Wireless Caching: Technical Misconceptions and Business Barriers

Georgios Paschos, Ejder Baştuğ, Ingmar Land, Giuseppe Caire, and Mérouane Debbah

The authors discuss several technical misconceptions with the aim of uncovering enabling research directions for caching in wireless systems. Ultimately, they make a speculative stakeholder analysis for wireless caching in 5G.

This research has been partly supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering) and the project BESTCOM.

ABSTRACT

Caching is a hot research topic and poised to develop into a key technology for the upcoming 5G wireless networks. However, the successful implementation of caching techniques crucially depends on joint research developments in different scientific domains such as networking, information theory, machine learning, and wireless communications. Moreover, there are business barriers related to the complex interactions between the involved stakeholders: users, cellular operators, and Internet content providers. In this article we discuss several technical misconceptions with the aim of uncovering enabling research directions for caching in wireless systems. Ultimately, we make a speculative stakeholder analysis for wireless caching in 5G.

INTRODUCTION

Caching is a mature idea from the domains of web caching, content delivery networks, and memory optimization in operating systems. Why is caching still an active topic of discussion? In the 1990s, the traffic in the web exploded, leading its inventor, Sir Tim Berners-Lee, to declare network congestion as one of the main challenges for the Internet of the future. The congestion was caused by the *dotcom boom*, specifically due to the client-server model of connectivity, whereby a web page was downloaded from the same network server by every Internet user in the world. The challenge was ultimately resolved by the invention of content delivery networks (CDNs) and the exploitation of web caching. The latter replicates popular content in many geographical areas and saves bandwidth by avoiding unnecessary multihop retransmissions. As a byproduct, it also decreases access time (latency) by decreasing the distance between two communicating entities.

Today, 30 years later, we are reviving the same challenge in the wireless domain. The latest report of Cisco [1] predicts a massive increase of Internet devices connected through wireless access, and warns of a steep increase in mobile traffic, which is expected to reach roughly 60 percent of total network traffic by 2018, the majority of which will be video. Wireless system designers strive to fortify fifth generation (5G) wireless networks with higher access rates on one hand and increased densification of network infrastructure on the other. Over the last three decades, these two approaches have been responsible for the majority of network capacity upgrade per unit area, successfully absorbing the wireless traffic growth. However, with the explosion of access rates and number of base stations, the backhaul of wireless networks will also become congested [2, 3], which motivates further use of caching: storing popular reusable information at base stations to reduce the load at the backhaul. Furthermore, a recent technique [4] combined caching with coding and revolutionized how goodput scales in bandwidth-limited networks. Therefore, caching has the potential to become the third key technology for wireless system sustainability.

The research community is converging to an enabling architecture as shown in Fig. 1. In the network of the future, memory units can be installed in gateway routers between the wireless network and the Internet (e.g., in 4G this is called the serving gateway, S-GW), in base stations of different sizes (small or regular size cells), and in end-user devices (mobile phones, laptops, routers, etc.). In this article, we discuss important topics such as:

- The characteristics of cacheable content and how this affects caching technologies in wireless
- Where to install memory
- The differences between wireless caching and legacy caching techniques

Last, we focus on business barriers that must be overcome for the successful adoption of wireless caching by the industry.

DEALING WITH MASSIVE CONTENT

Not all network traffic is cacheable. Interactive applications, gaming, voice calls, and remote control signals are examples of information objects that are not reusable and hence cannot be cached. Nevertheless, most network traffic today (an estimated 60 percent [1]) is deemed cacheable. We refer to cacheable information objects as *content* in this article. Since the performance of caching is inherently connected to the specifics of contents, this section is dedicated to the understanding of these specifics.

In particular, we focus on the following misconceptions:

- The static IRM model is sufficient for experimentation.

Georgios Paschos, Ingmar Land, Mérouane Debbah are with Huawei Technologies; Ejder Baştuğ is with Université Paris-Saclay; Giuseppe Caire is with Technische Universität Berlin

