The Role of Caching in Future Communication Systems and Networks

The Role of Caching in Future Communication Systems and Networks

Georgios S. Paschos, *Senior Member, IEEE*, George Iosifidis, Meixia Tao, *Senior Member, IEEE*, Don Towsley, *Fellow, IEEE*, and Giuseppe Caire, *Fellow, IEEE*

***Abstract*— This paper has the following ambitious goal: to convince the reader that *content caching* is an exciting research topic for the future communication systems and networks. Caching has been studied for more than 40 years, and has recently received increased attention from industry and academia. Novel caching techniques promise to push the network performance to unprecedented limits, but also pose significant technical challenges. This tutorial provides a brief overview of existing caching solutions, discusses seminal papers that open new directions in caching, and presents the contributions of this special issue. We analyze the challenges that caching needs to address today, also considering an industry perspective, and identify bottleneck issues that must be resolved to unleash the full potential of this promising technique.**

***Index Terms*— Caching, storage, 5G, Future Internet, wireless networks, video delivery, coded caching, edge caching, caching economics, content delivery networks.**

I. Introduction

TODAY storage resources and caching techniques permeate almost every area of network and communication technologies. From storage-assisted future Internet architectures and information-centric networks, to caching-enabled 5G wireless systems, caching promises to benefit both the network infrastructure (reducing costs) and the end-users (improving services). In light of pressing data traffic growth, and the increasing number of services that nowadays rely on timely delivery of (rich-media) content, the following questions are inevitably raised: *can caching deliver on these promises?* and if the answer is affirmative, *what are the required research advances to this end?*

In this tutorial paper we investigate these two questions in detail. We start with a brief discussion about the historical background of caching, and then present three key factors that, in our opinion, render caching research very important today. These factors relate to the constantly evolving user needs, novel demanding network services, but also to new technologies that can make caching very effective. In Section II we present the most active research areas in caching. We analyze seminal papers and discuss the latest developments in each area, and present the advances made by the papers appearing in this Special Issue. Our goal is to provide a unified view on the different (and often disconnected) research threads in caching.

In Section III we discuss several state-of-the-art caching systems, focusing on the research challenges they raise. We also present the latest wireless caching standardization efforts that pave the way for the design of new caching architectures. Finally, we analyze a set of key open challenges, i.e., bottleneck issues that need to be resolved in order to unleash the full potential of this promising tool. These issues range from the need to analyze the economic interactions in the complex caching ecosystem, to develop methods for coping with volatile content popularity, and to devise joint caching and computing solutions.

A. Historical Perspective

The term *cache* was introduced in computer systems to describe a memory with very fast access but typically small capacity. By exploiting correlations in memory access patterns, a small cache can significantly improve system performance. Several important results related to caching strategies can be found in papers from the 1970s. Prominent examples include the oracle MIN policy that maximizes hits under an arbitrary request sequence [1], and the analysis of Least-Recently-Used (LRU) policy under stationary sequences using a Markovian model [2] or using an efficient approximation [3].

The caching idea was later applied to the Internet: instead of retrieving a webpage from a central server, popular webpages were replicated in smaller servers (*caches*) around the world, reducing *(i)* network bandwidth usage, *(ii)* content access time, and *(iii)* server congestion. With the rapid Internet traffic growth in late 1990s, the management of theses caches became complicated. This led to the proliferation of *Content Delivery*

Manuscript received March 30, 2018; revised April 29, 2018; accepted May 30, 2018. Date of current version September 12, 2018. The work of G. Iosifidis was supported by the Science Foundation Ireland, under Grant 17/CDA/4760. The work of M. Tao was supported by the National Natural Science Foundation of China under Grant 61571299 and Grant 61521062. The work of D. Towsley was supported in part by the U.S. ARL and the U.K. MoD under Agreement W911NF-16-3-0001 and in part by the NSF under Grant NSF CNS-1617437.

G. S. Paschos is with the France Research Center, Huawei Technologies, 92100 Boulogne-Billancourt, France (e-mail: georgios.paschos@huawei.com).

G. Iosifidis is with the School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin, College Green, Dublin 2, D02PN40 Ireland (e-mail: george.iosifidis@tcd.ie).

M. Tao is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: mxtao@sjtu.edu.cn).

D. Towsley is with the School of Computer Science, University of Massachusetts, Amherst, MA 01002 USA (e-mail: towsley@cs.umass.edu).

G. Caire is with the Department of Electrical Engineering and Computer Science, Technical University of Berlin, 10623 Berlin, Germany (e-mail: caire@tu-berlin.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2844939

Networks (CDNs), an integral part of the Internet ecosystem that employ monitoring and control techniques to manage interconnected caches. Research in CDNs made progress in the last decades on investigating *(i)* where to deploy the servers (*server placement*) [4]; *(ii)* how much storage capacity to allocate to each server (*cache dimensioning*) [5]; *(iii)* which files to cache at each server (*content placement*); and *(iv)* how to route content from caches to end-users (*routing policy*). However, new questions arise today as CDNs need to support services with more stringent requirements.

Recently, caching has also been considered for improving content delivery in wireless networks [6]. Indeed, network capacity enhancement through the increase of physical layer access rate or the deployment of additional base stations is a costly approach, and outpaced by the fast-increasing mobile data traffic [7], [8]. Caching techniques promise to fill this gap, and several interesting ideas have been suggested: *(i)* deep caching at the evolved packet core (EPC) in order to reduce content delivery delay [9]; *(ii)* caching at the base stations to alleviate congestion in their throughput-limited backhaul links [10]; *(iii)* caching at the mobile devices to leverage device-to-device communications [11]; and *(iv)* coded caching for accelerating transmissions over a broadcast medium [12]. There are many open questions in this area, and several papers of this Special Issue focus on this topic.

B. Caching for Future Networks

There is growing consensus that caching is poised to play a central role in future communication systems and networks, and inevitably the following question arises: *Can we tackle the upcoming challenges in caching using existing tools?* We believe that the answer to this question is an emphatic "no", providing motivation to further study caching systems. Our view is based on three key arguments that can be summarized as follows.

1) Evolution of Content Demand Characteristics: Internet-based online video services gradually replace classical Television, and new specifications (4K, QHD, 360^o, etc.) increase the bandwidth consumption per content request. Furthermore, most video files in these services need to be available in different encoding format, and this *versioning* enlarges the caching requirements. These factors drive the explosion of video traffic, which is expected to surpass 80% of the total Internet traffic [7]. At the same time, new services are emerging, such as (mobile) Augmented and Virtual Reality with even tighter bandwidth and latency requirements than typical video streaming. These services aspire to feed the users with enormous amounts of personalized sensory information and hologram depictions in real time, and hence have to rely on edge caches. Finally, the proliferation of online social networks (OSNs) is placing users in the role of content creator, thus disrupting the traditional server-client content delivery model. OSNs increase the volatility of content popularity, and create often unforeseen spatio-temporal traffic spikes. In sum, the characteristics of *cache-able* content and content demand are rapidly changing, forcing us to revisit caching architectures and caching solutions.

TABLE I
CACHING TOPICS STATISTICS

Topic and Keywords	Papers
Information-theoretic Caching Analysis	16
Fundamental Limits of Caching	16
Coded Caching Design with Practical Constraints	12
Scaling Laws of Cache Networks	7
HetNet and Device-to-device Caching	18
Edge Caching, Cooperation and Femtocaching	32
Joint Caching, Scheduling and Routing	19
Secure Caching and Privacy Preservation	6
Content Caching and Delivery	56
Algorithms for Storage Placement	17
Video caching and Streaming	14
Caching Economics	13
Caching models for ICN	12
Popularity Models and Machine Learning	9

2) Memory as a Fundamental Resource: Recent developed techniques that combine caching with coding demonstrate revolutionary *goodput* scaling in bandwidth-limited cache-aided networks [12]. This motivated the fundamental question of how memory "interacts" with other types of resources. Indeed, the topic of *coded caching* started as a powerful tool for broadcast mediums, and is being currently expanded towards establishing an information theory for memory. The first results provide promising evidence that the throughput limits of cache-enabled communication systems are in fact way beyond what is achievable by current networks. Similarly, an interesting connection between memory and processing has been recently identified [13], creating novel opportunities for improving the performance of distributed and parallel computing systems. These lines of research have re-stirred the interest in joint consideration of bandwidth, processing, and memory, and promise novel caching systems with high performance gains.

3) Memory Cloudification and New Architectures: Finally, the advent of technologies such as Software-Defined Networking (SDN) and Network Function Virtualization (NFV) create new opportunities for leveraging caching. Namely, they enable the fine-grained and unified control of storage capacity, computing power and network bandwidth, and facilitate the deployment of in-network caching services. Besides, recent proposals for content-centric network architectures place storage and caching at a conspicuous place, but require a clean-slate design approach of caching techniques. At the same time, new business models are emerging today since new players are entering the content delivery market. Service providers like Facebook are acquiring their own CDNs, and network operators deploy in-network content servers to reduce their bandwidth expenditures. These new models create, unavoidably, new research questions for caching architectures and the caching economic ecosystem.

C. About This Issue

This Special Issue received a very large number of submissions verifying that caching is an active research topic in many areas: 237 authors from Asia/Pacific (50.6% of total), 121 from Europe, Middle East, Africa (25.9%), 105 from the United States and Canada (23.5%). These statistics show that

caching appears to be most popular in P.R. China, USA, Korea, UK and France. During the review process approximately 360 experts were involved, and this indicates the large body of researchers on this topic. Table I shows a breakdown of topics in the submissions where up to three were registered per paper from its authors. These numbers are indicative of the current popularity of each topic.

The final version of the "JSAC-caching" Special Issue comprises novel technical contributions aiming to address a wide span of caching challenges for communication systems and networks. The topics include *information theory and coded caching*, *caching networks*, *caching policies and storage control*, *wireless caching techniques*, *caching economics*, and *content-based architectures*. In the following section we visit each research direction in detail, explaining the main idea and discussing the new contributions made in this Special Issue.

II. Caching: Past and Present

This Section presents the background, seminal papers, and recent developments in important research areas of caching. Furthermore, we present the papers that appear in this Issue and explain how they advance the state-of-the-art literature.

