β. In general, the content popularity distribution changes at a much slower speed than the traffic variation of cellular networks. Consequently, it is usually approximated as constant over a long time (e.g., one week for movies, and two or three hours for news [1]). Another important fact is that global popularity in a large region (say in a city or even a country) is often different from local popularity in a small region (say on a campus) [8].

Popularity prediction has become an active research field recently, as it is beneficial to many applications such as network dimensioning and online marketing. Many prediction methods have been proposed, such as cumulative views statistics based on the popularity correlation over time [7]. A special difficulty in wireless networks is the fine spatial granularity at which content popularity often needs to be known. Specifically, predicting the content popularity in the coverage of a BS is challenging because the number of users associated with a BS is dynamic, and the number of cumulative requests is limited during the lifetime of popular content.

User Preference: The user preference profile comprises the probability that each content is requested by a specific user during a certain period, and differs among individuals. This comes from the fact that a user usually has strong preference toward specific content categories [8].

The preference of a user can be predicted by machine learning (say collaborative filtering) based on the historical content requests of the user and the similarity among users [9]. This has been extensively studied in recommendation systems and is a hot topic nowadays for more general applications.

Prediction Uncertainty: The prediction accuracy of the content popularity and user preference affects the performance of proactive caching. Specifically, erroneous information reduces the probability of finding the requested files in the cache, called *cache-hit probability*, a metric usually used to reflect caching performance. As a result, it reduces the performance gain from caching and introduces extra cost, such as the energy consumed at backhaul and BS for improper content placement.

VoD and File Downloading

Content delivery traffic includes file downloading (e.g., software or data library update, music or video download) and video streaming. For file downloading, the content is consumed after the complete file has been delivered, where download time is often used as a metric to reflect the QoE. For video streaming, the user starts to play the video immediately after sending the request, where low initial delay, requested video quality, and few stalls during playback should be guaranteed.

Dynamic adaptive streaming over HTTP is common to reduce stalling under varying network congestion by providing various quality levels of videos in a content server. Caching at the wireless edge to support video streaming service is more challenging. Multiple versions of a video with different quality levels need to be cached by either storing differently compressed versions of the same video or using difference encoding such that a better quality video can be reconstructed from the lower quality video plus an enhancement layer. Moreover, a mobile user may retrieve only a partial segment of a video from a BS according to the video playback process and the moving speed. With limited cache size, caching each complete video at every BS allows only a smaller number of distinct videos to be stored, while caching partial video may cause the cached part to be useless for catching up with the playback process.

Content Placement: Difference between Caching in Wireless and Wired Edge

In this section, we address the possible benefits, trade-offs, and research issues of content placement at the wireless edge, by highlighting the difference with that at the wired edge, which comes from the architecture of wireless networks, the nature of wireless channels, and user mobility.

Caching at BSs or Users: Benefits and Tradeoffs

In wireless networks, caches can be installed in macro BSs (MBSs), SBSs (say, pico or femto BSs), relays, and users' devices, see Fig. 1.

Caching at BSs: Compared to caching at the evolved packet core (EPC) or a higher level, caching at existing MBSs and SBSs essentially plays the role of replacing backhaul links, and hence alleviates backhaul congestion. Moreover, a new type of SBS without any backhaul connections, called *helpers* [1], enable flexible and cost-effective deployment to deliver popular contents.

Because increasing cache size can increase the cache-hit probability, and hence lower the required backhaul capacity, there is a trade-off

A special difficulty in wireless networks is the fine spatial granularity at which content popularity often needs to be known. Specifically, predicting the content popularity in the coverage of a BS is challenging because the users associated with a BS is dynamic and the number of cumulative requests is limited during the lifetime of popular content.