0163-6804/16/$25.00 © 2016 IEEE

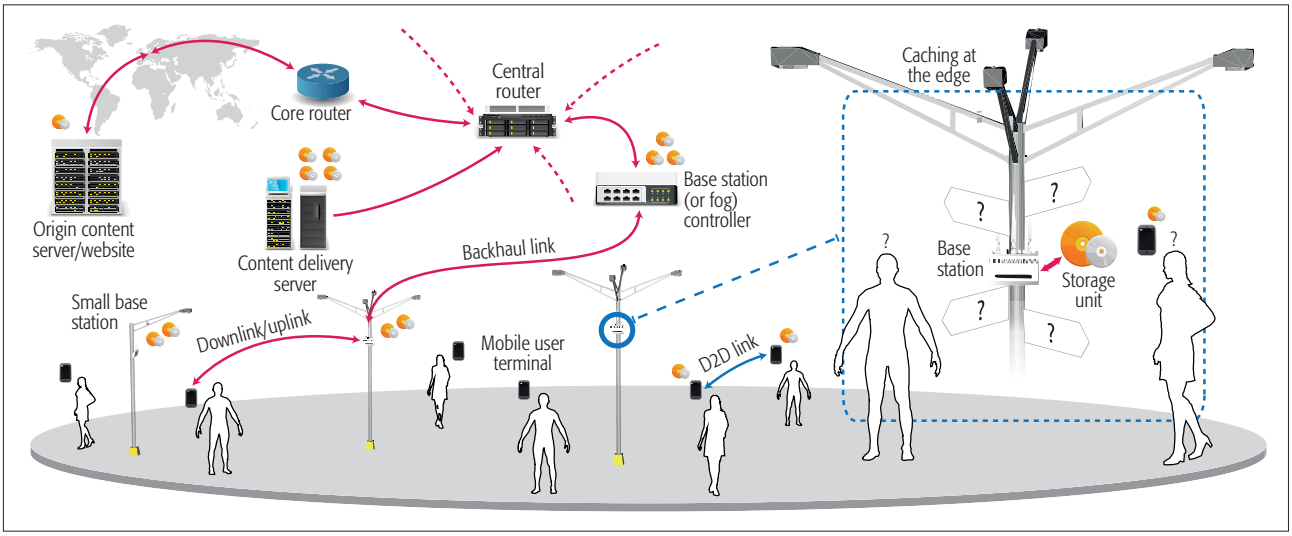

Figure 1. An illustration of caching in future wireless networks. Contents available in the origin server are cached at the base stations and user devices for offloading the backhaul and the wireless links.

- User information cannot be used for popularity estimation due to the vast number of users.
- Security issues precludes caching at the edge.

Insufficiency of Static Popularity Models

The standard approach to designing and analyzing caching systems involves a model for generating content requests to replace the actual request traces — this approach is often several orders of magnitude faster.

The de facto model for performance analysis of web caching is the independence reference model (IRM): content N is requested according to an independent Poisson process with rate λp_n, where p_n refers to the content popularity modeled by a power law (i.e., $p_n \propto n^{-\alpha}$, $\alpha > 0$). This well established model thrives due to its simplicity; it only has two parameters: λ to control the rate of requests, and α to control the skewness of the popularity.

Numerous studies fit the IRM to real traffic with satisfactory results [5], so why do we need to change it? The IRM assumes that the content popularity is static, which of course is not true. Trending tweets, breaking news, and the next episode of *Game of Thrones* are examples of ephemeral content with rapidly changing popularity; they appear, they become increasingly popular, and they gradually become unpopular again. In fact, [6] considers large YouTube and video on demand (VoD) datasets and discovers that time-varying models are more accurate than the IRM with respect to caching performance analysis; Fig. 2 reproduces the comparison when fitting YouTube data and shows the superiority of modeling the popularity as time-varying. In the inhomogeneous Poisson model proposed in [6], each content is associated with a "pulse" the duration of which reflects the content life span and the height of which denotes its instantaneous popularity. The model is called the shot noise model (SNM), mirroring the Poisson noise from electronics. While the shape of the pulse is not important, the study observes strong correlations between popularity and duration; apparently, popular contents prosper longer. Finally, a class-based model [6] can conveniently capture spatio-temporal correlations while allowing analytical tractability. Mobile users are especially keen on downloading ephemeral content; thus, it is expected that in the case of wireless content, the improvement in modeling accuracy will be even greater.

To optimize a cache one needs to track the changes in content popularity. For example, the classical web caching systems adopt dynamic eviction policies like least recently used (LRU) in order to combat time-varying content popularity in a heuristic manner. However, the joint consideration of popularity variations with wireless systems reveals a new challenge that renders LRU policies inefficient. While a typical CDN cache normally receives 50 requests/content/day, the corresponding figure for base station cache may be as low as 0.1 requests/content/day. With such a small number of requests, fast variations of popularity become very difficult to track, and classical LRU schemes fail.

This development motivates novel caching techniques that employ learning methodologies to accurately track the evolution of content popularity over time. A recent study [7] analyzes the SNM model and gives the optimal[1] policy for joint caching and popularity estimation. Additionally, [8] proposes as an alternative solution the use of LRU with prefilters.

How to Track Popularity Variations

Since content popularity is time-varying, caching operations can only be optimized if a fresh view of the system is maintained. This requires massive data collection and processing, and statistical inference from this data, which by itself is a complex task to handle. Additionally, user privacy is a concern that can limit the potential of collecting such information. So, can we promptly gather all this information in a wireless network?

Consider K users subscribed to telecom operator and L caches placed in the network (e.g., in

[1] The optimality is established for the restricted case of homogeneous rectangular pulses and asymptotically large content catalogs.