A. Information-Theoretic Caching Analysis

In 2013 Maddah-Ali and Nielsen studied the fundamental limits of broadcast transmissions in the presence of receiver caching [12]. They assumed a shared error-free medium connecting a source to K users, each one requesting a file of F bits. There are N available files in the system (in total NF bits) while each receiver can store MF bits in their cache, with $\frac{M}{N} \triangleq \gamma < 1$ denoting the relative cache size. A caching policy π performs two functions: *(i)* placement, where it decides the bits (or functions of bits) that will be stored at each cache spot; and *(ii)* delivery where it determines a sequence of multicast transmissions that ensure correct delivery, i.e. that every receiver obtains the requested file. We denote by R the vector of file requests, and by $T^{\pi}(R)$ the smallest number of transmissions under policy π such that all receivers have obtained their files indicated by R. The problem is to find π that attains $\inf_{\pi} \max_{R} T^{\pi}(R)$. Note that traditional caching policies would place a γ fraction of all files, requiring $K(1-\gamma)$ transmissions under any request.

Although this problem is largely intractable, the seminal paper [12] proposed a scheme that achieves a 12-approximation. The policy known as "centralized coded caching" is depicted in Fig. 1. During placement the policy splits the caches in parts corresponding to all subsets of users and caches different bits. During delivery, for any demand there is a XOR-based code that allows correct delivery of all demanded contents in at most $K(1-\gamma)/(1+\gamma K)$ transmissions. This provides a $1+\gamma K$ gain over classical caching. Further, the number of required transmissions converges to $\frac{1-\gamma}{\gamma}$ for $K \to \infty$. This implies that in a wireless downlink with finite resource blocks, an indefinite number of memory-equipped receivers can be simultaneously served (albeit at a small rate) which is in contrast to all previous broadcast schemes that can only serve a finite number of users in this

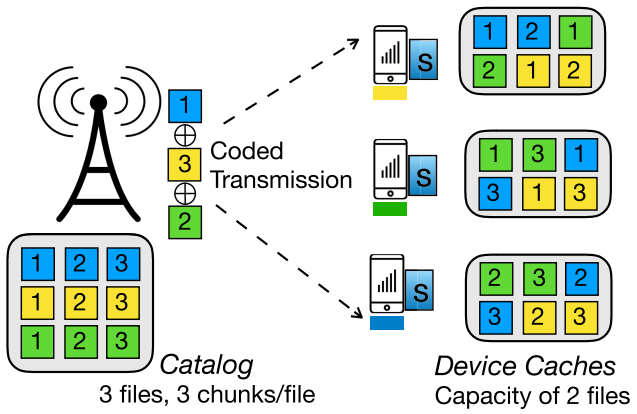


Fig. 1. Illustration of a 3 user example where coded caching offers a 3x gain over plain broadcasting by requiring only one transmission to satisfy the request. The colors depict contents with numbered chunks. The coded caching scheme consists in caching different chunks per content at each receiver, and then appropriately combining three chunks in XOR field according to the demand (the color under each user indicates the showcased demand, but we note that the scheme guarantees that one transmission is sufficient for any demand) [12].

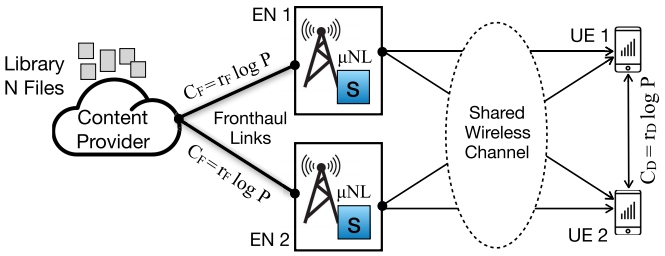
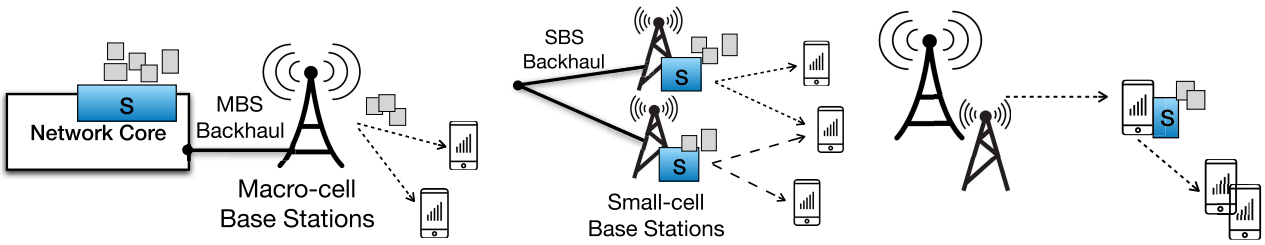


Fig. 2. Coded caching can be used in more general topologies; in the showcased example from [14] caching is used at the transmitter side to enhance the latency performance of a C-RAN-based wireless network [14].

setting. Several extensions of this scheme were subsequently considered, e.g., the scenario of adding storage capacity at the transmitters aiming to reduce latency [14], Fig. 2.

A large number of papers appearing in this Special Issue are related to coded caching. The classical coded caching scheme suffers from the subpacketization issue; the maximum gains can only be reached if the packets are split into 2^K pieces. Since an L-bit packet can be split at most L times (typically much less in practical systems), as K increases the practically observed gains diminish. This problem is studied in [15] which suggests the addition of antennas to the source. In particular it shows that W transmit antennas reduce the required subpacketization to approximately its W-th root. Similarly, [16] studies the throughput-delay trade-offs in an ad hoc network where each node moves according to a simplified *reshuffling mobility model*, and extends prior work to the case of subpacketization.

Reference [17] introduces a novel unification of two extreme and different approaches in coded caching, namely *(i)* the uncoded prefetching designed by [12], and the *(ii)* the coded prefetching designed in [18]. A scheme that generalizes both prior cases is proposed, and it is shown that it achieves new trade-offs. On the other hand [19] uses coded prefetching to



(a) Core-caching (b) Femtocaching (c) D2D Caching

Fig. 3. Different scenarios for placing storage ("S") at wireless networks. (a): Caching at the evolved packet core of a mobile network; (b): Caching at small cell base stations (Femtocaching); (c): Caching at the user device and device-to-device (D2D) content delivery.

achieve the rate-memory region of [18] with a much smaller field (of order 2^2 instead of $2^m, m \geq K \log_2 \ N$).

Other works in this issue focus on the interplay of coded caching with user demand aspects. In [20] the coded caching framework is extended to an online approach that designs placement and delivery while considering user demand asynchronism. In the same context, [21] extends the information theoretic analysis of caching to the case of C-RAN clouds with time-varying popularity, and looks at the normalized delivery time metric which captures both the delivery part of coded caching and the time needed to replenish the cache contents. In [22] the idea of coded caching is generalized to correlated content, showing how to exploit correlations in order to obtain a lower bound for the rate-memory trade-off. The paper [23] matches users to one of a subset of caches after the user request is revealed. It compares a scheme focusing on coded server transmissions while ignoring matching capabilities, with a scheme that focuses on adaptive matching while ignoring potential coding opportunities.

The paper [24] extends coded caching to wireless broadcast channels with Gaussian noise, and studies the optimization of energy efficiency. Reference [25] addresses the problem of combining network coding and caching in order to optimize a multicast session over a directed-acyclic graph. In [26] the authors study the fundamental limits of secretive coded caching, examining the case of partially known prefetched files. The work in [27] also considers secrecy constraints; it studies a two-hop network architecture known as a combination network, where a layer of relay nodes connects a server to a set of end users. A new centralized coded caching scheme is developed that jointly optimizes cache placement and delivery phase, and enables decomposing the combination network into a set virtual multicast sub-networks.

B. Caching in Wireless Systems

Beyond coded caching there are several recent proposals for cache-aided wireless network architectures, and for techniques that combine caching with other wireless communication decisions.

1) Femtocaching and D2D: Caching content at the very edge of wireless networks (base stations; user devices) is fundamentally different from caching techniques in CDNs, and raises novel challenges. Namely, in wireless networks the demand per edge cache is smaller in volume and varies rapidly with time as users move from one cell to another. Furthermore, caching decisions are coupled not only because caches share backhaul links, but also because users might be in range with multiple cache-enabled base stations. These characteristics, together with the inherent volatility of the wireless medium, render caching decisions particularly difficult to optimize and, oftentimes, less effective, e.g., in terms of the achieved cache hit ratio.

Nevertheless, several interesting proposals for wireless caching have recently appeared, Fig. 3. The seminal "femtocaching" paper [10] proposed the idea of proactive caching at small cell base stations as a solution to their capacity-limited backhaul links. The problem of minimizing the average content delivery delay was formulated and solved using submodular optimization. Many follow-up works focus on this architecture, including [28] in this Issue (discussed later). In a similar setting, [29] studied content dissemination through device-to-device (D2D) communications. It was shown that short-range D2D transmissions combined with content caching at user devices yield a throughput scaling law that is independent of the number of users.

2) Caching and Wireless Transmissions: The design of wireless transmission techniques changes significantly in the presence of caching. For example, caching at transmitters can turn an interference channel into a broadcast channel or X-channel [30]; and caching at both transmitters and receivers can turn an interference channel into a so-called cooperative X-multicast channel [31]. Clearly, physical-layer transmission and scheduling schemes have to be re-visited in cache-enabled wireless networks.

The new cache-aided designs induce a coupling between the transmissions and the caching strategies, and this gives rise to challenging mixed time-scale optimization problems. For example, [32] showed that by caching a portion of each file, the base stations can opportunistically employ cooperative multipoint (CoMP) transmission without expensive backhaul in MIMO interference networks, yielding the so-called cache-induced opportunistic CoMP gain; a technique that requires the joint optimization of MIMO precoding (small time scale) and caching policies (large time scale). The joint design of caching, base station clustering, and multicast beamforming can significantly improve the energy-backhaul trade-offs in C-RAN systems [33]. More complicated cross-time-scale interactions are investigated in [34]–[36] for either throughput maximization or service cost minimization.