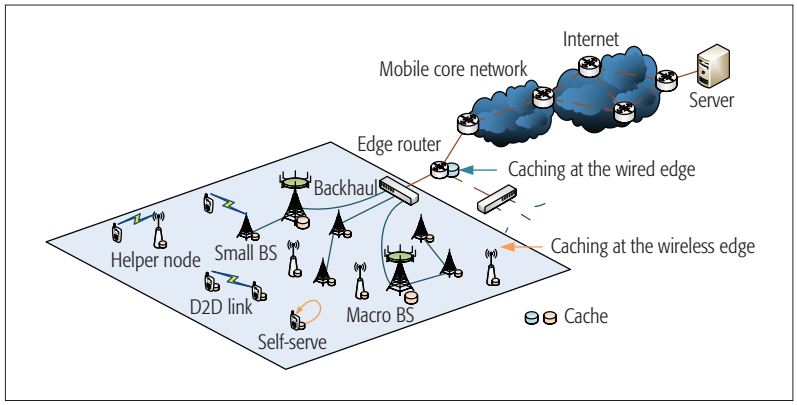

Figure 1. Local caching and content delivery at the wireless edge.

between cache size and backhaul capacity. When the backhaul capacity is a bottleneck, increasing cache size is able to increase throughput. This translates to a trade-off between the cache size and the network SE.

Since caching at the BSs can reduce the traffic in backhaul, mobile core network, and wired Internet, it can also improve the EE of the network if power-efficient cache hardware such as a high-speed solid state disk is used.

As an important difference to wired networks, the performance gain of caching at the wireless edge highly depends on the inter-cell interference (ICI) level. In one example, the maximal EE gain of caching over not caching will be 575 percent if the ICI can be completely removed and 250 percent without any ICI coordination [5]. When further considering overlapped coverage of densely deployed SBSs or heterogeneous networks and user mobility, the performance of caching in the wireless edge are yet to be fully exploited.

Caching at Users: Precaching contents at user terminals such as smartphones, tablets, and laptops has long been applied as a technique to improve QoE [10]. Recently, it is also proposed to offload wireless traffic.

With known content popularity, a BS can push the popular contents to all users via broadcast [11]. With known user preference, the BS can pre-download favorite contents to some users via unicast. According to whether the contents are sent to a user via unicast [10] or broadcast [11], in this article we classify *precaching* into *prefetching* and *pushing*.

When a user requests a content already cached at its local storage, the content can be retrieved from the cache (referred as self-serve as in Fig. 1) with zero delay and without generating interference to other users. Otherwise, the content can be conveyed to the user via unicast. In this way, precaching can improve QoE of all users either directly or indirectly, and improve wireless throughput by offloading.

In practice, a user may not be willing to contribute a large portion of its storage space for caching files. Then pushing according to content popularity in a cell may not yield high cache hit probability, p_h. For instance, when the content popularity is Zipf distributed with $N_f = 1000$ and $\beta = 0.8$, and 10 files can be pushed to a user,

$$p_h = \left(\sum_{i=1}^{10} i^{-0.8}\right) \Big/ \left(\sum_{i=1}^{N_f} i^{-0.8}\right) = 23\%.$$

How to motivate users to cache more contents deserves investigation.

By further exploiting D2D communications, some users can share their cached contents to help improve the QoE of other users in proximity and offload the traffic [1]. If the requested file is not cached at the local storage but cached at a nearby user, a D2D link can be established to deliver the content.

Recent results have demonstrated pronounced performance gain brought by cache-enabled D2D communications. Under many parameter settings, cache-enabled D2D provides network throughput that increases linearly with the number of users (or, equivalently, a per-user throughput independent of the number of users) [4]. This makes this technology one of the most promising methods of achieving order-of-magnitude improvements in SE. Besides, a trade-off between throughput and outage is possible in cache-enabled D2D networks [4].

However, a high offloading ratio comes at the cost of the energy consumed at users, especially those acting as D2D transmitters. This may lead to a trade-off between the offloading ratio and the energy consumption.

Caching Policy in the Wireless Edge: Unique Features

Similar to caching in wired networks, an appropriate caching policy in wireless networks is critical for achieving the potential gain.

A caching policy can be either *reactive* or *proactive*. A reactive caching policy determines whether to cache a particular content after it has been requested according to certain replacement algorithms [8]. A proactive caching policy determines which contents should be cached at each node before they are requested based on the predicted users' demand profile [1, 2, 8]. For proactive caching, the decision to cache among multiple nodes can be jointly optimized, and hence the caching gain is high with perfect prediction. With prediction errors, however, the cache-hit probability will be reduced, and proactive caching may perform worse than reactive caching [12].

Caching at BSs: Because the coverage of a BS is much smaller than an EPC or server, and the connectivity between BS and user is highly uncertain, caching policy design at BSs is more challenging.