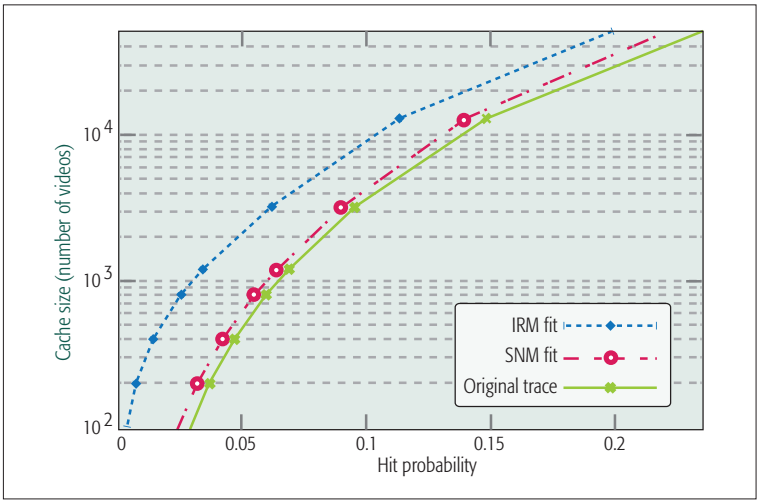

Figure 2. Hit probability comparison between best fit of the IRM, SNM, and YouTube traces (from [6]).

base stations or other locations). Each of these caches has the capability of storing M contents out of N contents in the catalog. Then let the matrix $\mathbf{P} \in \mathbb{R}^{K \times N}$ model the content access statistics where rows are users and columns are contents. In other words, each entry (or rating) in this matrix quantifies how popular content N is to user K. The popularity matrix $\mathbf{P}$ is large, sparse, and only partially known in practice, and has to be continuously estimated in order to enable the correct cache decisions at the base stations. At first, this seems to be an impossible feat.

To deal with the complexity of handling matrix $\mathbf{P}$, it is possible to use machine learning tools to estimate the unknown entries. Such an estimation is particularly efficient when the matrix has low spectral dimension, and the system can be described by a small number of "features"; fortunately, popularity correlation induces such behavior in matrices obtained from real data. For instance, low-rank matrix factorization methods, that is, $\mathbf{P} \approx \mathbf{K}^T\mathbf{N}$ where $\mathbf{K} \in \mathbb{R}^{r \times K}$ and $\mathbf{N} \in \mathbb{R}^{r \times N}$ are factor matrices, can be employed to construct the r-rank version of the matrix, using the fact that users' interests are correlated and predictable when r is small. This additionally allows the collected statistics to be stored in a more compact way. As a result, a big data platform installed in the operator network can provide efficient collection and processing of user access patterns from several locations, as evidenced in [9]. Further development of novel machine learning tools, such as clustering techniques, are needed to improve the estimation of the time-evolving content popularity matrix (i.e., $\mathbf{P}_l(t)$ for base station L), which may differ from base station to base station.

It is worth noting that a caching system has requirements similar to those of a recommendation system. For example, the well-known Netflix movie recommendation system exploits information of a user's past activity in order to predict which movie is likely to be scored high by the user. Similarly, a caching system exploits the request sequence to predict what contents are popular enough to be cached. In this context, user privacy regulations may affect the collection of these valuable data. A key topic of research in this direction is privacy-preserving mechanisms that can enable sufficient sampling of the time-evolving and location-dependent popularity matrix $\mathbf{P}_l(t)$ without compromising user privacy.

Security Is A Kind of Death

A common anti-caching argument relates to the operation of caching in a secure environment. The secure counterpart of HTTP, called HTTPS, was originally used to provide end-to-end (e2e) encryption for securing sensitive information like online banking transactions and authentication. Due to the recent adoption from traffic giants Netflix and YouTube, the HTTPS protocol is growing to soon exceed 50 percent of total network traffic. Content encryption poses an unsurmountable obstacle to in-network operations, including caching. Since encrypting the data makes them unique and not reusable, caching, or even statistically processing encrypted content, is impossible. Ironically, Tennessee Williams' statement "security is a kind of death" seems to squarely apply to wireless caching.

Security is definitely a precious good everyone welcomes. Although securing a video stream might seem an excessive measure, in some cases it may be well justified. Unfortunately, e2e encryption is clearly not in the Berners-Lee spirit since it prevents operators from optimizing their networks and reanimates the server-client ghost of congestion, a reality that equally no one can overlook. In fact, modern CDN systems resolve this issue by having "representatives" of the content provider at the edge of the Internet. These representatives are trusted entities that hold the user keys and are able to decrypt the requests and perform standard caching operations. Ultimately, this methodology is neither entirely secure for the user nor efficient for the network [10]. The need to make the system sustainable finally overrules the need for e2e encryption, which is an argument against HTTPS for video delivery. Given this situation, however, how can we realistically push caching deeper into the wireless access?

Currently, content providers install their own caching boxes in the operator network and intercept the related encrypted content requests deeper in the wireless access network. In this approach, the boxes are not controlled by the operator, which leads to several limitations:

- The caching boxes cannot perform complex tasks.
- It is difficult to apply learning techniques without context information from the operator.
- The caching approach is similar to CDNs and therefore does not exploit the performance opportunities specific to wireless caching, as we discuss below.

New security protocols have been proposed to enable operators to perform caching on encrypted requests [10]. This leads to an interesting research direction: to combine user security and privacy with facilitation of the network management operations, which are crucial for the sustainability of future wireless systems.