3) Caching in Stochastic Wireless Networks: Another line of research that has attracted great attention is the caching

optimization in stochastic wireless networks where node locations are modeled as independent spatial random processes, e.g., Poisson Point Process (PPP). Due to advances in stochastic geometry tools for cellular networks, cf. [37], this approach facilitates the analysis and design of large-scale wireless caching systems. Assuming that base stations cache the most popular contents, the works [38], [39] derived closed-form expressions for the outage probability and the average delivery rate of a typical user as a function of SINR, base station density, and file popularity. If base stations cache the contents randomly, the optimization of caching probabilities is considered in [40]. If base stations employ maximum distance separable (MDS) codes or random linear network codes for content caching, the optimization of caching parameters is considered in [41].

Reference [42] extends caching to communication scenarios with Unmanned Aerial Vehicles (UAVs). Specifically, it proposes policies that decide jointly caching and trajectories of UAVs in order to maximize the efficiency of content delivery. Zhang *et al.* [43] investigate video caching over heterogeneous networks modeled with PPP, and study the impact of different viewing quality requirements on energy efficiency. Reference [44] uses stochastic geometry to model the locations of base stations and study different cooperative caching techniques, including coded caching. The proposed solutions demonstrate superior energy efficiency for the network over selected benchmark schemes. The idea of improving caching decisions in small cells and user devices by considering social-layer characteristics, such as mutual user interests and mobility patterns, is proposed in [45]. Finally, a mixed time-scale problem is studied in [46] with cache dimensioning at the base stations and beamforming decisions for improving content delivery in C-RAN systems.

C. ICN Architectures

Information (or, Content) Centric Networking (ICN or CCN) is a research thread that aims to provide a clean-slate design of the Internet [47]. The main idea is to redesign basic communication functions (such as routing) based on content addressing, replacing the IP-based network paradigm, cf. this survey [48]. As the traffic volume of video and other types of content grow fast, ICN architectures are becoming increasingly relevant.

The core idea in ICN is to proactively publish content to all interested Internet entities using a multicast session. This type of communications are inherently related to caching and motivate novel technical questions: *(i)* how to perform en-route caching with mechanisms such as "leave a copy" [49]; *(ii)* how to design caching structures that can scale to handle large traffic volumes; and *(iii)* how much storage to deploy in the network, and at which nodes, so as to balance costs and performance gains [50].

Another crucial topic in ICN is content discovery. While collaboration of caches increases the hit performance (by fine-tuning how often each file is replicated), in ICN this promising architecture entangles the content discovery process. Namely, in a network of caches, although a replica can be cached closer to the user, discovering its actual location may take a significant amount of time and even violate the QoS criteria due to excessive delays. This problem of content delivery is studied in [51] that appears in this Issue, which proposes the *scope-flooding* technique to propagate control signals for content discovery by building multicast trees routed at the source node. Since replica frequency is expected to relate to popularity, the authors suggest tuning the discovery radius according to content popularity.

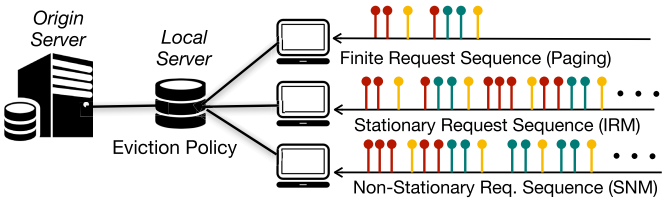


Fig. 4. A sequence of content requests towards the origin server is intercepted by a local server, which caches certain contents. Upon a request that is a miss, an eviction policy must decide which content will be evicted from the local server. The optimality of the eviction policy is measured in hits, and depends on the model for the request sequence.

Another important direction for ICN is certainly the efficient simulation of large caching installations. Due to the immense number of contents, it might be computationally-demanding (and even prohibitive) to model and simulate such systems, e.g., in order to assess the performance of different caching policies. The work of [52] revisits this problem and proposes model-driven techniques for simulating general cache networks. This solution leverages very accurate approximations of caching systems to be able to allow the simulation of hundreds of caches and trillions of contents, while employing complex routing and caching algorithms.

D. Online Caching Policies and Analytics

Online caching refers to the problem of devising a cache eviction policy in order to maximize the cache hit probability. This is a fundamental and well-studied topic, yet remains highly relevant today. The problem definition involves a cache of certain size, a file request sequence, and the eviction policy that determines which content should be removed when the cache overflows, Fig. 4. Typical versions of this problem consider: *(i)* finite request sequences and aim to devise eviction policies that maximize the number of hits; *(ii)* stationary request sequences, with the goal to maximize the hit probability (stationary behavior); and *(iii)* non-stationary sequences, where an additional challenge is to track the evolution of content popularity.

Various eviction policies have been proposed in the past, each one having different advantages. For example, the Least-Recently-Used (LRU) policy promotes file recency and optimizes performance under adversarial finite request sequences (achieves the optimal competitive ratio [53]). Similarly, Least-Frequently-Used (LFU) policies maximize hit ratio under stationary sequences by promoting the contents with the highest request frequency, while Time-To-Live (TTL) policies use timers to adjust the hit probability of each con-

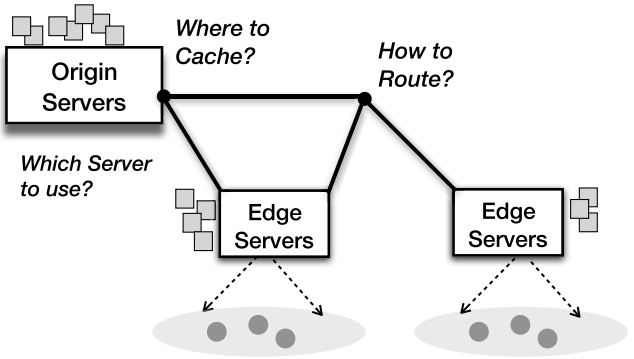


Fig. 5. Decisions in Caching Networks [135]. **Small time scale**: which server to use? where to cache? and how to route the content? The caching and routing decisions are inherently coupled, as a request can only be routed to caches where the requested item is available. **Large time scale**: where to place servers, and how to dimension the links connecting them?.

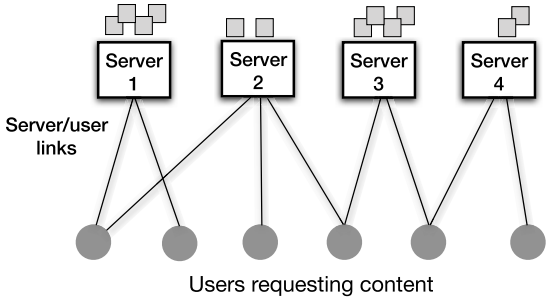


Fig. 6. Bipartite caching model [10]. A set of users is connected with caching servers. Every user can fetch content from each server with different cost, and servers cache possibly different items. The bipartite model is generic and captures several wired and wireless architectures, where link cost parameters can represent delay, energy, or monetary costs.

tent [54]. Lately, great emphasis is put on online caching with non-stationary request sequences, focusing on the practical, yet challenging, scenario of time-varying file popularity.

Indeed, when popularity varies with time caching performance cannot be determined solely based on the stationary hit ratio, and the eviction policy needs to be able to adapt to content popularity changes. In order to shed light on this aspect, [55] studies the mixing time of caching policies. It suggests that (most) eviction policies can be modeled as Markov chains, whose mixing times give us a figure of how "reactive" the policy is, or else how true to its stationary performance. The τ–distance [56] is leveraged for characterizing the learning error of caching policies. A practical lesson learned is that although multi-stage LRU policies offer tunable hit rate performance, they adapt slowly to popularity changes.

Another line of research employs prediction schemes to accurately exploit file popularity instead of relying on LRU-type eviction rules. Somuyiwa *et al.* [57] employ reinforcement learning in order to keep track of file popularity. They study proactive caching and content delivery in wireless networks and minimize the average energy cost. The model includes wireless links with time-varying channel quality, a time-evolving content catalog, and non-stationary file popularity. In [58], traces from a vehicular ad-hoc network are used, and it is demonstrated that prediction-enhanced content prefetching can indeed increase the network performance. In this case the predictions refer both to content popularity and the vehicles' location.

Reference [59] argues that users often have elastic needs and can be satisfied with similar content, if the requested items are not available at a local cache, which results in a *Soft Cache hit*. This work is in line with the recently proposed idea of leveraging recommendation systems that are embedded in several CDNs (e.g., YouTube) in order to steer user demand towards already cached content [60]. Finally, the problem of online cooperative caching (femtocaching) is considered in [61], which is essentially a multi-cache generalization of the classical paging problem. The authors propose the "lazy" qLRU policy, where only the cache that is serving a content item can update its state, and it does so with probability q. It is shown that as $q \to 0$, the performance of this policy achieves a stationary limit which is a local maximum, in the sense that no performance improvement can be obtained by changing only one content's placement frequency.

E. Content Caching and Delivery Techniques

Modern caching systems are essentially networks of interconnected caches. Therefore, the caching problem in its entirety includes decisions about *server placement* and *cache dimensioning*, *content placement*, and *content routing*, Fig. 5. One can consider additional design parameters for these *caching networks* (CNs) as, for example, dimensioning the links and the cache serving capacity.

1) Cache Deployment and Dimensioning: The storage (or, cache) deployment has been extensively studied and we refer the reader to [62] for a survey. The cache deployment problem aiming to minimize content delivery delay has been formulated as a K-median problem [4], [63], and as a facility location problem [64]. The work [65] studies a variation, considering the cost of syncing different caches (ensuring consistent copies). It is shown that increasing the number of caches beyond a certain threshold induces costs higher than the performance benefits. When the network has a tree structure these deployment problems can be solved efficiently using dynamic programming [66], [67].

2) Hierarchical Caching Networks: Indeed, CDNs or IPTV networks have often a tree-like form which facilitates the design of caching and routing policies. Caching at leaf nodes improves the access time, while caching at higher layers increases the cache hit ratio. Hierarchical networks are typically studied for 2-layers, often with the addition of a distant root server. The seminal work [68] presented a polynomial-time exact algorithm for the min-delay content placement problem when leaf caches can exchange their files. Motivated by an actual IPTV network, [69] studied a similar problem for a 3-level hierarchical caching network. Also [70] considers hierarchical caching for a wireless backhaul network and designed a distributed 2-approximation caching algorithm. Another important objective is to minimize the requests sent to the root server in order to reduce off-network bandwidth and

server congestion [71]. In many cases, these multi-tier CNs can be modeled also as bipartite caching networks Fig. 6, where the links capture the cost of the entire path connecting the user with each cache.