Low cache-hit probability: The size of cache and the number of requests at each BS are much smaller than those in the core network or Internet. As a result, the reactive caching policies designed for the Internet are not effective for caching at the BSs [8]. Designing proactive caching policy for each BS independently (e.g., each BS caching the most popular contents) may result in insufficient utilization of caches.

One way to cope with this problem is to enable BSs to share the cached contents through a backhaul link, that is, *cooperative caching* [6, 12]. If the requested content is not in the cache of the local BS, by retrieving the requested content from the caches of adjacent BSs instead of from the server, the delivery cost and latency can

be reduced, and the overall cache-hit probability can be improved. Such an approach is more likely to be viable for MBSs that have high-capacity fiber connection to share data, but is hard to implement for SBSs.

Due to the openness of wireless channels, the coverage of SBSs often overlaps. This indicates that a user is able to fetch contents from multiple caches, and hence the equivalent cache size seen from the user is increased. Based on this observation, caching policies for adjacent BSs can be jointly optimized to increase the cache-hit probability without data sharing over backhaul links, which is referred to as *distributed caching* [1].

Topology uncertainty: Different from wired networks with fixed and known node topology, in wireless networks it will not be known a priori which user will connect to which BS due to undetermined user locations.

One way to deal with this problem is to employ a probabilistic caching policy [13], rather than the deterministic caching policy used in wired networks. To reflect the uncertainty, this approach treats the user locations as a spatial random process, and then optimizes the probability of each content being cached at each BS.

A further complication arises when a user is moving from one cell to another during the duration of content delivery. If the user mobility pattern can be predicted, the caching decision can be optimized [14]. Otherwise, a less than ideal alternative is to model the user movement as a Markov chain with predictable probability transfer matrix.

Fading and interference: In most of the current literature, caching policy optimization fails to take into account channel fading and interference, which are essential to wireless networks. Consequently, the optimized policy may not perform well in practice. For example, distributed caching can increase cache-hit probability, which, however, may not improve SE and EE when path loss and ICI are considered [5].[1] Specifically, when the nearest BS of a user does not cache the requested content of the user but the second nearest BS does, the signal power from the second nearest BS is lower due to path loss. Even worse, the nearest BS is possibly generating strong interference toward the user.

Caching at Users: The content placement at users in wireless networks differs from that in wired networks in both caching policy and the way to download contents. For precaching, the difference comes from fading and interference. For cache-enabled D2D, the difference comes from all mentioned aspects in wireless networks as in caching at the BSs.

Precaching: Traditional prefetching is implemented by over-the-top operators for users who have installed applications (apps) [10]. Since the caching decision is made according to the predicted preference of a user and aims to improve QoE, the contents are naturally predownloaded to each user via unicast when the channel is in good condition. This may generate interference to other ongoing transmission.

When prefetching is implemented by a mobile operator who has knowledge of the congestion status of BSs and the channel conditions of users, the degradation of the network performance can be minimized by designing sophisticated scheduling. Given that the predownloading via unicast consumes energy at both the BS and user, and may cause interference to other users, accurately predicting the contents that a user will request and the time instance when the user will initiate the request are important to guarantee that the advantages exceed the harm brought by prefetching.

Considering that content popularity usually changes slowly, the most popular contents can be pushed by broadcast at off-peak time, which causes negligible or even no performance degradation to the network. For the contents (e.g., news) that update fast, a part of the network bandwidth can be reserved for broadcasting to dynamically push popular files to users [11]; on the downside this may degrade the performance of the network during peak time.

Cache-enabled D2D: Different from prefetching, where the way to select contents for pre-downloading is "selfish," the principle of designing a caching policy for D2D communications is to help other users.

Due to user mobility, devices that will be in range of each other when a transmission request occurs are dynamic, which makes probabilistic caching more appropriate. For a given popularity distribution, the caching distribution that maximizes the offloading ratio takes on a form similar to *water-filling* [4].

With probabilistic caching policy, content can be downloaded to users by a combination of multicast and unicast. To save wireless resources, the BS can transmit content to a user only when first requested, while other users "overhear" and cache content they find suitable; the gain of this approach is still unknown.

With probabilistic caching policy, content can be downloaded to users by a combination of multicast and unicast. To save wireless resources, the BS can transmit content to a user only when first requested, while other users "overhear" and cache content they find suitable; the gain of this approach is still unknown.