Toward a Unified Network Memory

The proposition of information-centric networking (ICN) as a candidate for the future Internet has also raised the subject of *where to install net-*

work memory [11]. The ICN approach proposes to equip routers with caches, and to allow content replication everywhere in the network. A recent work [12] came up with striking conclusions about the ICN approach: most of the caching benefits of ICN can be obtained by caching at the edges of the network using existing CDNs, and any extra caching in the core network brings only negligible improvements at very high costs. In the wireless domain, however, the question remains relevant: does it make sense to cache even closer to the user than CDN?

It is commonly believed that caching is very inefficient near the user, and thus should be done at CDN. Below we explain the main causes of inefficiency and argue that they can be overcome.

Caching Deeper than CDN

Mitigating backhaul and wireless link overload requires going beyond CDN and caching at the base stations and mobile users. However, the efficient operation of such caches is very challenging.

In particular, there are two main challenges:

a. Caches used in wireless networks are typically small compared to CDN caches.
b. The popularity profile of traffic is highly unpredictable when non-aggregated.

To understand point a, consider that the effectiveness of a cache is measured with the *hit probability* (i.e., the fraction of requests found in the cache). This can be upper bounded by the popularity sum $\sum_{n=1}^{M} p_n$, where p_n is the ordered popularity distribution with p_1 denoting the probability of requesting the most popular file.

For power-law popularity the sum can further be approximated by $(M/N)^{1-\alpha}$, where $\alpha < 1$ is the power-law exponent. A very small ratio M/N means that the hit probability becomes vanishingly small. For example, if we are caching Netflix (12.5 PB) in a mobile phone (10 GB), $M/N \sim 10^{-6}$, $\alpha \sim 0.8$, and the hit probability is less than 10 percent. However, base stations equipped with a disk array (40 TB) can be extremely effective when caching contents for a mobile video on demand (VoD) application. In this context there are three promising research directions:

- Restrict caching operations to a subset of the catalog while maintaining network neutrality.
- Store only parts of the content using partial caching techniques.
- Install massive memory at the edge in the form of small-sized data centers.

The third option will be realized by the *fog computing* paradigm. Table 1 provides some indicative numbers for the memory types available and the catalog sizes of reasonable applications.

To understand the unpredictable nature of sparse requests (formulated as challenge b above), consider as an example the delivery of breaking e-news in a city served by a single CDN node. Most users will download the news only once. The CDN system can quickly detect the rising popularity of the news, since it will receive many requests in a short timeframe. From the point of view of a mobile user, however, the detection of the popularity of the trending news becomes very difficult because the news is requested only once by a given user. This example shows that detection efficiency depends on the number of requests aggregated at the popularity learner. To illustrate this, Fig. 3 shows the optimal hit probability in a hierarchy of L base stations. Learning at the global CDN cache is shown to detect variations that are L times faster than those at local caches. To remedy the situation, it is possible to use an architecture that combines information obtained at different aggregation layers [7].

		Netflix catalog (12.5 PB)[1]	Torrents (1.5 PB)	Wireless VoD catalog (1 TB)
Disk	2 TB	~0.01%	~0.1%	100%
Disk array	40 TB	~0.3%	~2%	100%
Data center	150 PB	50%	100%	100%

[1] The entire catalog was anecdotally measured to contain 3.14 PB of content in 2013, which, however, we multiply by 4 since the same video is available in multiple formats.

Table 1. Typical data size values for normalized cache size M/N taken from the study of [8]. In practice, it is anticipated that wireless traffic is an 80–20 mix of torrent-like traffic and live VoD traffic tailored to wireless device capabilities.

Memory Is Cheap but Not Free

Although the cost of a small cache is dwarfed by the base station cost, the total amount of installed memory in a mobile network can be considerable; therefore, deciding to install wireless caching requires a careful cost analysis [8]. To compute the optimal size of memory to install at each location, one needs to know:

a. The cost coefficients
b. The skewness of content popularity
c. The local traffic distribution in cells

Predicting how a and b will evolve is quite challenging, but as in [8] a survey may determine a good set of parameters at any given time.

For c, the literature is extensively based on grid models, which in the case of future wireless networks might be off for a significant factor. More accurate models have recently been introduced from the field of stochastic geometry, where the cache-enabled base stations are distributed according to a spatial point process (often chosen to be 2D Poisson), thus enabling the problem to be handled analytically. The validity of such modeling compared to regular cellular models has been verified using extensive simulations. Additional insights for the deployment of cache-enabled base stations can be obtained by analytically characterizing the performance metrics, such as the outage probability and average delivery rate, for a given set of parameters such as given number of base stations, storage size, skewness of the distribution, transmit power, and target signal-to-interference-plus-noise ratio (SINR) [13]. Therefore, although storage units become increasingly cheaper, the question of how much storage we should place at each location should be studied together with realistic topological models.