3) General Caching Networks: There are also general CN models [72], [73] where the link delay increases non-linearly with the number of requests [74], or the objective functions are non-linear to the cache hit ratio [75]. In some of these cases the problem has a convenient convex or submodular structure and hence greedy algorithms ensure a 2-approximation ratio [74]. Finally, in the general case, routing can involve multihop or multipath decisions. This means that, if there are hard capacity constraints or load-dependent link costs, the routing decisions are not fully determined by the caching decisions (as e.g., in femtocaching), and therefore the routing and caching policies need to be jointly devised.

The paper [58] that appears in this Special Issue presents the interesting scenario of a 3-tier vehicle ad hoc network that includes origin servers, regional servers and road-side units. Focusing on general network architectures, [76] studies the minimum-cost joint routing and caching problem and shows the benefit over considering the two problems separately. The study includes both hop-by-hop routing and source-routing decisions, and proposes distributed and online solution algorithms that achieve constant approximation ratio in polynomial time. Finally, in [28] the femtocaching problem is extended to the setting where files can be stored for a limited duration at a cache and delivered with multicast. The authors provide performance guarantees for a greedy algorithm that selects jointly caching retention times and routing decisions.

F. Video Caching

Due to the popularity of video applications and the large size of the involved files, video content delivery is currently a very important research topic in caching. On the one hand, there are obviously tight delay constraints, especially for streaming services where successive video segments need to be delivered in sync so as to avoid playback stalls. On the other hand, caching decisions are perplexed due to the multiple encoding options. Each video comes in several versions, each of different size, and furthermore the users might have inelastic or elastic needs in terms of video quality. Finally, it is worth mentioning that in live video streaming requests can be predicted with higher precision [77], and this facilitates caching decisions.

The early work [78] proposed scheduling policies that leverage caching at the user side (buffering) to improve video delivery, and [79] suggested proactive caching of video segments in a peer-assisted CDN. The simplest scenario is that of video on demand (VoD) delivery where one needs to decide which version(s) of each video item to cache [80]. When the video versions are independent, the caching decisions are only coupled due to the fact that users may have elastic quality needs (hence, the video versions are complementary). When, however, scalable video coding (SVC) is used [81], additional constraints appear as users can fetch and combine layers of the same video file from different caches.

The work [75] studies joint video caching and routing in HetNets, aiming to reduce delivery delay and network expenditures, and [82] focuses on delay minimizing policies which are, therefore, also suitable for video streaming. Similarly, [83] formulates a min-cost routing and caching problem for a large VoD network, and [84] analyzes the benefits of SVC for caching networks that deliver video streaming services. More recently, the focus has been shifted to wireless networks with proposals for collaborative video caching [85], joint routing and caching [86], or network-coding assisted video caching [87].

Several papers appearing in this Issue study the delivery of multimedia content. The paper [88] considers a C-RAN edge caching solution for multimedia services and introduces a dynamic policy for jointly deciding the configuration of the virtual machines (storing the content), the caching decisions, and the user request routing policy. A different architecture is considered in [89] which studies mobile video delivery through D2D links. A base station seeds the devices with videos of possibly different encoding quality, and nearby devices collaborate by exchanging these files. This is formulated as a dynamic problem that maximizes the time-average video quality, through caching and D2D scheduling policies.

Finally, [90] analyzes HTTP-based live streaming services where mobile devices request ultra-high video quality. In these services the end-users often have deteriorated Quality-of-Experience due to the employed congestion control mechanisms in TCP. The authors propose a solution for this problem which employs context-aware transient holding of video segments at the cache-enabled mobile edge. This approach eliminates buffering and reduces the initial startup delay and live stream latency.

G. Caching Economics

The economics of caching is perhaps one of the least explored research areas, rapidly gaining momentum due to the advances in virtualization, that enhance the flexibility in managing storage resources. Prior works in this area can be broadly categorized to: *(i)* caching cooperation mechanisms, and *(ii)* pricing methods.

1) Cooperation Mechanisms: In previous works, e.g., see [68] and [91], the term "cooperative caching" was used to describe systems where content requests are served by any cache belonging to a set (or, network) of caches. These works, however, take for granted the cooperation among CDNs, mobile operators and users. In practice, these self-interested entities will share their storage resources and coordinate their caching policies, only if they will benefit from cooperation. Prior work shows that if the *incentive alignment* problem is not solved, caching systems experience significant performance loss [92]. Later, [93] proposed a cooperation mechanism (for a general CN) based on the Nash Bargaining Solution. The latter is attractive as it disperses the cooperation benefits proportionally to the performance each entity would have achieved under non-cooperation. A different suggestion is to use pricing where co-located caches pay for the content they receive from each other, e.g., [85].

Incentives may be also offered to users in order to assist the network. For example, [94] discusses the problem of

incentivizing users to exchange content by leveraging D2D communications. In a different example, [95] proposed a solution where an operator can lease storage and wireless bandwidth of residential access points. Such solutions that involve user equipment are important as their benefits scale with the demand. A related business model is proposed in this Issue [96], where a Mobile Network Operator (MNO) leases its edge caches to a Content Provider (CP). The latter aims to offload user requests at the edge by maximizing the edge cache hit-ratio with the minimum possible leasing cost. The authors introduce an analytical framework for optimizing CP decisions, which are conditioned on the user association policy of the network. This is an increasingly relevant scenario and follows proposals for deploying edge storage resources at mobile networks, namely at the EPC or base stations.

2) Pricing Mechanisms: The caching economic ecosystem is complex as it includes: payments from Content Providers (CP) to CDNs for content delivery, from CDNs to ISPs for bandwidth and in-network storage, and from users to CPs and ISPs. Pricing employed by the CDN affects how much content the CP places at the edge, which in turn impacts ISP costs and user-perceived performance. The work [97] studies revenue-maximizing CDN policies, while [98] proposes a flexible CDN pricing method. It was shown in [99] that a revenue-seeking cache owner should offer both best effort and guaranteed content delivery services. On the other hand, [100] has shown that ISPs can increase their marginal profits by imposing data plan caps to users, thus inducing CPs to charge the users with lower prices. These interactions are further complicated by the new business models such as Telco-CDNs, CP-CDNs, or elastic CDNs [101].

Finally, it is crucial to make the distinction between *popular* content items (that typical caching algorithms consider), and *important* items that yield higher revenue. For example, [54] models caching as a utility maximization problem instead of cache hit-ratio or delay optimization problem. This allows us to capture the different importance (and hence price) of each content item. Going a further step, [102] proposed dynamic pricing and content prefetching techniques for wireless networks, assuming that operators directly charge the end-users for content access.

III. Open Issues in Caching

In this Section we present a set of important open problems in caching. We first discuss representative state-of-the-art caching systems and the challenges they bring. The solution of these problems, clearly, is of high priority and motivates certain research directions that we further analyze.

A. Notable Existing Caching Systems

1) Akamai Intelligent Platform: Akamai owns one of the largest CDNs, delivering today 20% of the Internet traffic. The 216K caching servers of its *intelligent platform* [103] are dispersed at network edges (Points-of-Presence, PoPs) offering low-latency (1-10msec) content access around the globe. Several technical challenges arise in such large delivery platforms. First, it is necessary to protect websites from Distributed-Denial-of-Service attacks [104], and this need motivates the development of caching and filtering techniques that can deal with large volume of requests. Second, the idea of *deep* (or, edge) caching in PoPs improves the CDN performance but reduces user demand per cache, and hence makes the file popularity at the local level highly volatile [105]. This requirement stirs research in the area of edge caching, where the goal is to achieve a high hit ratio in caches placed very close to demand (end-user).

Finally, Akamai, among others, uses the idea of cloud or *elastic* CDN where storage resources are dynamically adapted to meet demand [106]. This architecture couples storage deployment and caching decisions. Hence, it renders imperative the efficient design of joint storage allocation and content caching policies, and also gives rise to new business models for content caching.

2) Google: The Google Global Cache (GCC) system comprises caches installed at ISP premises. The goal of GCC is to serve locally requests for YouTube content, reducing this way off-network bandwidth costs [107]. This system grew substantially after YouTube adopted https traffic encryption in 2013. The importance of GCC motivates the study of peering relations between content providers and network operators, and in particular the design of pricing models for leasing in-network caching capacity at operator premises. Another related challenge is security. Prior work has proposed schemes for caching with content confidentiality [108], which allows transparent caching for encrypted flows. The dominance of end-to-end encryption motivates further research on the topic of caching encrypted content.

3) Netflix Open Connect: Similarly to GCC, the Netflix CDN is partially deployed within ISPs [109]. However, Netflix video caching faces different challenges from YouTube, mainly because its catalogue is much smaller and the file popularity more predictable. As such, Netflix has been very innovative in studying spatio-temporal request profiles, popularity prediction mechanisms, and mechanisms to preload the caches overnight and reduce the daylight traffic footprint. An open research challenge in this context is the early detection of popularity changes, and online classification of video files as to whether they are cache-worthy or not.

4) Facebook Photo CDN: Facebook uses its own hierarchical CDN for delivering pictures to its users. The system leverages web browser caches on user devices, edge regional servers, and the origin caches [110]. Notably, browser caches serve almost 60% of traffic requests, due to the fact that users view the same content multiple times. Edge caches serve 20% of the traffic (i.e., approximately 50% of the traffic not served by browser caches), and hence offer important off-network bandwidth savings by locally serving the user sessions. Finally, the remaining 20% of content requests are served at the origin, using a combination of slow back-end storage and a fast origin-cache. The information flow in the Facebook network involves the generation and exchange of content among users, which is the prototypical example of ICN systems. It is therefore of interest to study how ICN caching techniques can improve this architecture.