Content Delivery: Impact of Caching on Wireless Transmission

To exploit the cache resource and traffic characteristics, some transmit strategies (essentially cross-layer) need to be re-designed. Since caching at BSs can replace backhaul and reduce latency while caching at users can even offload wireless traffic, new criteria emerge for optimization, such as minimizing backhaul traffic or end-to-end delay, and maximizing offloading ratio.

User Association

In heterogeneous networks with overlapped coverage of different BSs, caching may change the method of user association. To reduce end-to-end delay or balance the traffic load in backhaul links, users are not always associated with the nearest BS. Instead, associating the users with the BSs that cache the requested contents may be more beneficial. In this case, the closest BS to the user may generate interference stronger than the desired signal. Such a problem has been identified and studied in the literature.

Coordinated Multipoint Transmission

Coordinated multipoint joint transmission (CoMP-JT) is hard to implement in traditional networks due to the high-capacity backhaul required for exchanging data among BSs. Caching at SBSs makes CoMP-JT a possible cost-ef-

[1] Cache-hit probability can reflect the cache utilization efficiency, but does not necessarily reflect wireless resource utilization efficiency. This is an example of how content placement and content delivery are strongly entwined.

	Caching at BSs	Caching at users	
		Cache-enabled D2D	Precaching
Trade-offs	SE and cache size		
	EE and cache size	Offloading gain and energy cost	/
	Backhaul capacity and cache size	Throughput and outage	
Features in content placement	Small cache size and number of requests		Small cache size
	Uncertainty in topology		/
	Openness of wireless channel: a) overlapped coverage		b) multicast/ broadcast
	Fading and interference		
	/	Limited battery capacity	
Impacts on wireless transmission	Optimization criteria: delay, backhaul traffic, offloading ratio	User incentive to act as helpers	/
	Transmission strategies: user association, CoMP-JT, multicast	Balancing offloading and energy	

Table 1. Trade-offs, features, and impacts of caching in wireless networks.

fective way to alleviate ICI, another limiting factor on wireless transmission capacity of small cell networks. If the requested contents for the users are cached at several adjacent BSs, the BSs can serve the users with CoMP-JT by only exchanging channel information [15]. For helpers that completely lack backhaul, channel information needs to be shared with the assistance of the user device, or non-coherent CoMP-JT can be employed to avoid ICI and enhance the signal (but without any multiplexing gain).

Considering the limited cache size at SBSs and the dynamic nature of users' requests, cache-enabled CoMP-JT can only operate opportunistically. To enjoy the gain from CoMP, the probability that the content desired by a user can be found in local caches of multiple adjacent BSs should be increased. A simple way to do this is caching popular contents in every SBS, which, however, inevitably reduces the overall cache-hit probability. To maximize SE of the cache-enabled CoMP-JT with given cache size, the caching policy needs to be carefully designed. Because not all users in each cell can be jointly served by adjacent SBSs with the requested contents locally cached, the SBSs not using CoMP-JT will generate interference to the CoMP users, which will limit the overall throughput gain.

Multicast

Multicast is a mechanism to exploit content popularity for reducing duplicated transmission, with which a BS can serve multiple users requesting identical contents if they send the requests at the same time as happens for live video events.

The requests of users for content delivery are highly asynchronous. For file downloading that is delay-tolerant, this is not a big issue, where traditional multicast technology can be used to let the users wait for each other before starting transmission. For VoD streaming that has the particular QoE provision, this results in a large initial delay, which degrades the user perceived QoE. Although schemes such as harmonic broadcasting partly overcome this problem, they do not provide maximal SE.

This dilemma can be solved by jointly optimizing the content placement and content delivery. In [3], an ingenious coded-multicast strategy was proposed, which precaches partial contents possibly requested by all the users with carefully designed network coding. Then, during the content delivery phase, different requests can be satisfied with a single multicast transmission. Despite the theoretical beauty and importance of this strategy, practical challenges remain, since the coding complexity grows exponentially and the download delay for each user increases with the number of users.