Wireless ≠ Wired

Web caching has traditionally been studied by the networking community. A common misconception says that caching is a network layer tech-

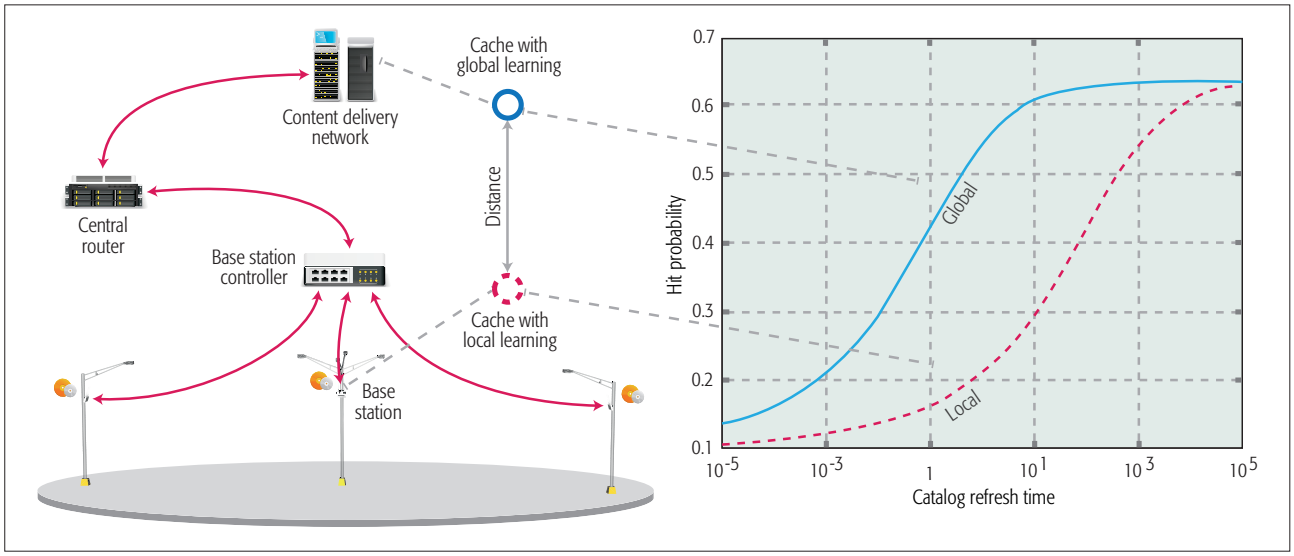

Figure 3. Optimal hit probability comparison between observing the aggregate request process at the CDN level (global) and observing the individual request process at each base station cache (local), when refreshing the catalog. The hit probability performance depends on how fast the time-varying popularities can be learned: global is faster than local.

nique, and hence the web caching approaches are sufficient for wireless caching as well.

However, following the fundamental work of Maddah-Ali and Niesen [4], the idea of caching has penetrated the information theory community with a new twist called *coded caching*, which promises unprecedented gains. In the following, we discuss the differences between wired and wireless caching.

Wireless Caching Lies at Both the Network and PHY Layers

Suppose that a base station wants to deliver information to K users at a rate of 1 Mb/s each for streaming a video. If the video is the same for all users (broadcast video), this might be possible for an arbitrarily large number of users. For example, the base station could use an omnidirectional antenna, exploit the broadcast characteristic of the wireless medium, and transmit at 1 Mb/s to all users simultaneously. When the videos are different, this is clearly not possible: the base station needs to multiplex the users over frequency, time, or codes, where each such *resource block* is then associated with a single user. Since the resource blocks are finite, ultimately the base station can serve 1 Mb/s videos up to a maximum number of users K_{max}, after which the resources are exhausted. To increase K_{max}, physical layer researchers propose ways to increase the resource blocks of a given spectrum (i.e., increase spectral efficiency) or install more base stations so that there are more resource blocks per unit area, referred to as network densification.

The novel paradigm of [4] shows a surprising fact: exploiting caching in a smart way, an unbounded number of users watching different videos can be accomodated. How is this made possible? During off-peak operation of the network, users can cheaply populate their caches with *parts of popular contents*. This is a perfectly reasonable assumption since the question of sustainability that caching is trying to tackle refers to the hours of the day when the network experiences peak traffic. The content parts are appropriately chosen according to a caching code, which ensures symmetric properties. Then at request time, a coding technique called index coding is employed to minimize the number of transmissions to satisfy all users.[2] The combination of these schemes is shown in [4] to yield required resource blocks equal to $K\,(1 - M/N)/(1 + KM/N)$, where K is the number of users, M the cache size, and N the catalog size. Hence, if the cacheable fraction of the catalog M/N is kept fixed, the required number of resource blocks does not increase with the number of users K; this can be verified by taking the limit $K \to \infty$ whereby the above quantity converges to a constant. The result is summarized in Fig. 4.

More recently, it has been shown that in order to achieve such "order of K" gain over conventional unicast (with possible legacy uncoded caching) systems, the content objects must be split into an $O(\exp(K))$ number of subpackets; for networks of practical size, this gain is not achievable. The optimal trade-off between coded caching gain and content object size is a very interesting topic of current research.