5) Amazon AWS: Part of AWS is the Amazon Cloudfront, a virtual CDN which utilizes the cloud storage to provide CDN services. Storing 1TB is priced at $20 [111], and Amazon allows one to dynamically rent caching resources by changing the storage size every one hour. This cloud or elastic CDN architecture, along with similar solutions proposed by Akamai and others, motivate further research on the arising business models, as well as on the dynamic cache placement and dimensioning.

6) Cadami: The Munich-based startup Cadami was the first to implement and evaluate coded caching in a real system [112]. The company demonstrated live streaming to 30 nodes, producing wireless transmission gains of $\times 3$ with realistic wireless channels, file subpacketization, and coding overheads. The most promising applications for such solutions are entertainment systems in long-haul flights, and satellite broadcast systems, and call for further research in coded caching, a topic well-represented in this Special Issue.

7) 3GPP Standards: Employing caching in upcoming 5G wireless networks has been discussed and proposed by many companies in the scope of 3GPP. For example, T-DOC R3-160688 proposes to place an edge cache at an LTE base station either embedded in eNodeB or standalone. T-DOC R3-160828 explores the benefits of local caching. In the scope of 5G-RAN R.14, the report 3GPP TR 36.933 (published in 03-2017) describes the different caching modules that are included in 5G base stations. These standardization efforts pave the road for further research in wireless caching.

B. Caching and Cloud Computing

Caching techniques are expected to play an important role in the upcoming cloud-based network architectures. We describe below two research directions on the topic of cloud memory.

1) Coded Distributed Computing: In 2004, Dean and Ghemawat (Google) proposed *map-reduce*, where a large-scale computing task is broken into simple processing of key/value pairs and assigned to parallel processors in an interconnected cluster. The map-reduce paradigm uses the *map* phase to assign processing tasks, and the *reduce* phase to combine answers in order to produce the final result, significantly reducing the total computation time of certain operations (e.g., matrix inversion). The idea of *coded distributed computing* [13] suggests to use coded caching on the reduce step. Combined with careful task assignment, this can dramatically decrease the communication bandwidth consumed during the reduce phase. This approach essentially allows us to trade-off node storage with link bandwidth, and can thus accelerate the network-limited map-reduce systems. An example is shown in Fig. 7. This recent finding reveals a hitherto hidden connection, or *interoperability*, among bandwidth, storage, and processing; and creates new possibilities for improving network performance through their joint consideration.

2) Caching and Virtualization: The network virtualization techniques continue to gain momentum, and SDN/NFV are considered key enablers for the next generation of cloud-based networks. In this context, CDNs are also expected to migrate to clouds. In particular, caching functionality will

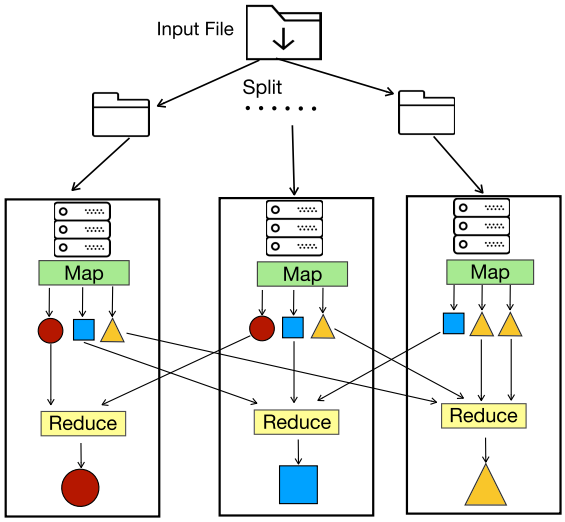


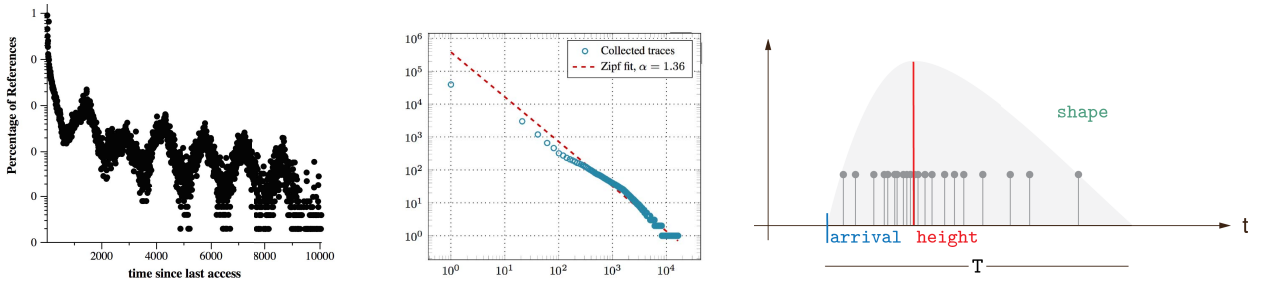
Fig. 7. An example of coded distributed computing.

be implemented as a Virtual Network Function (VNF) by means of software. Caching VNFs will provide a very flexible environment; they will be instantiated, executed, scaled and destroyed on the fly. Allocating resources for cache VNFs falls into the general framework of network slicing [113], with some special constraints. For example, populating a cache is time-demanding and bandwidth-consuming. Importantly, caches do not satisfy flow conservation; the incoming traffic is partially served by the cache, and only a fraction of traffic continues towards the origin server. Specifically, the larger the caching resource of the VNF, the greater the flow compression. Therefore, VNF embedding for caching must be generalized to include flow compression/decompression constraints [137]. These new considerations call for a generalization of the available theory for caching networks.

C. Caching in 5G Wireless Networks and Beyond

In the wireless domain, the interoperability of caching, computing and communication techniques opens exciting research directions. Caching trades scarce wireless communication bandwidth and transmission power with (the more cost-effective) memory storage by introducing traffic time-reversal into the system. Caching also enables edge computing capabilities by pre-installing necessary computing software and datasets at the wireless edge nodes. As such, investigating the interplay between these resources is essential for the development of future wireless networks, and several important research questions in this area have been already identified.

1) Performance Characterization of Cache-Enabled Wireless Networks: The first question is information-theory oriented and is related to defining and characterizing the performance limits of cache-enabled wireless networks. In traditional wireless networks, the transmission rate has been a universal performance metric, expressed as a function of signal-to-noise ratio. In the emerging cache-enabled wireless networks, due to the additional memory resource, which varies in size and location, previously adopted performance metrics have become diversified. They include hit probability [114], [38], delivery rate [29], [38], delivery



(a) Impact of Time on Popularity (b) Popularity Distribution (c) Poisson Shot Noise Model

Fig. 8. (a):Percentage of references as a function of the time since last access to the same document by the same user (log-log scale), from [118]. (b): http requests in mobile operator [136] (2015); (c): Poisson Shot noise model [120].

latency [10], [115], [31], [14], and traffic load [12]. Whether we need a universal metric that can capture, in a satisfactory fashion, the performance as a function of multi-dimensional resources is a question worth investigating. And if the answer is affirmative, we may need to expand the classic Shannon-type network information theory to study its limiting performance.

2) Tools for Wireless Optimization: The second research direction concerns the development of efficient and effective algorithms for the optimization of cache-enabled wireless networks. The joint optimization of cache placement and physical layer transmission is often NP-hard [10] and involves mixed time-scale optimization [32]. This makes these problems particularly challenging to solve, even more since their scale is typically very large and they need to be solved very fast (for enabling dynamic decisions). Furthermore, in wireless networks there are often multiple (collaborating) caches in range with the users, e.g., in multi-tier HetNets, and many possible paths connecting them. Despite the many efforts for designing algorithms that can solve these combinatorial problems (e.g., dual methods, pipage rounding, randomized rounding, etc.), practical challenges prohibit currently the application of these techniques in real systems and further progress needs to be made.

3) Support of Emerging Wireless Services: Finally, another key research thread is to explore the interplay between caching, computing and communications to boost future emerging wireless services, such as mobile AR/VR applications and *vehicle-to-everything* (V2X) communications. These services will often rely on cooperative caching and this raises additional technical questions. Namely, in multi-access caching, finding the route to nearest content replica is a practical challenge, since these services have very limited tolerance in route discovery delay. Therefore, it is important to simplify routing decisions and design them jointly with content discovery. Another interesting aspect is that these services often involve multicast or broadcast transmissions which can greatly benefit from caching. For example, delayed broadcast is currently implemented with parallel unicast sessions, but could be more bandwidth-efficient if caching is employed.

D. Caching With Popularity Dynamics

Understanding content popularity is essential to cache optimization; it affects the deployment of caching networks and the design of caching policies, shaping to large extent the overall network performance. In fact, the very notion of diverse popularity of the different content items is what motivated the idea of caching in the first place: "cache popular items to create a large effect with a small cache". Yet, understanding, tracking, and predicting the evolution of file popularity in real world is complex and, often, misinterpreted. The community is actively seeking answers to these questions.

1) Accurate Popularity Models: A large part of the literature employs the well-known *Independent Reference Model* (IRM), which assumes that the content requests are drawn in an i.i.d. fashion from a given distribution. Admittedly, IRM leads to tractable caching models but often at the expense of accuracy. For example, using IRM, we can draw power-law samples in an i.i.d. fashion to depict quite accurately the request in a real system during a short time interval. Indeed, the power-law models have been shown to characterize very accurately the popularity within a short time frame [116], i.e., in an interval when popularity can be assumed fixed. However, content popularity is in reality far from stationary, and might change significantly even over few hours. For example, requests of Wikipedia articles have a rapid day-to-day change in popularity rank: half of the top 25 contents change in a single day [117]. In Fig. 8(a) the phenomenon of "temporal locality" is demonstrated, where recently requested contents tend to be more popular [118] (note the decrease in the request rate envelope). *In summary, applying IRM to a large time-scale analysis is clearly problematic.*

The importance of content popularity dynamics is reflected in the proliferation of online caching policies such as LRU and LFU, which adapt caching decisions to temporal locality and popularity fluctuations. These policies do not necessarily provide the best performance, but they are championed in practical engineering systems because they capture some aspects of non-stationarity and they are easy to implement. These policies are often analyzed with stationary popularity or adversarial models. For example, LFU is optimal under IRM (it converges to "cache the most popular"), LRU over IRM can be analyzed with the characteristic time approximation [119], and LRU has optimal competitive ratio when the requests are chosen adversarially [53]. Recently, a number of works studied the performance of dynamic policies with non-stationary popularity. In [120], a inhomogeneous Poisson model was proposed to capture non-stationary popularity, called the Poisson Shot Noise (PSN) model. Under PSN, the LRU performance is

provided in [121], while [122] gives the optimal policy called "age-based threshold" which takes into account the frequency and the age of a content. However, a problem with PSN is that it has too many degrees of freedom, making it quite cumbersome for fitting to real data and optimizing caching systems. The quest for the right non-stationary model is still open.