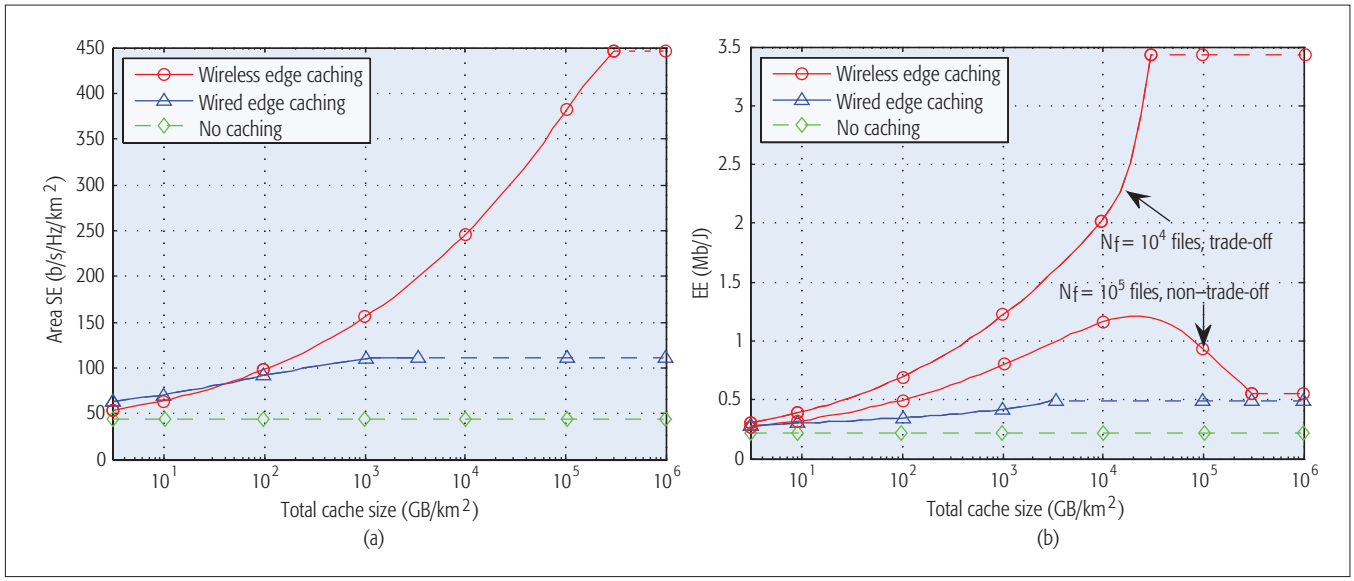

Figure 2. SE and EE comparison between caching at the wireless and wired edges. The solid curves end when all the files have been cached at each cache node.

Device-to-Device Communications

Traditional D2D communications can only offload peer-to-peer (P2P) traffic between source and destination in proximity at the time they wish to communicate. Cache-enabled D2D creates new opportunities to offload client/server traffic [1].

Nonetheless, different from wired P2P networks, in D2D transmission there is not only a question of which users in the vicinity are willing to help, but also the willingness of the users might change with time, and specifically with the state of their battery. Consequently, incentivizing users to act as helpers is an important issue. To justify the feasibility of cache-enabled D2D communications, the energy consumed by D2D transmitters needs to be evaluated, and the caching policy needs to be jointly optimized with communication protocol to balance the throughput gain and the energy cost. For readers' convenience, the key points of earlier sections are summarized in Table 1.

Simulation Results

To illustrate the design aspects in content placement and delivery, the performance difference between caching in wired and wireless edges, and quantitative results for some trade-offs, we consider two representative systems. The simulation parameters are summarized in Table 2.

To compare the SE and EE gains of caching at wireless and wired edges over not caching, and show the SE/EE-cache size trade-off, we consider a small cell network. Caches are either deployed at SBSs or the edge router, as shown in Fig. 1. Each cache node caches the most popular files until it is full.

It is shown from Fig. 2 that both SE and EE benefit more from caching at the wireless edge than at the wired edge. Specifically, the maximal SE gains are about 200 and 900 percent for wired and wireless edge caching, and the maximal EE gain are 200 percent and 500 percent, respectively. This is because compared to caching at the wired edge, which can only relieve congestion in the mobile core network and the Internet, caching at BSs can also relieve backhaul congestion, although the cache hit probability is lower. As the total cache size increases, SE increases for both wireless edge and wired edge caching. In both cases SE saturates, although the saturation point occurs at higher total cache size for wireless edge caching, since popular files need to be stored at multiple locations. EE generally also increases with total cache size, although EE first increases and then decreases for the case of very large library size (and thus large number of unpopular files), since the caches consume circuit power, and caching rarely used files requires more energy than it saves [5].