From the implementation point of view, promising research directions include extensions to capture system aspects such as:

- Popularity skewness
- Asynchronous requests
- Content objects of finite size
- Cache sizes that scale slower than N

Assuming that these practical challenges are resolved, caching for wireless systems will become intertwined with physical layer techniques employed at the base station and handheld devices.

One Cache Analysis Is Not Sufficient

A contemporary mobile receives the signals of more than 10 base stations simultaneously. In future densified cellular networks, the mobile will be connected to several femto-, pico-, or nanocells. The phenomenon of wireless multi-access opens a new horizon in caching exploitation [14].

[2] In fact, finding the optimal index code is a very difficult problem, and hence the proposed approach resorts to efficient heuristics..

Since a user can retrieve the requested content from many network endpoints, neighboring caches should cooperate and avoid storing the same objects multiple times.

Content placement optimizations of wireless caching typically boil down to a set cover problem in a bipartite graph connecting the users to the reachable caches. Therefore, finding what contents to store at each cache is a difficult problem even if the popularities are assumed known [14]. It is possible to relax the problem to convex optimization by the use of distributed storage codes, where each cache stores coded combinations of contents [14], or obtaining a fractional placement by time sharing different integral placements. These ideas lead to several interesting algorithms in the literature of cooperative caching.

What is the gain from these approaches? Cooperative caching typically saves space in the cache by avoiding caching the same popular contents in neighboring caches. Equivalently, we may think of multiplying the cache size M by a small number, at best say a gain of 3–5. With respect to hit probability, this can correspond to very different levels of gain, depending on the value of M/N. Due to the skewness of the popularity distribution, marginal hit probability gain[3] is high when M/N is small, and very small when M/N is large. Since in wireless we expect the former, high gains are expected from cooperative wireless caching.

The current proposals on cooperative caching assume static popularity, and therefore a promising direction of research along these lines is to design caching schemes that combine cooperation with learning of the time-varying popularity. The time to search and retrieve the content from a nearby cache may also be significant; hence, intelligent hash-based filtering and routing schemes are required [15].

A Stakeholder Analysis for Wireless Caching

The business of wireless caching involves three key stakeholders that together form a complex ecosystem.

The users of telecommunication services are primarily the customers and consumers of the content, but in the case of wireless caching they are also active stakeholders.

Users might be requested to help in the form of contributing with their own resource (e.g., in the case of coded caching it will be memory and processing, or in device-to-device, D2D, caching it will also be relaying transmissions), and they will end up spending energy for the benefit of better performance. On the other hand, one could envision users employing D2D technology to enable caching without the participation of other stakeholders. Due to the complexities mentioned above, however, efficient wireless caching will require heavy coordination and extensive monitoring/processing. Hence, D2D approaches will be limited to restricted environments.

The operators of telecommunication networks are well placed for wireless caching. Due to the particularities of coded caching and multi-access caching, operators are in a unique position to implement new protocols in base stations, affect the standards for new mobile devices, and develop big data processing infrastructure that can realize wireless caching. Nevertheless, for reasons related to encryption, privacy, and global popularity estimation, operators might not be able to install these technologies without the cooperation of the other two stakeholders.

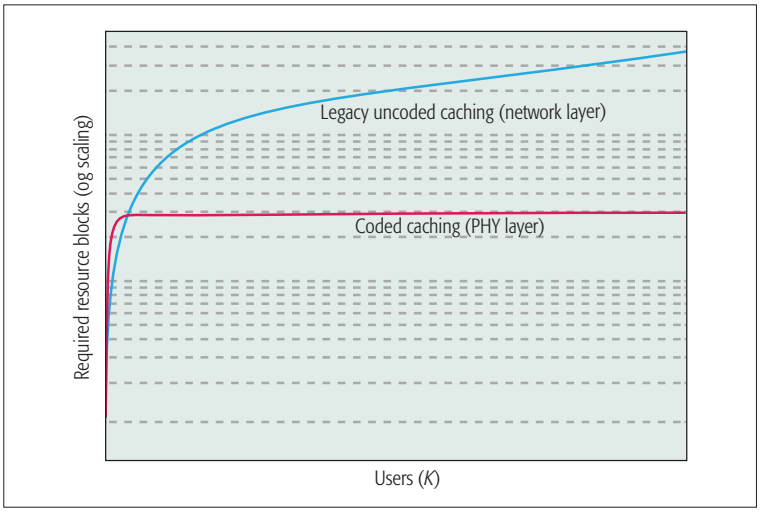

Figure 4. Required resource blocks for K mobile users with unicast demands, when the caches fit 30 percent of the catalog. Coded caching can serve an arbitrarily large population of users with a fixed number of resource blocks.

The providers of Internet content are champions of trust from the user community. Apart from the security keys, they also hold extensive expertise in implementing caching techniques in core networks. From this advantageous position, they can positively affect the progressive evolution of caching in wireless networks. On the other hand, content-provider-only solutions cannot unleash the full potential of wireless caching, since they are limited to alienated boxes in the operator network that can perform caching only with legacy CDN techniques. The deeper the caches go into the wireless network, the less efficient they will be if they stick to legacy CDN techniques.

