1 Introduction

Ejder Baştuğ, Thang X. Vu, Symeon Chatzinotas, and Tony Q.S. Quek

1.1 History of Caching

The idea of caching can be traced back to the sixties in context of fast memory access in operating systems. According to [1], the most effective online cache replacement strategy is earliest deadline first (EDF) policy that replaces contents in memory blocks that are not going to be requested in the nearest future, provided that the future content demand profile is perfectly available. Other examples include analysis of least recently used (LRU) cache placement policy under stationary demand, either using a Markovian model [2] or an approximate cache placement method [3]. Implementation of these methods in the systems of that era are largely based on LRU, least frequently used (LFU), hybrid, and/or randomized approaches, while maintaining low-complexity design. The main objective therein is to maximize cache hits subject to ultra-low memory capacity constraints, thus achieving the goal of fast content access in such a localized setting.

The nineties witnessed an explosion of interconnected devices and systems over the internet, with the rise of world wide web contributing to creation of millions of websites, startups, and projects. This growth together with the *dot-com bubble* introduced a huge congestion over the web infrastructure, even leading its inventor, Sir Tim Berners-Lee, to mention the network congestion and the conventional client–server connectivity model as the main bottlenecks for the scalability of the web-based internet. Regarding the client–server model by which a web page is downloaded from the same centralized server by every internet client/user (often) multiple times, a workaround solution to alleviate the bottleneck was to replicate contents across several proxy servers that are geographically close to users, mostly supporting heuristic caching placements of unencrypted traffic and static contents. This allowed websites and content providers to minimize their global network bandwidth, provide rapid content access, and offload their servers.

The automation of this technical process and new business models ultimately led to creating content delivery networks (CDNs) around the late 1990s (see [4] for a brief literature review), and over the time, this evolution yielded complex infrastructures supporting various features, like the caching of large video streaming, social network data, and high-traffic websites, mostly over secure connections and distributed around the planet. Several cache placement strategies for CDN have been proposed [9] and implemented in this regard, allowing content providers to minimize access delays to the

https://doi.org/10.1017/9781108691277.002 Published online by Cambridge University Press

requested contents. The questions that have been targeted in this scope are (1) where to deploy the cache servers, (2) how much cache capacity is required for each node, (3) which content to cache, and (4) how and when to redirect/route the contents to end users. Concepts like content centric networking (CCN) and information centric networking (ICN) emerged during this period (see [5, 6] for examples), aiming to fundamentally shift the way that content is stored and accessed on the internet and looking to pave the way of successful implementations. Most of CDN-like caching technologies are now crucial elements of the world networking infrastructure.

Nowadays, all these networking-level caching challenges are being revisited in the wireless domain, mostly driven by the steep increase of mobile data traffic and billions of devices/users connected to the mobile networks, where telecom operators and organizations are looking for innovative ways to design and deploy cellular networks of the future. The aforementioned technical challenges of caching are not only being revisited but also taking a twisted step, with majority of traditional caching problems taking into account limited backhaul capabilities in dense cellular deployments, base station cooperation and coordination, coded/uncoded techniques, learning in large scale, mobility, economics, ultra-performance demanding new applications, level of content placement at the wireless edge (namely at base stations and user devices), and others. The research community is growing and aims to bring the wireless caching into reality (see [7, 8, 10] for examples), also with industrial and startup activities taking place. The aim of our book is to cover most of these technical aspects of wireless edge caching, with the help of experts and researchers active in the domain.

1.2 Summary of the Book

The book presents a collection of invited chapters on a wide range of issues and open challenges related to edge caching applied to future wireless networks, which are coherently presented in four parts:

- Part I: Optimal Cache Placement and Delivery
- Part II: Proactive Caching
- Part III: Cache-Aided Interference and Physical Layer Management
- Part IV: Energy-Efficiency, Security, Economics, and Deployment