To illustrate the design aspects except the cache size, we consider a cache-enabled D2D network, where each user sends N_r requests sequentially in a period, and only the users within a collaboration distance r_c can establish D2D links. We also show the trade-off between *offloading ratio*, the amount of data conveyed by D2D divided by the total amount of data transmitted in the cell, and *energy consumption*, which is the total transmit and circuit power consumed at each D2D transmitter averaged over channel fading, user location, and file request. We consider *optimal caching* proposed in [4], *uniform caching* where each user caches files with identical probability, and *popularity caching* where each user caches a file with probability equal to the file popularity. We consider maximal transmit power ($P_{\max}$) and optimal transmit power (P^*) at each D2D transmitter, where P^* is obtained numerically by maximizing the offloading ratio.

It is shown from Fig. 3 that optimizing the caching policy can improve the offloading ratio and reduce energy consumption, while optimizing the collaboration distance r_c and transmit power are more critical. Furthermore, there is

Systems	Caching in SBSs	Cache-enabled D2D
File catalog size, N_f	10^5	10^3
File size	30 MB	
Zipf distribution skewness parameter, β	0.8	
Wireless transmission bandwidth	20 MHz (without interference coordination)	
Transmit power of SBS or user	200 mW	
Other power consumption parameters	See [5] and its reference [9]	Circuit power, 100 mW
Considered region	1 km × 1 km	0.5 km × 0.5 km
Number of SBSs (each with four antennas)	100 (uniform-located)	/
Number of single-antenna users	300 (uniform-located)	2500 (uniform-located)
Capacity of backhaul	30 Mb/s (microwave)	/
Transport bandwidth of wired network	1 Gb/s	/
Maximal energy allowed for transmitting one file	/	10% fraction of user battery capacity (1800 mAh)

Table 2. Simulation parameters.

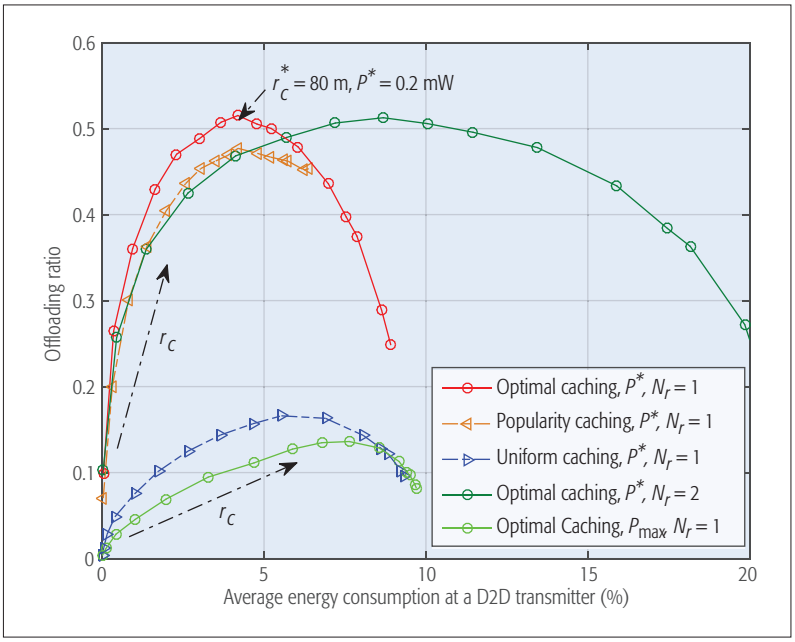

Figure 3. Offloading ratio and energy consumption. Each curve starts from r_c = 10 m and ends at r_c = 400 m, and r_c^* is the optimal collaboration distance that maximizes the offloading ratio.

To exploit the full potential of wireless edge caching, the unique limitations in wireless networks due to architecture and channel, such as topology, interference, users' mobility and limited battery, must be considered for both content placement and delivery, and accurate predictions of popularity distributions and user preferences are critical.

a trade-off between the offloading ratio and the energy consumption when r_c is small, as mentioned earlier.

Concluding Remarks

Caching at the wireless edge can significantly improve SE and EE of wireless networks compared to caching at the wired edge. The gain comes from saving bandwidth and energy for both getting the files from the servers to the wireless infrastructure as well as the transmission from wireless infrastructure to users. To exploit the full potential of wireless edge caching, the unique limitations in wireless networks due to architecture and channel, such as topology, interference, users' mobility, and limited battery, must be considered for both content placement and delivery, and accurate predictions of popularity distributions and user preferences are critical.