2) Content Popularity Prediction: Rather than following such reactive techniques, a recent research trend aims to predict content popularity and then optimize accordingly the content placement. For example several past papers look at how a trending file will evolve [123], or how social networks can be used to predict file popularity [124]. More recently, several machine learning techniques have been proposed to assimilate popularity changes, namely bandit models [125], recommendations [60], Q-learning [126], transfer learning [127], etc. However, due to its non-stationary nature, popularity is not easily predicted. In this Special Issue alone, there were 14 submissions on this topic, which reflects how inspiring this challenge is, but also how many different viewpoints are taken on this subject. We mention here some practical challenges: *(i)* apart from the content popularity, the catalogue of contents is also evolving; *(ii)* the learning rate depends on the volume of observed samples, and consequently on the aggregation layer of the cache. Learning the popularity at the edge is thus very challenging; *(iii)* the content popularity depends on the user community characteristics, and geographical clustering of caches has the potentially to improve learning [122].

E. Cooperation, Incentives, and Pricing

As the caching ecosystem grows more complex, it becomes imperative to align the interests of the key stakeholders so as to alleviate market inefficiencies. Indeed, similarly to other networking areas, is also true that in caching many technical issues can be solved with economic mechanisms.

1) Pricing Cached Content and Elasticity: User demand often exhibits elasticity that the network can exploit to improve the services and reduce content delivery costs. Users, for example, can often delay their requests and download large content files during off-peak hours, or can use non-congested network paths (e.g., Wi-Fi links). Moreover, the users can submit their content requests in advance so as to assist the network in serving them proactively [102]. They can even adapt their requests, e.g., selecting a lower video quality or an already cached video [60]. There are two important open questions here: how to better exploit this elasticity so as to maximize caching performance (or minimize costs) and how to incentivize users to comply accordingly.

These questions open the discussion about *smart pricing* techniques for cached content that extend beyond managing network congestion [128]. There is clearly an opportunity to couple caching policies with the importance of each content file, measured in terms of revenue (user payments). First steps towards this direction have been made, e.g., see [54], [129] where content popularity is not the sole caching criterion. Charging the cached content delivery in proportion to the induced bandwidth consumption, inversely proportional to its expected cache hit ratio, or based on the service quality improvement it offers to the user, are only some first intuitive suggestions worthwhile investigating.

2) Network and Cache Sharing: The deployment of infrastructure entails huge capital and operational costs which constitute high market-entry barriers. A solution to this problem is to virtualize and share storage resources, e.g., different CDNs can jointly deploy and manage edge servers. These architectures require mechanisms for deciding: *(i)* how much capital each entity should contribute for the shared storage? *(ii)* how to allocate the virtualized capacity? There are (at least) two levels of cooperation: agree to share the physical resources (*virtualized storage*), and share the cached content (*virtualized content*). The latter option brings higher benefits if the CDNs design jointly their caching policies, and this is one of the most interesting open scenarios in this topic.

Furthermore, cooperation of CDNs with ISPs can bring significant performance and economic benefits. Selecting jointly, for example, the server and route for each content request can reduce both the service delay and network congestion [133]. This coordination is expected to yield significant benefits for wireless edge caching where the network state and demand are highly volatile. Yet, we need to explore how this coordination can be realized, meaning we have to study how to solve these joint optimization problems (caching is already a complex one), and how to disperse the benefits to the collaborating CDNs and ISPs. Finally, elastic CDNs create a new set of problems where cache dimensioning and content caching decisions are devised in the same time scale [101]. The flexibility of these architectures enables the very frequent update of these decisions, and therefore it is important to optimize long-term performance criteria for a collection of policies (instead of single-policy metrics).

3) Incentive Provision for Hybrid Architectures: User-owned equipment has increasing capacity and can be considered as an effective caching element of the network. The idea of hybrid CDN-P2P systems is an excellent example in this direction where content delivery (e.g., software patches) is assisted by the end-users [134]. In future networks this model can deliver even higher benefits (e.g., asymptotic laws of D2D caching) as it transforms negative externalities (congestion) to positive externalities (through users' collaboration). Yet, this solution requires the design of incentive mechanisms for the users, a problem that has been studied for connectivity services [132], [131]. Nevertheless, in case of user-assisted caching new questions arise: How to charge for content that is cached at the user devices? How is time affecting the content price (*freshness*)? How to minimize the cost users incur when delivering the content?

IV. Conclusions

Caching techniques have a central role in future communication systems and networks. Their beneficial effects are expected to crucially impact core parts of our communication infrastructure, including clouds, 5G wireless systems, and Internet computing at large. At the same time, the ecosystem of caching is ever-changing, constantly requiring ideas for novel

architectures and techniques for optimizing their performance. These developments, combined with the recent advances in the domain of resource interactions between storage, bandwidth and processing, create a fascinating research agenda for the years to come.

ACKNOWLEDGMENTS

The opinions expressed in this paper are of the authors alone, and do not represent an official position of Huawei Technologies.

The authors would like to acknowledge the excellent work of all reviewers who participated in this double JSAC Issue; and the great support they received from Max Henrique Machado Costa, JSAC Senior Editor, and Laurel Greenidge, Executive Editor.

REFERENCES

[1] L. A. Belady, "A study of replacement algorithms for a virtual-storage computer," *IBM Syst. J.*, vol. 5, no. 2, pp. 78–101, 1966.

[2] W. F. King, III, "Analysis of demand paging algorithms," in *Proc. IFIP Congr.* Amsterdam, OX, USA: North Holland, 1971, pp. 485–490.

[3] R. Fagin, "Asymptotic miss ratios over independent references," *J. Comput. Syst. Sci.*, vol. 14, no. 2, pp. 222–250, 1977.

[4] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the placement of Web server replicas," in *Proc. IEEE INFOCOM*, Apr. 2001, pp. 1587–1596.

[5] T. Kelly and D. Reeves, "Optimal Web cache sizing: Scalable methods for exact solutions," *Comput. Commun.*, vol. 24, no. 2, pp. 163–173, 2001.

[6] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[7] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast*, document 1454457600805266, Cisco Public Information, Mar. 28, 2017.

[8] P. Cerwall *et al.*, "The Ericsson mobility report," Ericsson, Stockholm, Sweden, White Paper EAB-18:004510 Uen, Revision A, Jun. 2018.

[9] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. ACM MobiSys*, 2013, pp. 319–332.

[10] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[11] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[12] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[13] S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, "A fundamental tradeoff between computation and communication in distributed computing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 109–128, Jan. 2018.

[14] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[15] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1176–1188, 2018.

[16] A. Malik, S. H. Lim, and W.-Y. Shin, "On the effects of subpacketization in content-centric mobile networks," *IEEE J. Sel. Areas Commun.*, to be published.

[17] K. Zhang and C. Tian, "Fundamental limits of coded caching: From uncoded prefetching to coded prefetching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1153–1164, 2018.

[18] C. Tian and J. Chen. (Apr. 2016). "Caching and delivery via interference elimination." [Online]. Available: https://arxiv.org/abs/1604.08600

[19] J. Gómez-Vilardebó, "A novel centralized coded caching scheme with coded prefetching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1165–1175, 2018.

[20] Y. Lu, W. Chen, and H. V. Poor, "Coded joint pushing and caching with asynchronous user requests," *IEEE J. Sel. Areas Commun.*, to be published.

[21] S. M. Azimi, O. Simeone, A. Sengupta, and R. Tandon, "Online edge caching and wireless delivery in fog-aided networks with dynamic content popularity," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1189–1202, 2018.

[22] P. Hassanzadeh, A. M. Tulino, J. Llorca, and E. Erkip, "On coding for cache-aided delivery of dynamic correlated content," *IEEE J. Sel. Areas Commun.*, to be published.

[23] J. Hachem, N. Karamchandani, S. Moharir, and S. Diggavi, "Caching with partial adaptive matching," *IEEE J. Sel. Areas Commun.*, to be published.

[24] M. M. Amiri and D. Gündüz, "Caching and coded delivery over Gaussian broadcast channels for energy efficiency," *IEEE J. Sel. Areas Commun.*, to be published.

[25] M. Mahdian, N. Prakash, M. Médard, and E. Yeh, "Updating content in cache-aided coded multicast," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1203–1216, 2018.

[26] Y. P. Wei, K. A. Banawan, and S. Ulukus, "Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1126–1139, 2018.

[27] A. A. Zewail and A. Yener, "Combination networks with or without secrecy constraints: The impact of caching relays," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1140–1152, 2018.

[28] S. Shukla, O. Bhardwaj, A. A. Abouzeid, T. Salonidis, and T. He, "Proactive retention-aware caching with multi-path routing for wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1286–1299, 2018.

[29] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, Dec. 2015.

[30] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE ISIT*, Jun. 2015, pp. 809–813.

[31] F. Xu, M. Tao, and K. Liu, "Fundamental trade-off between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[32] A. Liu and V. K. N. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[33] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[34] N. Abedini and S. Shakkottai, "Content caching and scheduling in wireless networks with elastic and inelastic traffic," *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 864–874, Jun. 2014.

[35] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[36] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6284–6297, Sep. 2016.

[37] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[38] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *J. Wireless Commun. Netw.*, vol. 41, 2015.

[39] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[40] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4341–4354, May 2016.

[41] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[42] X. Xu, Y. Zeng, Y. L. Guan, and R. Zhang, "Overcoming endurance issue: UAV-enabled communications with proactive caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1231–1244, 2018.

[43] X. Zhang *et al.*, "Energy-efficient caching for scalable videos in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, to be published.

[44] N. Deng and M. Haenggi, "The benefits of hybrid caching in Gauss-Poisson D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1217–1230, 2018.