We summarize what each stakeholder *offers* and *needs* in Fig. 5. What are the prospects of the required collaboration among the stakeholders? Operators and content providers seek a "best friends forever" union in order to mutually harvest benefits in the digital value chain while keeping their users happy. This is a favorable environment for the scenario of wireless caching. In fact, if telecom operators enable caching capabilities at the edge of their networks, their infrastructure will become sustainable while they gain access to new business models. On the other hand, content providers can benefit from a caching collaboration since:

- Traffic will be intercepted earlier and the content transport cost will be reduced.
- User demand will not be held back by sustainability issues.
- Costs associated with the deployment of large memory units will be avoided.
- They will be able to reach closer to their users and extend computing infrastructures to the fog paradigm.

Lastly, it is foreseeable that in some situations

[3] Gain obtained in hit probability when increasing M slightly

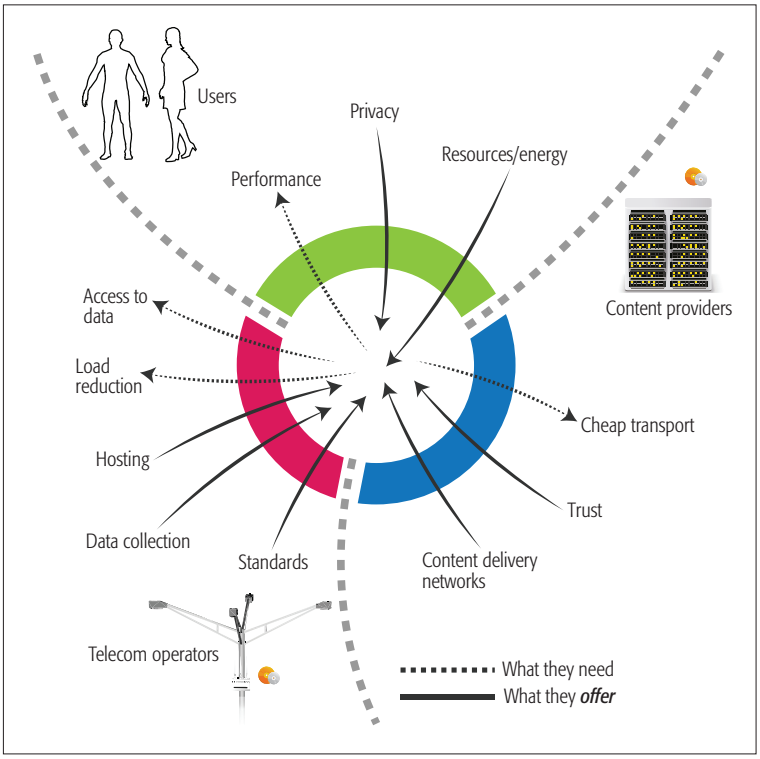

Figure 5. A stakeholder analysis.

the roles of the content provider and the wireless operator may converge.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019," White Paper, 2015, available: http://goo.gl/tZ6QMk

[2] E. Baştuğ, 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] 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.

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

[5] L. Breslau *et al.*, "Web Caching and Zipf-Like Distributions: Evidence and Implications," *IEEE INFOCOM*, Mar 1999.

[6] S. Traverso *et al.*, "Temporal Locality in Today's Content Caching: Why It Matters and How to Model It," *ACM SIGCOMM*, 2013.

[7] M. Leconte *et al.*, "Placing Dynamic Content in Caches with Small Population," *IEEE INFOCOM*, 2016.

[8] S.-E. Elayoubi and J. Roberts, "Performance and Cost Effectiveness of Caching in Mobile Access Networks," *Proc. 2nd ACM Int'l. Conf. Information-Centric Networking*, ser. ICN '15, 2015, pp. 79–88.

[9] E. Baştuğ *et al.*, "Big Data Meets Telcos: A Proactive Caching Perspective," *J. Commun. and Networks*, Special Issue on Big Data Networking-Challenges and Applications, vol. 17, no. 6, Dec. 2015, pp. 549–58.

[10] D. Naylor *et al.*, "Multi-Context TLS (mcTLS): Enabling Secure in-Network Functionality in TLS," *ACM SIGCOMM*, 2015.

[11] J. Roberts and N. Sbihi, "Exploring the Memory-Bandwidth Tradeoff in an Information-Centric Network," *ITC*, 2013.

[12] S. K. Fayazbakhsh *et al.*, "Less Pain, Most of the Gain: Incrementally Deployable ICN," *ACM SIGCOMM*, 2013.

[13] E. Baştuğ *et al.*, "Cache-Enabled Small Cell Networks: Modeling and Tradeoffs," *EURASIP J. Wireless Commun. and Networking*, no. 1, Feb. 2015, p. 41.

[14] N. Golrezaei *et al.*, "FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers," *IEEE Trans. Info. Theory*, vol. 59, no. 12, 2013, pp. 8402–13.

[15] M. Tao *et al.*, "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN," arXiv preprint arXiv: 1512.06938, 2015.