Part I provides a comprehensive view on optimal cache design for both placement and delivery phases. The five chapters in this part cover most advanced techniques in coded caching, cache-aided device-to-device communications, and cooperative caching. More specifically, Chapter 2 provides the comprehensive performance analysis of coded caching in heterogeneous wireless networks via a joint design of storage and delivery and the optimal trade-off between the cache memory size and the broadcast delivery rate. Chapter 3 investigates the performance of cache-aided device-to-device networks under both uncoded and coded caching strategies. Chapter 4 proposes a cooperative hierarchical caching framework in cloud radio access networks (C-RAN) and explores the synergies of the in-network computing and storage resources. Chapter 5 proposes the concept

https://doi.org/10.1017/9781108691277.002 Published online by Cambridge University Press

of stochastic caching in large wireless networks and analyzes the three main performance metrics: cache-hit probability, successful delivery probability, and content delivery latency. Chapter 6 studies the edge caching via a joint design of caching, routing, and channel assignment for video delivery over coordinated small-cell cellular systems.

Part II provides key aspects in designing proactive caching algorithms with users' behavior prediction and learning techniques. In particular, Chapter 7 proposes a novel popularity-predicting-based caching procedure based on raw video data to determine an optimal cache placement policy, which deals with both published and unpublished videos. Chapter 8 studies wireless edge caching paradigms for mobile social networks to improve reliable and low-latency communication services for mobile users on social networking. In Chapter 9, a big data analytic-based framework is proposed for content popularity estimations and proactive caching at base stations. Chapter 10 investigates the impact of mobility on edge caching and proposes the optimal cache in both static and mobile user scenarios.

Part III consists of four chapters that provide the cross-layer cache-aided design for interference and physical layer resources management under both coded caching and uncoded methods. More precisely, Chapter 11 studies the caching effects on multicast-enabled access downlinks and proposes cache-aware joint designs of the content-centric base station clustering and multicast beamforming. Chapter 12 analyzes the impact of caching in the interference networks under both fully and partially connected topologies. Chapter 13 studies the performance enhancement brought by the caching capabilities in full-duplex radios in the context of ultra-dense networks. Chapter 14 investigates the impact of edge caching in mobile millimeter wave systems via a mobility management framework that exploits broadband millimeter wave connectivity to cache the contents of interest.

Part IV highlights the edge caching in future wireless networks from various aspects such as energy efficiency, security, and economics as well as the edge caching deployment in unmanned aerial vehicle (UAV) and virtual reality systems. Chapter 15 investigates the energy-efficiency and delivery time performance of the wireless edge caching systems via both uncoded and coded caching strategies. Chapter 16 studies the cache-enabled UAVs in C-RAN to reduce the content delivery latency and improve the users' quality of experience. Chapter 17 investigates the application of edge caching to enhance the physical layer security of cellular networks via proactive content sharing policy across a subset of base stations. Chapter 18 proposes a framework for the delivery of 360°-navigable videos to 5G virtual reality wireless clients in future cooperative multi-cellular systems. Finally, Chapter 19 investigates the elastic wireless edge caching that reveals the economic interactions of different stake holders in the network and provides the key differences between in-network and edge caching.

References

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

https://doi.org/10.1017/9781108691277.002 Published online by Cambridge University Press

[2] A. J. Smith, "Analysis of the optimal, look-ahead demand paging algorithms," *SIAM Journal on Computing*, vol. 5, no. 4, pp. 743–757, 1976.
[3] R. Fagin, "Asymptotic miss ratios over independent references," *Journal of Computer and System Sciences*, vol. 14, no. 2, pp. 222–250, 1977.
[4] J. Wang, "A survey of web caching schemes for the internet," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 5, pp. 36–46, 1999.
[5] A. Araldo, M. Mangili, F. Martignon, and D. Rossi, "Cost-aware caching: optimizing cache provisioning and object placement in ICN," arXiv:1406.5935, 2014.
[6] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26–36, 2012.
[7] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "The role of caching in future communication systems and networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1111–1125, 2018.
[8] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "Guest editorial caching for communication systems and networks—part II," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 8, pp. 1663–1665, Aug. 2018.
[9] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *INFOCOM, 2010 Proceedings IEEE*, IEEE, 2010, pp. 1–9.
[10] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 16–22, 2016.

https://doi.org/10.1017/9781108691277.002 Published online by Cambridge University Press