[45] X. Li, X. Wang, P.-J. Wan, Z. Han, and V. C. M. Leung, "Hierarchical edge caching in device-to-device aided mobile networks: Modeling, optimization, and design," *IEEE J. Sel. Areas Commun.*, to be published.

[46] B. Dai, Y.-F. Liu, and W. Yu, "Optimized base-station cache allocation for cloud radio access network with multicast backhaul," *IEEE J. Sel. Areas Commun.*, to be published.

[47] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. ACM CoNEXT*, 2009, pp. 1–12.

[48] G. Xylomenos *et al.*, "A survey of information-centric networking research," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 1024–1049, 2nd Quart., 2014.

[49] V. Sourlas, L. Gkatzikis, P. Flegkas, and L. Tassiulas, "Distributed cache management in information-centric networks," *IEEE Trans. Netw. Service Manage.*, vol. 10, no. 3, pp. 286–299, Sep. 2013.

[50] S.-E. Elayoubi and J. Roberts, "Performance and cost effectiveness of caching in mobile access networks," in *Proc. ACM ICN*, 2015, pp. 79–88.

[51] L. Wang, S. Bayhan, J. Ott, J. Kangasharju, and J. Crowcroft, "Understanding scoped-flooding for content discovery and caching in content networks," *IEEE J. Sel. Areas Commun.*, to be published.

[52] M. Tortelli, D. Rossi, and E. Leonardi, "Parallel simulation of very large-scale general cache networks," *IEEE J. Sel. Areas Commun.*, to be published.

[53] D. D. Sleator and R. E. Tarjan, "Amortized efficiency of list update and paging rules," *Commun. ACM*, vol. 28, pp. 202–208, Feb. 1985.

[54] M. Dehghan, L. Massoulie, D. Towsley, D. Menasche, and Y. C. Tay, "A utility optimization approach to network cache design," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[55] J. Li, S. G. Shakkottai, J. C. S. Lui, and V. Subramanian, "Accurate learning or fast mixing? Dynamic adaptability of caching algorithms," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1314–1330, 2018.

[56] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM J. Discrete Math.*, vol. 17, no. 1, pp. 134–160, 2003.

[57] S. O. Somuyiwa, A. György, and D. Gündüz, "A reinforcement-learning approach to proactive caching in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1331–1344, 2018.

[58] Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. I. Braun, and S. Sargento, "Mobility prediction-assisted over-the-top edge prefetching for hierarchical VANETs," *IEEE J. Sel. Areas Commun.*, to be published.

[59] P. Sermpezis, T. Giannakas, T. Spyropoulos, and L. Vigneri, "Soft cache hits: Improving performance through recommendation and delivery of related content," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1300–1313, 2018.

[60] L. E. Chatzieleftheriou, M. Karaliopoulos, and I. Koutsopoulos, "Caching-aware recommendations: Nudging user preferences towards better caching performance," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[61] E. Leonardi and G. Neglia, "Implicit coordination of caches in small cell networks under unknown popularity profiles," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1276–1285, 2018.

[62] J. Sahoo *et al.*, "A survey on replica server placement algorithms for content delivery networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1002–1026, 2nd Quart., 2017.

[63] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 568–582, Oct. 2000.

[64] M. Bateni and M. Hajiaghayi, "Assignment problem in content distribution networks: Unsplittable hard-capacitated facility location," *ACM Trans. Algorithms*, vol. 8, no. 3, 2012, Art. no. 20.

[65] E. Cronin, S. Jamin, C. Jin, A. R. Kurc, D. Raz, and Y. Shavitt, "Constrained mirror placement on the Internet," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1369–1382, Sep. 2002.

[66] A. Benoit, V. Rehn-Sonigo, and Y. Robert, "Replica placement and access policies in tree networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 12, pp. 1614–1627, Dec. 2008.

[67] B. Li, M. J. Golin, G. F. Italiano, X. Deng, and K. Sohraby, "On the optimal placement of Web proxies in the Internet," in *Proc. IEEE INFOCOM*, Mar. 1999, pp. 1282–1290.

[68] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, "Placement algorithms for hierarchical cooperative caching," *J. Algorithms*, vol. 38, no. 1, pp. 260–302, 2001.

[69] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li, "Collaborative hierarchical caching with dynamic request routing for massive content distribution," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2444–2452.

[70] V. Pacifici, S. Jošilo, and G. Dán, "Distributed algorithms for content caching in mobile backhaul networks," in *Proc. ITC*, Sep. 2016, pp. 313–321.

[71] K. Poularakis and L. Tassiulas, "On the complexity of optimal content placement in hierarchical caching networks," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2092–2103, Mar. 2016.

[72] T. Bektaş, J.-F. Cordeau, E. Erkut, and G. Laporte, "Exact algorithms for the joint object placement and request routing problem in content distribution networks," *Comput. Oper. Res.*, vol. 35, no. 12, pp. 3860–3884, Dec. 2008.

[73] G. Carofiglio, L. Mekinda, and L. Muscariello, "Joint forwarding and caching with latency awareness in information-centric networking," *Comput. Netw.*, vol. 110, pp. 133–153, Dec. 2016.

[74] M. Dehghan *et al.*, "On the complexity of optimal request routing and content caching in heterogeneous cache networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1635–1648, Jun. 2017.

[75] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proc. IEEE INFOCOM*, Apr./May 2014, pp. 1078–1086.

[76] S. Ioannidis and E. Yeh, "Jointly optimal routing and caching for arbitrary network topologies," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1258–1275, 2018.

[77] D. Tsilimantos, T. Karagkioules, and S. Valentin. (Mar. 2018). "Classifying flows and buffer state for YouTube's HTTP adaptive streaming service in mobile networks." [Online]. Available: https://arxiv.org/abs/1803.00303

[78] Z.-L. Zhang, J. Kurose, J. D. Salehi, and D. Towsley, "Smoothing, statistical multiplexing, and call admission control for stored video," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1148–1166, Aug. 1997.

[79] K. Suh *et al.*, "Push-to-peer video-on-demand system: Design and evaluation," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 9, pp. 1706–1716, Dec. 2007.

[80] F. Hartanto, J. Kangasharju, M. Reisslein, and K. Ross, "Caching video objects: Layers vs versions?" *Multimedia Tools Appl.*, vol. 31, no. 2, pp. 221–245, 2006.

[81] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[82] K. Poularakis, G. Iosifidis, A. Argyriou, I. Koutsopoulos, and L. Tassiulas, "Caching and operator cooperation policies for layered video content delivery," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[83] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. K. Ramakrishnan, "Optimal content placement for a large-scale VoD system," in *Proc. ACM CoNEXT*, 2010, Art. no. 4.

[84] Y. Sanchez *et al.*, "Efficient HTTP-based streaming using scalable video coding," *Signal Process., Image Commun.*, vol. 27, pp. 329–342, Apr. 2012.

[85] J. Dai, F. Liu, B. Li, B. Li, and J. Liu, "Collaborative caching in wireless video streaming through resource auctions," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 458–466, Feb. 2012.

[86] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[87] P. Ostovari, A. Khreishah, and J. Wu, "Multi-layer video streaming with helper nodes using network coding," in *Proc. IEEE MASS*, Oct. 2013, pp. 524–532.

[88] L. Pu, L. Jiao, X. Chen, L. Wang, Q. Xie, and J. Xu, "Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, to be published.

[89] M. Choi, J. Kim, and J. Moon, "Wireless video caching and dynamic streaming under differentiated quality requirements," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1245–1257, 2018.

[90] C. Ge, N. Wang, W. K. Chai, R. Bradbury, and H. Hellwagner, "QoE-assured 4K HTTP live streaming via transient segment holding at mobile edge," *IEEE J. Sel. Areas Commun.*, to be published.

[91] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

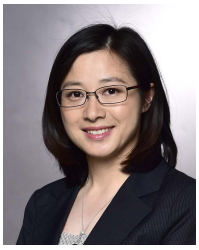
[92] B.-G. Chun, K. Chaudhuri, H. Wee, M. Barreno, C. H. Papadimitriou, and J. Kubiatowicz, "Selfish caching in distributed systems: A game-theoretic analysis," in *Proc. ACM PODC*, 2004, pp. 21–30.

[93] L. Wang, *et al.*, "Milking the cache cow with fairness in mind," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2686–2700, Oct. 2017.