Biographies

GEORGIOS S. PASCHOS [S'01, M'06, SM'15] (georgios.paschos@huawei.com) received his electrical and communications engineering diploma (2002) and Ph.D. (2006) from Aristotle University of Thessaloniki and University of Patras, respectively, both in Greece. He was an ERCIM postdoctoral fellow for one year at VTT, Finland. From 2008 to 2012, he was affiliated with CERTH-ITI, Greece, and also taught as an adjunct lecturer in the Electrical and Communications Engineering Department of the University of Thessaly. In 2012–2014 he was a postdoctoral associate at LIDS, MIT. Since 2014, he has been a principal researcher in the Mathematical and Algorithmic Sciences Lab of Huawei Technologies, Paris, France, leading the Network Control and Resource Allocation team. His main interests are in the area of communications networks, caching, and SDN. He serves on Committees of INFOCOM and WiOPT conferences, and serves as an Associate Editor for *IEEE/ACM Transactions on Networking*.

EJDER BAŞTUĞ (ejder.bastug@centralesupelec.fr) is currently a postdoctoral researcher at CentraleSupélec. He obtained his Ph.D. from CentraleSupélec in December 2015, under the guidance of Prof. Mérouane Debbah and Prof. Jean-Claude Belfiore. His Ph.D. topic was on distributed caching methods in small cell networks, whereas his main research interests are related to stochastic geometry and machine learning tools for wireless communications. He has been on the Executive Committee of IEEE WCNC 2014, Organizing Committee of IEEE BlackSeaCom 2015, Technical Program Committee of EuCNC 2015, and Chair of several international workshops. He is on the Organizing Committee of IEEE ICC 2017, to be held in Paris. He is the recipient of the Supélec Foundation's Best Ph.D. Thesis/Publications Prize 2015 and the IEEE Communications Society Best Tutorial Paper Award 2016.

MÉROUANE DEBBAH [F] (merouane.debbah@huawei.com) entered the Ecole Normale Supérieure de Cachan, France, in 1996, where he received his M.Sc and Ph.D. degrees. He worked for Motorola Labs, Saclay, France, from 1999 to 2002 and the Vienna Research Center for Telecommunications (Vienna, Austria) until 2003. From 2003 to 2007, he was with the Mobile Communications Department of the Institut Eurecom, Sophia Antipolis, France, as an assistant professor. Since 2007, he has been a full professor at CentraleSupélec, Gif-sur-Yvette, France. From 2007 to 2014, he was the director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he has been vice-president of the Huawei France R&D Center and director of the Mathematical and Algorithmic Sciences Lab. His research interests lie in fundamental mathematics, algorithms, statistics, information, and communication sciences research. He is an Associate Editor-in-Chief of the journal *Random Matrix: Theory and Applications* and was an Associate and Senior Area Editor for *IEEE Transactions on Signal Processing* in 2011-2013 and 2013–2014, respectively. He is a recipient of the ERC grant MORE (Advanced Mathematical Tools for Complex Network Engineering). He is a WWRF Fellow and a member of the academic senate of Paris-Saclay. He has managed 8 EU projects, and more than 24 national and international projects. He has received 15 best paper awards. He was the recipient of the Mario Boella award in 2005, the IEEE Glavieux Prize Award in 2011, and the Qualcomm Innovation Prize Award in 2012. He is the co-founder of Ximinds and Ulanta.

INGMAR LAND (ingmar.land@huawei.com) is a principal researcher at the Mathematical and Algorithmic Sciences Lab, French Research Center, Huawei Technologies Co. Ltd., Paris. Before joining Huawei, he held positions as senior research fellow and research fellow at the Institute for Telecommunications Research, University of South Australia, Adelaide, 2007–2014, and as assistant professor at Aalborg University, Denmark, 2005–2006. He received his Dr.-Ing. degree in 2004 from the University of Kiel, Germany, and studied for his Dipl.-Ing. degree at the University of Ulm and the University of Erlangen-Nürnberg, Germany. His research interests are coding and information theory in wireless and wired communications, cooperative communications, multiuser communications, physical-layer security, distributed source coding, and distributed storage.

GIUSEPPE CAIRE [S '92, M '94, SM '03, F '05] (giuseppe.caire@tu-berlin.de) received his B.Sc. in electrical engineering from Politecnico di Torino, Italy, in 1990, his M.Sc. in electrical engineering from Princeton University in 1992, and his Ph.D. from Politecnico di Torino in 1994. He was a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994–1995, an assistant professor in telecommunications at the Politecnico di Torino, an associate professor at the University of Parma, Italy, a professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, and is currently a professor of electrical engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles and an Alexander von Humboldt Professor with the Electrical Engineering and Computer Science Department of the Technical University of Berlin, Germany. He served as Associate Editor for *IEEE Transactions on Communications*, 1998–2001, and for *IEEE Transactions on Information Theory*, 2001–2003. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, and the Vodafone Innovation Prize in 2015. He served on the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as an officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the fields of communications theory, information theory, and channel and source coding, with particular focus on wireless communications.