References

[1] N. Golrezaei *et al.*, "Femtocaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, Apr. 2013, pp. 142–49.

[2] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 82–89.

[3] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Trans. Info. Theory*, vol. 60, no. 5, May 2014, pp. 2856–67.

[4] M. Ji, G. Caire, and A. Molisch, "Wireless Device-to-Device Caching Networks: Basic Principles and System Performance," *IEEE JSAC*, vol. 34, no. 1, Jan. 2016, pp. 176–89.

[5] D. Liu and C. Yang, "Energy Efficiency of Downlink Networks with Caching at Base Stations," *IEEE JSAC*, vol. 34, no. 4, Apr. 2016, pp. 907–22.

[6] 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.

[7] A. Tatar *et al.*, "A Survey on Predicting the Popularity of Web Content," *Springer J. Internet Services and Applications*, vol. 5, no. 1, 2014, pp. 1–20.

[8] H. Ahlehagh and S. Dey, "Video-Aware Scheduling and Caching in the Radio Access Network," *IEEE Trans. Net.*, vol. 22, no. 5, Oct. 2014, pp. 1444–62.

[9] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges," *ACM Comp. Surveys*, vol. 47, no. 1, May 2014, pp. 3:1–3:45.

[10] B. D. Higgins *et al.*, "Informed Mobile Prefetching," *ACM MobiSys*, 2012.

[11] K. Wang, Z. Chen, and H. Liu, "Push-Based Wireless Converged Networks for Massive Multimedia Content Delivery," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2894–2905.

[12] A. Gharaibeh *et al.*, "A Provably Efficient Online Collaborative Caching Algorithm for Multicell-Coordinated Systems," *IEEE Trans. Mobile Computing*.

[13] B. Blaszczyszyn and A. Giovanidis, "Optimal Geographic Caching in Cellular Networks," *IEEE ICC*, 2015.

[14] K. Poularakis and L. Tassiulas, "Exploiting User Mobility for Wireless Content Delivery," *IEEE ISIT*, 2013.

[15] A. Liu and V. K. Lau, "Mixed-Time Scale Precoding and Cache Control in cached MIMO Interference Network," *IEEE Trans. Signal Processing*, vol. 61, no. 24, Dec. 2013, pp. 6320–32.

Biographies

Dong Liu [S'13] (dliu@buaa.edu.cn) received his B.S. degree in electronics engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), China, in 2013. He is currently pursuing his Ph.D. degree in signal and information processing with the School of Electronics and Information Engineering, Beihang University. His research interests lie in the area of caching in wireless network and cooperative communications.

Binqiang Chen [S'14] (chenbq@buaa.edu.cn) received his B.S. degree in electronics engineering in 2012 and now is pursuing his Ph.D. degree in signal and information processing, both from the School of Electronics and Information Engineering, Beihang University. His research interests include interference management, cooperative communications, device-to-device communications, and content-centric networks.

Chenyang Yang [M'99, SM'08] (cyyang@buaa.edu.cn) received her Ph.D. degree in electrical engineering from Beihang University in 1997. She has been a full professor with the School of Electronics and Information Engineering, Beihang University, since 1999. Her research interests include green radio, local caching, and other emerging techniques for next generation wireless networks. She was the Chair of the IEEE Communications Society Beijing chapter from 2008 to 2012. She has served as a Technical Program Committee member for numerous IEEE conferences. She has been an Associate Editor or a Guest Editor of several IEEE journals. She was nominated as an Outstanding Young Professor of Beijing in 1995 and was supported by the 1st Teaching and Research Award Program for Outstanding Young Teachers of Higher Education Institutions by the Ministry of Education of China from 1999 to 2004.

Andreas F. Molisch [S'89, M'95, SM'00, F'05] (molisch@usc.edu) is a professor of electrical engineering at the University of Southern California. His current research interests are the measurement and modeling of mobile radio channels, ultrawideband communications and localization, cooperative communications, multiple-input-multiple-output systems, wireless systems for healthcare, and novel cellular architectures. He is a Fellow of NAI, Fellow of AAAS, Fellow of IET, and member of the Austrian Academy of Sciences, as well as a recipient of numerous awards.