[94] M. Taghizadeh, K. Micinski, C. Ofria, E. Torng, and S. Biswas, "Distributed cooperative caching in social wireless networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 6, pp. 1037–1053, Jun. 2013.
[95] K. Poularakis, G. Iosifidis, I. Pefkianakis, L. Tassiulas, and M. May, "Mobile data offloading through caching in residential 802.11 wireless networks," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 1, pp. 71–84, Mar. 2016.
[96] J. Krolikowski, A. Giovanidis, and M. Di Renzo, "A decomposition framework for optimal edge-cache leasing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1345–1359, 2018.
[97] E. Gourdin, P. Maillé, G. Simon, and B. Tuffin, "The economics of CDNs and their impact on service fairness," *IEEE Trans. Netw. Service Manage.*, vol. 14, no. 1, pp. 22–33, Mar. 2017.
[98] R. T. B. Ma and D. Towsley, "Cashing in on caching: On-demand contract design with linear pricing," in *Proc. ACM CoNEXT*, 2015, Art. no. 8.
[99] K. Hosanagar, R. Krishnan, J. Chuang, and V. Choudhary, "Pricing and resource allocation in caching services with multiple levels of quality of service," *Manage. Sci.*, vol. 51, no. 12, pp. 1844–1859, 2005.
[100] N. Economides and B. E. Hermalin, "The strategic use of download limits by a monopoly platform," *RAND J. Econ.*, vol. 46, no. 2, pp. 297–327, 2015.
[101] J. Kwak, G. Paschos, and G. Iosifidis, "Dynamic cache rental and content caching in elastic wireless CDNs," in *Proc. WiOpt*, May 2018, pp. 1–8.
[102] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Joint smart pricing and proactive content caching for mobile services," *IEEE/ACM Trans. Netw.*, vol. 24, no. 4, pp. 2357–2371, Aug. 2016.
[103] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai network: A platform for high-performance Internet applications," *ACM SIGOPS Oper. Syst. Rev.*, vol. 44, no. 3, pp. 2–19, 2010.
[104] O. Katz, R. Perets, and G. Matzliach, "Digging deeper—An in-depth analysis of a fast flux network," Akamai, Cambridge, MA, USA, White Paper, 2017.
[105] "The case for a virtualized CDN (vCDN) for delivering operator OTT video," Akamai, Cambridge, MA, USA, White Paper, Oct. 2017. [Online]. Available: https://community.akamai.com/docs
[106] *The Elastic CDN Solution*, document 3510532-001-EN, Akamai and Juniper, Dec. 2014.
[107] Google Global Cache. [Online]. Available: https://peering.google.com/#/infrastructure
[108] J. Leguay, G. S. Paschos, E. A. Quaglia, and B. Smyth, "CryptoCache: Network caching with confidentiality," in *Proc. IEEE ICC*, May 2017, pp. 1–6.
[109] K. Florence, "How netflix works with ISPs around the globe to deliver a great viewing experience," Netflix Blog, Tech. Rep., 2016.
[110] Q. Huang, K. Birman, R. van Renesse, W. Lloyd, S. Kumar, and H. C. Li, "An analysis of Facebook photo caching," in *Proc. ACM SOSP*, 2013, pp. 167–181.
[111] Amazon CloudFront Pricing. Accessed: May 25, 2018. [Online]. Available: https://aws.amazon.com/cloudfront/pricing/
[112] Cadami. *Media Content Distribution*. Accessed: May 25, 2018. [Online]. Available: http://cadami.net/
[113] S. Vassilaras *et al.*, "The algorithmic aspects of network slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, Aug. 2017.
[114] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 3358–3363.
[115] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. IEEE CISS*, Mar. 2016, pp. 320–325.
[116] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, Mar. 1999, pp. 126–134.
[117] G. Hasslinger, K. Ntougias, F. Hasslinger, and O. Hohlfeld, "Performance evaluation for new Web caching strategies combining LRU with score based object selection," *Comput. Netw.*, vol. 125, pp. 172–186, Oct. 2017.
[118] P. Cao and S. Irani, "Cost-aware WWW proxy caching algorithms," in *Proc. USENIX USITS*, 1997, p. 18.
[119] H. Che, Y. Tung, and Z. Wang, "Hierarchical Web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.
[120] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching: Why it matters and how to model it," in *Proc. ACM SIGCOMM*, 2013, pp. 5–12.
[121] E. Leonardi and G. Torrisi, "Least recently used caches under the shot noise model," in *Proc. IEEE INFOCOM*, Apr. 2015, pp. 2281–2289.
[122] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.
[123] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of Web content," *J. Internet Services Appl.*, vol. 5, p. 8, Dec. 2014.
[124] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. IEEE/WIC/ACM WI-IAT*, 2010, pp. 492–499.
[125] P. Blasco and D. Gündüz, "Multi-access communications with energy harvesting: A multi-armed bandit model and the optimality of the myopic policy," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 585–597, Mar. 2015.
[126] A. Sadeghi, F. Sheikholeslami, and G. B. Giannakis, "Optimal and scalable caching for 5G using reinforcement learning of space-time popularities," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 180–190, Feb. 2018.
[127] E. Bastug, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *Proc. WiOPT*, 2015, pp. 161–166.
[128] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart data pricing: Using economics to manage network congestion," *ACM Commun.*, vol. 58, no. 12, pp. 86–93, 2015.
[129] T. Giannakas, P. Sermpezis, and T. Spyropoulos, "Show me the cache: Optimizing cache-friendly recommendations for sequential content access," in *Proc. IEEE WoWMoM*, 2018.
[130] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.
[131] G. Iosifidis, L. Gao, J. Huang, and L. Tassiulas, "Incentive mechanisms for user-provided networks," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 20–27, Sep. 2014.
[132] V. Misra, S. Ioannidis, A. Chaintreau, and L. Massoulie, "Incentivizing peer-assisted services: A fluid shapley value approach," in *Proc. ACM SIGMETRICS*, 2010, pp. 215–226.
[133] B. Frank *et al.*, "Pushing CDN-ISP collaboration to the limit," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 2, pp. 34–44, 2013.
[134] Peer5 Tech Blog. *Akamai's Peer-to-peer Love Story*. Accessed: May 25, 2018. [Online]. Available: https://blog.peer5.com/akamais-peer-to-peer-love-story/
[135] T. Bektas, O. Oguz, and I. Ouveysi, "Designing cost-effective content distribution networks," *Comput. Oper. Res.*, vol. 34, no. 8, pp. 2436–2449, Aug. 2007.
[136] E. Baştuğ *et al.*, "Big data meets telcos: A proactive caching perspective," *J. Commun. Netw.*, vol. 17, no. 6, pp. 549–557, Dec. 2015.
[137] B. Addis, D. Belabed, M. Bouet, and S. Secci, "Virtual network functions placement and routing optimization," in *Proc. IEEE CloudNet*, Oct. 2015, pp. 171–177.

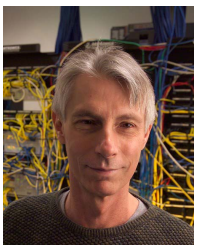
Georgios S. Paschos received the Diploma degree in electrical and computer engineering (ECE) from the Aristotle University of Thessaloniki, Greece, in 2002, and the Ph.D. degree in wireless networks from the ECE Department, University of Patras, Greece, in 2006. He held research positions at VTT, Finland, from 2007 to 2008, CERTH-ITI, Greece, from 2008 to 2012, and LIDS, MIT, USA, from 2012 to 2014. From 2009 to 2012, he was with the ECE Department, University of Thessaly. Since 2014, he has been a Principal Researcher with Huawei Technologies, Paris, France, where he has been leading the Network Control and Resource Allocation Team. He is a Technical Program Committee Member of INFOCOM, WiOPT, and Netsoft. Two of his papers received Best Paper Awards in GLOBECOM 2007 and IFIP Wireless Days 2009. He was an Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS special issue for content caching and delivery, and he actively serves as an Associate Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING, the IEEE NETWORKING LETTERS.

George Iosifidis received the Diploma degree in electronics and communications from the Greek Air Force Academy, Athens, in 2000, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Thessaly, in 2012. He was a Post-Doctoral Researcher with CERTH-ITI, Greece, from 2012 to 2014, and a Post-Doctoral/Associate Research Scientist with Yale University from 2014 to 2017. He is currently the Ussher Assistant Professor in Future Networks with the School of Computer Science and Statistics, Trinity College Dublin, Ireland. He was a co-recipient of the Best Paper Awards in WiOPT 2013 and the IEEE INFOCOM 2017 conferences, and received an SFI Career Development Award in 2018.



Meixia Tao (S'00–M'04–SM'10) received the B.S. degree in electronic engineering from Fudan University, Shanghai, China, in 1999, and the Ph.D. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology in 2003. She was a member of Professional Staff with the Hong Kong Applied Science and Technology Research Institute from 2003 to 2004, and a Teaching Fellow and then as an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore from 2004 to 2007. She is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. Her current research interests include wireless caching, physical-layer multicasting, resource allocation, and interference management.

Dr. Tao currently serves as a member of the Executive Editorial Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. She was on the Editorial Board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2007 to 2011, the IEEE COMMUNICATIONS LETTERS from 2009 to 2012, and the IEEE WIRELESS COMMUNICATIONS LETTERS from 2011 to 2015. She received the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2009. She was a recipient of the WCSP'12 Best Paper Award in 2012, the IEEE Heinrich Hertz Award for Best Communications Letters in 2013, and the IEEE/CIC ICCC'15 Best Paper Award in 2015.



Don Towsley received the B.A. degree in physics and the Ph.D. degree in computer science from the University of Texas in 1971 and 1975, respectively. He has held visiting positions at numerous universities and research labs, including the University of Paris VI, IBM Research, AT&T Research, Microsoft Research, and INRIA. He is currently a Distinguished Professor with the College of Information and Computer Sciences, University of Massachusetts. His research interests include security, quantum communication, and networks and performance evaluation. He is a fellow of the ACM. He is a Corresponding Member of the Brazilian Academy of Sciences. He has received numerous IEEE and ACM awards, including the 2007 IEEE Koji Kobayashi Award, the 2007 ACM SIGMETRICS Achievement Award, and the 2008 ACM SIGCOMM Achievement Award. He has also received numerous best paper awards, including the IEEE Communications Society 1998 William Bennett Paper Award, the 2008 ACM SIGCOMM Test of Time Award, the 10+ Year 2010 DASFAA Best Paper Award, the 2012 ACM SIGMETRICS Test of Time Award, and five ACM SIGMETRICS Best Paper Awards. He has served as a Program Co-Chair for numerous conferences and on the program committees of many other. He is a Co-Founder of the *ACM Transactions on Modeling and Performance Evaluation of Computing Systems* and served as one of its first Co-Editor-in-Chiefs. He served as the Editor-in-Chief for the IEEE/ACM TRANSACTIONS ON NETWORKING and on numerous other editorial boards.

Giuseppe Caire (S'92–M'94–SM'03–F'05) was born in Turin, Italy, in 1965. He received the B.Sc. degree in electrical engineering from the Politecnico di Torino, Italy, in 1990, the M.Sc. degree in electrical engineering from Princeton University in 1992, and the Ph.D. degree from the Politecnico di Torino in 1994. He was a Post-Doctoral Research Fellow with European Space Agency, ESTEC, Noordwijk, The Netherlands, from 1994 to 1995, an Assistant Professor in telecommunications with the Politecnico di Torino, an Associate Professor with the University of Parma, Italy, and a Professor with the Department of Mobile Communications, Eurecom Institute, Sophia-Antipolis, France. He 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, Technical University of Berlin, Germany. His main research interests are in the field of communications theory, information theory, and channel and source coding with particular focus on wireless communication. 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 has 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 the President of the IEEE Information Theory Society in 2011. He served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 1998 to 2001 and the IEEE TRANSACTIONS ON INFORMATION THEORY from 2001 to 2003.